Contribution of three components to individual cancer risk predicting breast cancer risk in Italy

P Boyle[1], M Mezzetti[2], C La Vecchia[3,4], S Franceschi[5], A Decarli[6] and C Robertson[7]

We used data from a multicentre case–control study conducted in Italy between 1991 and 1994 on over 2500 cases of breast cancer and a comparable number of controls, and estimates of breast cancer incidence in Italy to compute individual breast cancer risk for Italian women. The estimated probabilities between age 50 and 80 ranged from approximately 5% (for a woman with no family history and low modifiable risk profile) to about 30% (for a woman with young family history and high modifiable risk) on the basis of various women's baseline characteristics. Expected numbers of breast cancer cases using the present model were compared with those based on the USA Gail model, and with the observed ones in the comparison group of the Italian Tamoxifen Trial. These show a closer agreement between the observed and the expected total numbers of breast cancers than the USA Gail model. Thus, the Gail model can be improved for use in other populations by using estimates of incidence and risk which are more appropriate to the target population. ***European***

Journal of Cancer Prevention **13:000–000 © 2004 Lippincott Williams & Wilkins.**

European Journal of Cancer Prevention 2004, **13**:000–000

Keywords: Alcohol, anthropometry, breast neoplasms, diet, family history, hormones, physical activity, risk assessment

[1]Division of Epidemiology and Biostatistics, European Institute of Oncology, via Ripamonti 435, 20141 Milan, Italy, [2]Istituto di Metodi Quantitativi, Università Bocconi, Viale Isonzo, 250, 20135 Milan, Italy, [3]Istituto Ricerche Farmacologiche "Mario Negri", via Eritrea 62, 20157 Milan, Italy, [4]Istituto di Biometria e Statistica Medica, Universita di Milano, via Venezian 1, 20133 Milan, Italy, [5]International Agency for Research on Cancer, 150 Cours Albert Thomas, 69372 Lyon Cedex, France, [6]Sezione Statistica Medica e Biometria, Università di Brescia, 25123 Brescia, Italy and [7]Department of Statistics and Modelling Science, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH, Scotland.

Correspondence to: C La Vecchia, Istituto di Ricerche Farmacologiche "Mario Negri", Via Eritrea 62, 20157 Milano, Italy.
E-mail: lavecchia@marionegri.it

Received 15 January 2004 **Accepted** 15 February 2004

Introduction

Historically, attempts have been made to estimate the burden of cancer in populations (Waterhouse *et al.*, 1977, 1982; Muir *et al.*, 1987; Parkin *et al.*, 1993, 1997, 2002) and then to attribute the causes of cancer in population groups (Wynder, 1952a,b,c; Higginson and Muir, 1977; Doll and Peto, 1981). A more recent phenomenon has been to attempt to calculate the lifetime risk of developing certain forms of cancer for individuals. The majority of this development has undoubtedly stemmed from the identification of genetic markers of high risk (Easton *et al.*, 1995; Gayther *et al.*, 1997; Struewing *et al.*, 1997; Thorlacius *et al.*, 1998) and the attempt to estimate the likely benefits of potential interventions (Schrag *et al.*, 1997).

The situation has been developing quickly in regard to breast cancer risk where, following increasing public awareness of breast cancer risk factors – particularly having a close family member with the disease – has created a demand for counselling. Apart from depending on the availability and utility of available control options, the decision to undertake such options also depends critically on an individualized estimate of the probability of developing breast cancer in a defined period. A method has been developed (Gail *et al.*, 1989) ('Gail Score') to estimate the chance that a woman of a given age and risk factor set has of developing breast cancer over a given period. This method has been widely used, for example, in identifying women at high risk of breast cancer for intervention studies (Breast Cancer Risk Tool, 1999; Costantino *et al.*, 1999; Rockhill *et al.*, 2001; Clamp *et al.*, 2002; Tartter *et al.*, 2002; Freedman *et al.*, 2003) or for women with family history of breast cancer (McTiernan *et al.*, 2001).

The first published version of the 'Gail Score' (Gail *et al.*, 1989) was calculated using information on the age at menarche of the woman, the age of the woman at the time she gave birth to her first child, the number of first-degree relatives of the woman who have breast cancer, and the number of previous breast biopsies that the woman has had. The subsequent revision of the score (Breast Cancer Risk Tool, 1999) includes a diagnosis of atypical hyperplasia as an additional factor. One of the factors conveys genetic information, two are reproductive and two are associated with previous or previously suspected breast disease.

When faced with an individual woman without cancer, it is clear that there are three components to her risk of developing a particular cancer that need to be identified

0959-8278 © 2004 Lippincott Williams & Wilkins DOI: 10.1097/01.cej.0000130014.83901.53

and considered separately. The first component is the genetic risk conferred at conception which creates a human being with a genetic mutation that carries an identified high risk of developing cancer of a particular type in her lifetime. The second component relates to those accepted risk factors which have already taken place and are irreversible: these include age at menarche, age at first and last birth, number of full-term pregnancies and the factors that may have been already set depending on the age of the woman. The third component is the set of risk factors that are – at least theoretically – available to alteration, such as diet, body mass index, use of hormone replacement therapy and alcohol consumption (D'Avanzo *et al.*, 1996; Mezzetti *et al.*, 1998). In a certain sense, the first two components are fixed and prospects for altering the risk of disease relate either to altering those risk factors in the third component which are amenable to change or to modify risk by surgical or chemical intervention.

We have data from a multicentre case–control study in Italy between 1991 and 1994 on over 2500 breast cancer cases and a comparable number of controls (Talamini *et al.*, 1996). To address the issues outlined above, we have attempted to verify the performance of the Gail model (Gail *et al.*, 1989) in separating cases and controls in this dataset and to attempt to identify and attribute the three components (genetic, fixed exposures and exposures which could be altered) of breast cancer risk. The aim would be to define a method to identify women and increased risk of breast cancer in the Italian population.

We also have data from the Italian chemoprevention study comparing tamoxifen with placebo in a randomized trial (Veronesi *et al.*, 2002, 2003). Following the report of the NSPAB1 trial showing that tamoxifen had a significant chemopreventative effect (Fisher *et al.*, 1998), the Italian study published its results to date showing that there was no overall evidence of tamoxifen reducing the risk of breast cancer in the group of women studied, but a favourable effect in oestrogen receptor-positive women (Veronesi *et al.*, 2002, 2003). We use these data as an independent validation of the Gail model and its application to the Italian population of women.

Materials and methods

The data were derived from a multicentre case–control study of breast cancer conducted between June 1991 and February 1994 in six Italian areas: Greater Milan, the urban area of Genoa, the provinces of Pordenone and Gorizia, the province of Forlí, the province of Latina and the urban area of Naples (La Vecchia *et al.*, 1995). Cases were 2569 women, aged 23–74 years (median age 55 years), admitted to the major teaching and general hospitals of the study areas with histologically confirmed breast cancer diagnosed within the year before interview, and no previous history of cancer.

Controls were 2588 women, aged 20–74 years (median age 56 years), and admitted to hospitals in the same catchment areas of cases for acute conditions. Women admitted for gynaecological, hormonal or neoplastic diseases, or for diseases related to known risk factors for breast cancer were not included. Twenty-two per cent of the controls were admitted for traumas, mostly fractures and sprains, 33% for non-traumatic orthopaedic diseases, 15% for surgical conditions, 18% for eye diseases and 12% for miscellaneous other conditions, such as ear, nose and throat, skin and dental conditions. The distributions of cases and controls in terms of age and area of residence were similar, although cases and controls were not individually matched. Less than 4% of cases and controls refused to participate.

The interviewers were trained centrally, the same structured questionnaire and coding manual were used in all study centres. The questionnaire included information on sociodemographic characteristics, such as education, occupation and socio-economic indicators; lifelong smoking habits; physical activity at selected ages; anthropometric measures before diagnosis and weight at various ages; alcohol and coffee consumption; dietary habits; personal medical history and selected questions regarding family history of cancer; gynaecological and reproductive history; and history of use of oral contraceptives, hormone replacement therapy (HRT), and female hormone preparations for other indications.

A validated food frequency questionnaire (FFQ) was used to assess the usual diet 2 years before diagnosis, in order to estimate the mean daily intake of calories and selected nutrients. Subjects had to report their average weekly consumption of 78 foods or food groups, including the most common recipes of the Italian diet, and of several types of alcoholic beverages. Reproducibility and validity of the FFQ were satisfactory (Franceschi *et al.*, 1995; Decarli *et al.*, 1996; Ferraroni *et al.*, 1996). To compute energy and nutrient intake, Italian food composition databases were used for about 80% of food items, and were integrated with other sources and information (Salvini *et al.*, 1996). Daily intakes of alcohol from different beverages were computed using the number of days per week each type of beverages was consumed, and the average number of drinks per day (Ferraroni *et al.*, 1998).

Two diet scores were constructed, though only one of them was used in the statistical model at one time. The first was a composite risk score obtained from beta-carotene and vitamin E, and the second one was based upon the number of portions of fresh fruit, weighted for

the months of availability, together with fresh and cooked vegetables (Franceschi *et al.*, 1996a, 1998; Negri *et al.*, 1996). Two scores are used because one is based upon nutrients that come from many different sources and the other is based upon observable portions of specific food types. We believe that the first score is based upon the nutrients that are associated with breast cancer risk while the second one is the main source of these nutrients. The latter is easy to modify by eating more fruits and vegetables, while modifying the first score is more complex.

The subjects were specifically asked how many sisters and brothers they had and whether parents, siblings, children, grandparents or spouse had ever had cancer (except skin cancer). For each relative with history of cancer, the subjects were asked to specify whether the relative was still alive at the time of interview, current age or age at death, site of the tumour, and age at diagnosis. In this model we only used information on breast cancer among female first-degree relatives.

Physical activity has also been identified as a risk factor for breast cancer in Italy (D'Avanzo *et al.*, 1996) and our previous analysis has suggested that low activity is associated with a 14% attributable risk among postmenopausal women but only 7% among premenopausal women (Mezzetti *et al.*, 1998). We used occupational physical at ages 30–39 as the best single measure for physical activity. The original questionnaire response was recorded on a 5-point scale, was included in this model on a 3-point scale. We only considered physical activity as a risk factor among postmenopausal women as the attributable risk among premenopausal women is low (D'Avanzo *et al.*, 1996).

Data from the Italian randomized trial for the prevention of breast cancer were used to validate the model. This trial began in October 1992 and is a double-blind placebo-controlled trial of tamoxifen at 20 mg per day for 5 years among women without breast cancer and who had previously had a hysterectomy. Recruitment to the trial ended in June 1997 and a total of 5408 women were randomized. The median follow-up time for breast cancer is 82 months as of February 2001 and 79 cases for breast cancer have been reported, with 58 during the first 5 years.

The women in the trial are all hysterectomized, and did not have any benign breast disease prior to entry to the study. Information on all of the other risk factors in the Gail model was collected. Detailed information was also collected on the presence of breast cancer in the mother, sisters or daughters of the participants in the trial. Only the presence or absence of cancer in the sisters or daughters was recorded, so the total number of first-degree relatives with breast cancer calculated in this paper will be a slight underestimate of the total. A similar problem occurred with previous breast biopsies. Information was collected only if there had been any previous biopsies (11% of women had at least one biopsy prior to the study), but the number of biopsies was not recorded. Again this will cause a slight underestimate of the effect of biopsies.

The overall approach with the data sets is to use the case–control study for the estimation of the parameters and some cross validation. The randomized chemoprevention trial is used only as part of the validation exercise of all the models. Cancer registry data are used to estimate the age specific incidence of breast cancer.

Statistical methods

Odds ratios (OR) and the corresponding 95% confidence intervals (CI) in strata of age, younger than 50 years old and older than 50, were computed using unconditional multiple logistic regression models (Breslow and Day, 1980). The fitted model included terms for centre, education, energy intake, age at menarche, age at first birth, alcohol intake, family history, age of diagnosis in relatives, and one of the two diet scores. Body mass index and HRT were included only for women older than 50 years (La Vecchia *et al.*, 1995; Franceschi *et al.*, 1996b). Both diet scores were weighted for the excess risk obtained from a logistic regression model adjusting also for age, family history, centre, education and energy intake. Allowance for total energy intake was made by using the residual regression method (Willett and Stampfer, 1986).

Mutations in known and unknown breast cancer susceptibility genes account for an estimated 5–10% of cases of breast cancer. The two major breast cancer genes are the *BRCA1* gene on chromosome 17q and *BRCA2* on chromosome 13q. Mutations in these genes are rare in the general population. Studies of the Icelandic population have shown that this mutation is found in 7.7% (5.7–9.7%) of unselected female breast cancer patients and in an estimated 0.6% (0.1–1.7%) of the general population (Thorlacius *et al.*, 1998). Given the rarity of this gene mutation, we believe it is not necessary to include it in the calculation of the attributable fraction of the exposure variables. We included as a risk factor having a first-degree relative with breast cancer at younger age as a proxy variable for the possible presence of a *BRCA* mutation in the family (Negri *et al.*, 1997).

The projections of individual probabilities are based on an assumption of a proportional hazard model. The baseline breast cancer specific hazard $h_1(t)$ at age or age group t, for a subject without identified risk factors, will be estimated according to Gail *et al.* (1989) and Benichou and Gail (1995), by:

$$h_1(t) = h^*(t)(1 - \mathrm{AR}(t))$$

where $h^*(t)$ is the overall age-specific incidence rate for breast cancer irrespective of risk group, and $\mathrm{AR}(t)$ is the attributable risk estimate, according to Bruzzi *et al.* (1985), of the specified factors which, in turn, define the different risk groups.

After obtaining $h_1(t)$, it is possible to estimate the probability of developing breast cancer between the ages a_1 and a_2 when the following values are known for each age (or age group) t: the relative risk of developing breast cancer for each risk group compared with the reference group; the mortality rate from all causes of death, except breast cancer, in the population; the probability of surviving all other causes of death in the population.

The age-specific survival and mortality rates from all other causes are obtained from population registers of death rates, using the observed rates and without making any parametric assumptions. All the other causes are considered as competitive causes of death. The relative risks are obtained from the analysis of the case–control study.

The attributable risk of the specified factors which, in turn, defines the different risk groups (Bruzzi *et al.*, 1985; Mezzetti *et al.*, 1996) was based upon the inversion of a decomposition of the overall incidence rate into the contributions of the individual risk groups (Benichou and Gail, 1995).

The confidence intervals of the projected probabilities are obtained using the parametric bootstrap (Efron and Tibshirani, 1993). They are obtained considering the relative risk and the population attributable fraction as the two major source of uncertainties. The logistic coefficients are sampled from a multivariate normal distribution and the proportions of cases are sampled from a multinomial distribution. This produces a sampling distribution for the population attributable fraction allowing the calculation of the standard error, and hence the confidence interval using the percentile method; 5000 bootstrap simulations were used. We did not take into account any sampling variation in the other causes mortality and overall incidence rates, as the sampling error of these is much less than in the other two components.

The validation of the USA Gail model was carried out by calculating the 5-year probability of breast cancer for each woman in the study, using the supplied disk (Breast Cancer Risk Tool, 1999), and correcting this for the length of time each woman had been in the study using linear interpolation. A slight error is introduced by this linear interpolation. The supplied disk uses the breast cancer rates and overall mortality rates for women in the USA and also the relative risk calculated from a population of women in the USA. A second validation of the Gail model, the Italian-Gail model, comes from using the breast cancer and overall incidence rates for women in Italy and using relative risks estimated from the Italian case–control study. This is carried out on an individual year basis and no interpolation is necessary. In both cases we calculated the expected number of cases of breast cancer in various subgroups of women and compare this to the observed number. The results were not separated out into tamoxifen and placebo, and are reported separately for total follow-up and for follow-up to a maximum of 5 years.

The breast cancer incidence rates were obtained as a weighted average of the age-specific breast cancer incidence rates in 1988–1992, from the registries in the five centres of the multicentre case–control study (Zanetti *et al.*, 1997). Thus, the incidence rates are derived from Italian registries and correspond to the population from which we estimated the odds ratios.

Results

Individual risk model for Italy with modifiable factors

The risk factors and their levels are given in Table 1, together with the distribution among cases and controls. The corresponding odds ratios are given in Table 2. Once the variables have been coded in classes, they are included in the logistic regression models as ordinal variables. A comparison with the relative risks obtained from the same variable considered as a categorical variable with separate relative risks for each class (results not shown) justified this procedure. Investigation of interaction terms showed that it was necessary to include interactions between age at menarche and age at first birth in women under 50, and the interaction between the beta-carotene and vitamin E diet score and body mass index in women over or equal 50. No interaction between BMI and the fruit and vegetable diet score was needed. As one might anticipate, there are little differences in the relative risks in the two separate models using the different diet scores.

The population attributable risk fraction for the indicated risk factors is 59% in women younger than 50 years old, and 54% for women older than 50 years old, using the beta-carotene and vitamin E diet score. The corresponding figures for the fruit and vegetable diet score are 54% and 60%. We consider this estimation of the population attributable fraction as the major source of uncertainty, as it involves the estimation of the joint effect of many exposure factors. Our bootstrap procedure for the estimation of confidence interval of individual projection includes this uncertainty.

Table 1 Numbers (%) of cases and controls according to selected variables. The scores used for the variables in the logistic regression are shown in brackets

	Younger than 50 years old		At least 50 years old	
	Cases	Controls	Cases	Controls
Age at menarche				
(2) ≤ 12	446 (53.1%)	381 (49.5%)	677 (39.1%)	687 (37.8%)
(1) 13–14	338 (40.3%)	326 (42.4%)	741 (42.8%)	772 (42.4%)
(0) ≥ 15	55 (6.6%)	62 (8.1%)	312 (18.1%)	360 (19.8%)
Age at first birth				
(1) ≤ 20	90 (10.7%)	151 (19.6%)	141 (8.2%)	226 (12.4%)
(2) 20–24	239 (28.5%)	256 (33.3%)	432 (25.0%)	547 (30.1%)
(3) 25–29 none[a]	399 (47.6%)	317 (41.2%)	857 (49.5%)	794 (43.7%)
(4) ≥ 30	111 (13.2%)	45 (5.9%)	300 (17.3%)	252 (13.9%)
Family history				
(0) No	729 (86.9%)	731 (95.1%)	1541 (89.1%)	1722 (94.7%)
(1) Yes	110 (13.1%)	38 (4.9%)	189 (10.9%)	97 (5.3%)
Age of relative at diagnosis				
(1) ≤ 40	18 (2.1%)	3 (0.4%)	28 (1.6%)	13 (0.7%)
(0) >40 or none[b]	821 (97.9%)	766 (99.6%)	1702 (98.4%)	1806 (99.3%)
Diet score beta-carotene and vitamin E[c]				
(4) 1 quintile	162 (19.3%)	192 (25.0%)	285 (16.5%)	328 (18.0%)
(3) 2 quintile	135 (16.1%)	118 (15.3%)	229 (13.2%)	263 (14.5%)
(2) 3 quintile	194 (23.1%)	161 (20.9%)	362 (20.9%)	391 (21.5%)
(1) 4 quintile	205 (24.4%)	186 (24.2%)	427 (24.7%)	370 (20.3%)
(0) 5 quintile	143 (17.1%)	112 (14.6%)	427 (24.7%)	467 (25.7%)
Diet score fruit and vegetables[c]				
(4) 1 quintile	167 (18.7%)	168 (21.8%)	265 (15.3%)	342 (18.8%)
(3) 2 quintile	186 (22.1%)	147 (19.1%)	322 (18.6%)	373 (20.5%)
(2) 3 quintile	155 (18.5%)	160 (20.8%)	314 (18.2%)	329 (18.1%)
(1) 4 quintile	170 (20.3%)	163 (21.2%)	430 (24.9%)	375 (20.6%)
(0) 5 quintile	171 (20.4%)	131 (17.0%)	399 (23.1%)	400 (22.0%)
Alcohol (g/day)[d]				
(0) Non-drinker	298 (35.5%)	362 (47.1%)	617 (35.7%)	686 (37.7%)
(1) $>15.3+$ ex	273 (32.5%)	223 (29.0%)	599 (34.6%)	621 (34.1%)
(2) ≥ 15.3	268 (32.0%)	184 (23.9%)	514 (29.7%)	512 (28.2%)
Body mass index (kg/m^2)				
(0) <23.3			480 (27.8%)	520 (28.6%)
(1) 23.3–26.6			587 (33.9%)	608 (33.4%)
(2) ≥ 26.6			663 (38.3%)	691 (38.0%)
Hormone replacement therapy				
(0) Never			1553 (89.8%)	1651 (90.8%)
(1) Used			177 (10.2%)	168 (9.2%)
Physical activity[e]				
(0) High			288 (16.7%)	366 (20.1%
(1) Moderate			1285 (74.5%)	1346 (74.0%)
(2) Low			153 (8.9%)	106 (5.8%)

[a]Nulliparous women included in this category.
[b]No family history included in this category.
[c]Quintiles on the weighted by excess risk sum of the residual on the regression.
[d]On the residual regression method.
[e]Occupational physical activity aged 30–39.

The projections in Table 3 are based on the assumption that the age-specific hazard of dying of causes other than breast cancer, obtained from the Italian population register, is the same for all subjects. The desired probability is obtained by considering breast cancer and the other causes of dying as competing risks. The differences between the estimated probabilities using the two diet scores is due to differences in the estimated logistic regression coefficients principally for diet, family history, the age at which the affected relative was diagnosed with cancer and body mass index. The estimated probabilities between age 50 and 80 ranged from approximately 5% to about 30%, on the basis of various women's baseline characteristics.

Validation of Gail model (USA) and Gail model (Italy)

The expected numbers of cases of breast cancer for the Gail model (USA) are presented in Table 4 for the subgroups of women with different values on the risk factors in the model. Although the numbers of invasive breast cancers predicted is greater than that observed, there is no significant difference between the observed and expected ($P = 0.32$). The ratio of observed to expected is 0.89 with exact 95% confidence limits of 0.70–1.09. These show reasonably good agreement, even allowing for the low power of this validation, in view of the low frequency of breast cancer cases. The over-estimation is greatest among older women. Using only the data up to a 5-year follow-up, we have a ratio of observed

Table 2 Odds ratios (and 95% confidence intervals) of breast cancer in relation to selected covariates. Italy, 1991–1994. The score for the variables are listed in Table 1

	Odds ratios (95% confidence intervals)[a]	
	Younger than 50 years old	At least 50 years old
(a) Beta-carotene and vitamin E diet score		
Age at menarche	0.83 (0.60–1.17)	1.04 (0.94–1.14)
Age at 1st birth	1.05 (0.77–1.43)	1.24 (1.14–1.34)
Age at menarche and age at 1st birth interaction	1.25 (1.03–1.53)	
Family history	2.23 (1.39–3.60)	2.05 (1.53–2.74)
Age relative ≤ 40	2.27 (0.67–7.65)	1.30 (0.66–2.55)
Diet score (beta-carotene and vitamin E)	1.16 (1.06–1.27)	1.07 (0.98–1.16)
BMI		0.98 (0.83–1.15)
Diet score and BMI interaction		1.06 (1.00–1.13)
Alcohol intake	1.35 (1.19–1.53)	1.05 (0.97–1.15)
HRT use		1.16 (0.92–1.46)
Physical activity		1.17 (1.01–1.35)
(b) Fruits and vegetables diet score		
Age at menarche	0.83 (0.59–1.16)	1.04 (0.95–1.14)
Age at 1st birth	1.04 (0.76–1.42)	1.23 (1.13–1.33)
Age at menarche and age at 1st birth interaction	1.26 (1.03–1.53)	
Family history	2.44 (1.61–3.69)	2.12 (1.61–2.79)
Age relative ≤ 40	3.14 (0.85–11.68)	1.16 (0.56–2.37)
Diet score (fruits and vegetables)	1.06 (0.98–1.14)	1.12 (1.06–1.18)
BMI		1.11 (1.02–1.21)
Alcohol intake	1.36 (1.20–1.55)	1.05 (0.96–1.15)
HRT use		1.15 (0.92–1.45)
Physical activity		1.15 (0.99–1.32)

[a]Derived from multiple logistic regression equations including terms for centre, calorie intake, education plus the above variables.

to expected breast cancer cases of 0.86 (95% CI 0.64, 1.08).

The Gail model (Italy) was calculated for all women in the case–control study. There were some differences between the distribution of variables and risk factors in the Italian data and 'Gail population'. Only 1.45% of Italian women had a history of previous biopsy compared with 20% in the data which were used to construct the Gail model (Gail *et al.*, 1988). Similarly, 92.6% of Italian women had no family history of breast cancer compared with 81.6% of the Gail sample. Also, 0.6% of Italian women had more than one relative with breast cancer compared with 2% in Gail's women.

The expected numbers of cases of breast cancer from the Gail model (Italy) are also presented in Table 4. These show a closer agreement between the observed and expected total numbers of breast cancers than the USA Gail model. The observed to expected ratio is 0.96 (95% CI 0.75, 1.16), for the complete follow-up and 0.92 (95% CI 0.68, 1.16) for a maximum of 5 years follow-up. This implies that part of the slight overestimation in risk associated with the USA Gail model can be explained by differences in the incidence rates and coefficients between Italy and the USA.

Finally, we compared the Gail model (Italy) and our new individual risk models with modifiable factors. We calculated a score for each subject and compared the scores of the cases and controls through the area under the ROC curve. The two models are comparable: the Gail model gives an area of 0.582 and our model, based upon the beta-carotene and vitamin E diet score, gives a slightly higher area of 0.593, and 0.600 when the fruit and vegetable diet score is used. A cross-validation procedure has been implemented. The two models are fitted in 70% of the subjects, randomly selected, and the area under the ROC curve is estimated in the subjects left. Using 500 cross-validation samples yields a mean area of 0.565 (0.0194) for the Gail model (Italy) and an area of 0.589 (0.0196) for our model.

We constructed five equal groups according to the distribution of the score in the entire population. We then used this score as the only predictor variable in a logistic regression model of breast cancer risk in the case–control study with the baseline risk set to 1 in the lowest fifth, adjusting for centre and education. The risk of breast cancer associated with the Gail score (Italy) rose from 1 (referent category) through to attain 2.1 in the highest quintile (Table 5). The corresponding relative risk for the highest quintiles of the new risk model based upon the nutrient diet score is 2.8 and for the score based upon fruits and vegetables is also 2.8. This implies that the introduction of a diet component leads to a better discrimination between those at higher risk and those at moderate and lower risk.

Discussion

This study has shown that the Gail model (USA, Breast Cancer Risk Tool, 1999) is reasonably valid in Italy, though with a propensity to overestimate risk. Admittedly, the validation exercise has a low power as few cancers have been observed. Also, none of the women in the validation data set had atypical hyperplasia or a history of breast disease. The USA Gail model, which was developed for USA women undergoing a screening programme (Gail *et al.*, 1989), can be improved for use in Italy by using estimates which are more appropriate to Italy, for example Italian incidence rates. This confirms observation made in other populations, such as African–American women (Bondy and Newman, 2003).

Some uncertainty in our estimates is given by the application of national incidence estimates in a cohort of women who had undergone hysterectomy, whose breast cancer risk may be lower (by about 20%), than that of the general population (Parazzini *et al.*, 1997). This may explain the small deficit of observed cases in

Table 3 Estimates of probabilities (% and 95% bootstrap confidence intervals) of developing breast cancer for selected group of subjects

	Age	Beta-carotene and vitamin E diet score	Fruit and vegetable diet score
Subject 1: Family history, late menarche, early first birth, average modifiable risk, no HRT			
AM 15+; AF ≤ 20; FH yes;	30–50	2.58 (2.34, 2.83)	2.22 (2.05, 2.40)
AR 41+; DS 2; AL 1; PE 1;	50–60	2.30 (2.21, 2.40)	2.28 (2.18, 2.37)
BMI 25; HR no	50–70	4.80 (4.61, 5.00)	4.75 (4.56, 4.94)
	50–80	6.89 (6.62, 7.17)	6.81 (6.54, 7.09)
Subject 2: Young family history, late menarche, early first birth, average modifiable risk, no HRT			
AM 15+; AF ≤ 20; FH yes;	30–50	5.00 (2.92, 8.57)	4.99 (2.59, 9.64)
AR ≤ 40; DS 2; AL 1; PE 1;	50–60	2.99 (2.37, 3.76)	2.63 (2.04, 3.38)
BMI 25; HR no	50–70	6.20 (4.93, 7.78)	5.47 (4.27, 7.00)
	50–80	8.85 (7.08, 11.07)	7.83 (6.14, 9.98)
Subject 3: Average menarche, late age at first birth no family history, high modifiable risk, no HRT			
AM 13–14; AF ≥ 30; FH no;	30–50	4.61 (4.36, 4.87)	3.38 (3.21, 3.56)
AR –; DS 3; AL 2; PE 2;	50–60	4.35 (4.25, 4.45)	3.76 (3.68, 3.84)
BMI>27; HR no	50–70	8.96 (8.76, 9.17)	7.77 (7.61, 7.93)
	50–80	12.70 (12.43, 12.99)	11.05 (10.83, 11.28)
Subject 4: Average menarche, late age at first birth, no family history, low to moderate modifiable risk, no HRT			
AM 13–14; AF ≥ 30; FH no;	30–50	2.55 (2.41, 2.70)	2.14 (2.02, 2.26)
AR –; DS 1; AL 1; PE 1;	50–60	2.39 (2.35, 2.44)	2.27 (2.23, 2.31)
BMI 25; HR no	50–70	4.98 (4.89, 5.08)	4.73 (4.65, 4.82)
	50–80	7.15 (7.02, 7.28)	6.79 (6.68, 6.91)
Subject 5: Average menarche, no children, no family history, average modifiable risk, HRT			
AM 13–14; AF none; FH no;	30–50	2.43 (2.33, 2.52)	1.96 (1.89, 2.03)
AR –; DS 2; AL 1; PE 1;	50–60	2.53 (2.46, 2.60)	2.38 (2.31, 2.45)
BMI 25; HR yes	50–70	5.27 (5.12, 5.42)	4.96 (4.82, 5.11)
	50–80	7.55 (7.34, 7.76)	7.11 (6.91, 7.32)
Subject 6: Average menarche, no children, no family history, low modifiable risk, HRT			
AM 13–14; AF none; FH no;	30–50	1.61 (1.55, 1.68)	1.33 (1.28, 1.39)
AR –; DS 1; AL 0; PE 0;	50–60	1.76 (1.70, 1.82)	1.60 (1.55, 1.65)
BMI <23; HR yes	50–70	3.68 (3.57, 3.79)	3.35 (3.25, 3.46)
	50–80	5.29 (5.13, 5.46)	4.83 (4.68, 4.98)
Subject 7: Family history, age menarche 12, first birth at 25–29, no HRT, highest diet quintile, no alcohol, high exercise, low BMI			
AM ≤ 12; AF 25–29; FH yes;	45–50	1.26 (1.20, 1.32)	1.14 (1.09, 1.19)
AR 41+; DS 0; AL 0; PE 0; BMI <23; HR no	45–90	10.52 (10.07, 10.99)	9.53 (9.15, 9.94)
Subject 8: Young family history, highest diet quintile, no alcohol, high exercise, low BMI			
AM ≤ 12; AF 25–29; FH yes;	45–50	1.63 (1.29, 2.07)	1.31 (1.01, 1.70)
AR ≤ 40; DS 0; AL 0; PE 0; BMI < 23; HR no	45–90	13.42 (10.79, 16.70)	10.92 (8.57, 13.91)
Subject 9: Young family history, middle diet quintile, no alcohol, high exercise, average BMI			
AM ≤ 12; AF 25–29; FH yes;	45–50	2.38 (1.88, 3.00)	2.07 (1.61, 2.68)
AR ≤ 40; DS 2; AL 0; PE 0; BMI 25; HR no	45–90	18.91 (15.35, 23.31)	16.72 (13.25, 21.09)
Subject 10: Young family history, middle diet quintile, moderate alcohol, moderate exercise, average BMI			
AM ≤ 12; AF 25–29; FH yes;	45–50	2.50 (1.98, 3.16)	2.18 (1.69, 2.81)
AR ≤ 40; DS 2; AL 1; PE 1; BMI 25; HR no	45–90	19.80 (16.09, 24.37)	17.48 (13.87, 22.01)
Subject 11: Young family history, lowest diet quintile, moderate alcohol, moderate exercise, average BMI			
AM ≤ 12; AF 25–29; FH yes;	45–50	3.18 (2.53, 4.01)	2.71 (2.10, 3.50)
AR ≤ 40; DS 4; AL 1; PE 1; BMI 25; HR no	45–90	24.47 (20.03, 29.90)	21.27 (16.99, 26.63)
Subject 12: Young family history, lowest diet quintile, high alcohol, low exercise, average BMI			
AM ≤ 12; AF 25–29; FH yes;	45–50	3.91 (3.11, 4.93)	3.25 (2.52, 4.19)
AR ≤ 40; DS 4; AL 2; PE 2; BMI 25; HR no	45–90	29.16 (24.05, 35.37)	24.91 (20.01, 31.00)

AM, age at menarche; AF, age at first birth; FH, family history; AR, age at diagnosis for relative with breast cancer; DS, diet score; AL, alcohol; PE, physical exercise; BM, body mass index; HR, hormone replacement therapy.

our validation study. More important, our projected probabilities of getting breast cancer have not been validated on an independent database and this is an important task before widespread use of such a model.

Using the structure of the original Gail model (USA) we have developed an individual risk model for women in Italy which has three components, one of which is potentially modifiable. This gives an important tool for individuals who are involved with breast cancer risk counselling. With the current Gail model, the only options for women who are perceived to be at high risk are some form of chemoprevention. Using the NSAPB1 results a counsellor might suggest that 5 years on tamoxifen would reduce an individual's 5 year risk of breast cancer by 40%. With the new model, the counsellor has the opportunity to look at the lifestyle of the women and possibly suggest that this be modified as an alternative to chemoprevention, or in conjunction with chemoprevention. The results show that the effect of changing diet from highest quintile to 3rd is associated

Table 4 Comparison of observed (O) and expected (E) numbers of breast cancer cases in the Italian tamoxifen trial. The numbers of women (*N*) and person years at risk (PY) are also presented

		5-year follow-up				Total			
	N	PY	E-US	E-IT	O	PY	E-US	E-IT	O
Overall Age	5383	33722	88.4	82.5	79	25827	67.6	63.1	58
<50	1871	11632	22.3	23.0	24	8964	17.1	17.7	20
≥50	3512	22090	66.1	59.5	55	16864	50.5	45.4	38
Age at menarche									
≥14	1466	9361	23.5	22.4	20	7076	17.8	16.9	16
12–13	2615	16204	42.5	39.3	39	12517	32.7	30.2	26
<12	1302	8157	22.4	20.8	20	6234	17.1	15.9	16
Age at first birth									
<20	420	2518	4.2	3.3	4	1993	3.4	2.6	3
20–24	2157	13522	30.8	26.1	25	10387	23.7	20.1	16
25–29+none	2368	14955	43.5	42.4	39	11376	33.1	32.2	30
≥30	2727	9.8	10.6	11	2072	7.5	8.1	9	
Number of first degree relatives									
0	4744	29744	69.1	63.8	63	22769	52.8	48.7	44
1	609	3790	17.6	17.0	16	2911	13.5	13.0	14
2	30	188	1.7	1.7	0	147	1.3	1.3	0
Biopsy									
No	4578	28701	71.0	63.5	61	21979	54.4	48.6	45
es	805	5021	17.3	18.9	18	3848	13.2	14.5	13

E-US, expected numbers of breast cancers based upon the Gail model US data; E-IT, Gail model re-estimated coefficients, incidence and mortality using Italian data.

Table 5 Odds ratios (95% confidence intervals of breast cancer) based upon a logistic regression of case–control status on the quintile of the probability score

	Gail model re-estimated using Italian data	Model based on beta-carotene and vitamin E diet score	Model based on fruit and vegetable diet score
1 quintile	1	1	1
2 quintile	1.05 (0.88–1.26)	1.40 (1.17–1.66)	1.33 (1.11, 1.58)
3 quintile	0.96 (0.81–1.15)	1.43 (1.20–1.71)	1.56 (1.31, 1.86)
4 quintile	1.28 (1.08–1.52)	1.94 (1.63–2.32)	1.96 (1.64, 2.34)
5 quintile	2.10 (1.75–2.52)	2.77 (2.31–3.32)	2.83 (2.36, 3.40)

+Reference category.

with a reduction in the risk of breast cancer by 34% in younger women and of 14% in older ones. For women in the highest group of alcohol consumption eliminating alcohol is associated with a reduction in breast cancer risk of 19% in younger women and only 10% in older ones.

The new models for predicting the probability of developing breast cancer, which include the modifiable factors such as diet and alcohol consumption, could not be validated using data from the breast cancer chemoprevention trial as no dietary information was recorded. We used therefore cross-validation techniques on the case–control data to investigate the predictions of this model.

These limitations notwithstanding, the findings of the present work indicate that the Gail model can be improved for use in populations other than the American one, by using breast cancer incidence and relative risk estimates for risk factors of interest which are more appropriate to the target population.

Acknowledgements

This work was conducted with the contributions of the Associazione Italiana per la Ricerca sul Cancro and the Italian League against Cancer.

References

Benichou J, Gail MH (1995). Methods of inference for estimates of absolute risk derived from population-based case–control studies. *Biometrics* **51**: 182–194.

Bondy ML, Newman LA (2003). Breast cancer risk assessment models: applicability to African–American women. *Cancer* **97**: 230–235.

Breast Cancer Risk Tool (1999). *National Cancer Institute*. (GENERIC) Ref Type: Computer Program.

Breslow NE, Day NE (1980). *Statistical Methods in Cancer Research*, Vol 1: *The Analysis of Case–Control Studies*. Lyon: IARC.

Bruzzi P, Green SB, Byar DP, *et al.* (1985). Estimating the population attributable risk for multiple risk factors using case–control data. *Am J Epidemiol* **122**: 904–914.

Clamp A, Danson S, Clemons M (2002). Hormonal risk factors for breast cancer: identification, chemoprevention, and other intervention strategies. *Lancet Oncol* **3**: 611–619.

Costantino JP, Gail MH, Pee D, *et al.* (1999). Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J Natl Cancer Inst* **91**: 1541–1548.

D'Avanzo B, Nanni O, La Vecchia C, *et al.* (1996). Physical activity and breast cancer risk. *Cancer Epidemiol Biomarkers Prev* **5**: 155–160.

Decarli A, Franceschi S, Ferraroni M, *et al.* (1996). Validation of a food-frequency questionnaire to assess dietary intakes in cancer studies in Italy. Results for specific nutrients. *Ann Epidemiol* **6**: 110–118.

Doll R, Peto R (1981). The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *J Natl Cancer Inst* **66**: 1191–1308.

Easton DF, Ford D, Bishop DT (1995). Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *Am J Hum Genet* **56**: 265–271.

Efron B, Tibshirani RJ (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Ferraroni M, Decarli A, Franceschi S, *et al.* (1996). Validity and reproducibility of alcohol consumption in Italy. *Int J Epidemiol* **25**: 775–782.

Ferraroni M, Decarli A, Franceschi S, La Vecchia C (1998). Alcohol consumption and risk of breast cancer: a multicentre Italian case–control study. *Eur J Cancer* **34**: 1403–1409.

Fisher B, Costantino JP, Wickerham DL, *et al.* (1998). Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst* **90**: 1371–1388.

Franceschi S, Barbone F, Negri E, *et al.* (1995). Reproducibility of an Italian food frequency questionnaire for cancer studies. Results for specific nutrients. *Ann Epidemiol* **5**: 69–75.

Franceschi S, Favero A, Decarli A, *et al.* (1996a). Intake of macronutrients and risk of breast cancer. *Lancet* **347**: 1351–1356.

Franceschi S, Favero A, La Vecchia C, *et al.* (1996b). Body size indices and breast cancer risk before and after menopause. *Int J Cancer* **67**: 181–186.

Franceschi S, Parpinel M, La Vecchia C, *et al.* (1998). Role of different types of vegetables and fruit in the prevention of cancer of the colon, rectum, and breast. *Epidemiology* **9**: 338–341.

Freedman AN, Graubard BI, Rao SR, *et al.* (2003). Estimates of the number of US women who could benefit from tamoxifen for breast cancer chemoprevention. *J Natl Cancer Inst* **95**: 526–532.

Gail MH, Brinton LA, Byar DP, *et al.* (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* **81**: 1879–1886.

Gayther SA, Mangion J, Russell P, *et al.* (1997). Variation of risks of breast and ovarian cancer associated with different germline mutations of the BRCA2 gene. *Nat Genet* **15**: 103–105.

La Vecchia C, Negri E, Franceschi S, *et al.* (1995). Hormone replacement treatment and breast cancer risk: a cooperative Italian study. *Br J Cancer* **72**: 244–248.

Mezzetti M, Ferraroni M, Decarli A, La Vecchia C, Benichou J (1996). Software for attributable risk and confidence interval estimation in case–control studies. *Comput Biomed Res* **29**: 63–75.

Mezzetti M, La Vecchia C, Decarli A, *et al.* (1998). Population attributable risk for breast cancer: diet, nutrition, and physical exercise. *J Natl Cancer Instit* **90**: 389–394.

McTiernan A, Kuniyuki A, Yasui Y, *et al.* (2001). Comparisons of two breast cancer risk estimates in women with a family history of breast cancer. *Cancer Epidemiol Biomarkers Prev* **10**: 333–338.

Negri E, La Vecchia C, Franceschi S, *et al.* (1996). Intake of selected micronutrients and the risk of breast cancer. *Int J Cancer* **65**: 140–144.

Negri E, Braga C, La Vecchia C, Franceschi S, Parazzini F (1997). Family history of cancer and risk of breast cancer. *Int J Cancer* **72**: 735–738.

Parazzini F, Braga C, La Vecchia C, *et al.* (1997). Hysterectomy, oophorectomy in premenopause, and risk of breast cancer. *Obstet Gynecol* **90**: 453–456.

Parkin DM, Muir CS, Whelan SL, *et al.* (1993). *Cancer Incidence in Five Continents*, Vol VI. Lyon: IARC Scientific Publications.

Parkin DM, Whelan SL, Ferlay J, Raymond L, Young J (1997). *Cancer Incidence in Five Continents*, Vol VII. Lyon: IARC Scientific Publications.

Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB. (2002). *Cancer Incidence in Five Continents*, Vol VIII. Lyon: IARC Scientific Publications.

Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA (2001). Validation of the Gail *et al.* model of breast cancer risk prediction and implications for chemoprevention. *J Natl Cancer Inst* **93**: 358–66.

Salvini S, Gnagnarella P, Parpinel M, *et al.* (1996). The food composition database for an Italian food frequency questionnaire. *J Food Composition Analysis* **9**: 57–71.

Schrag D, Kuntz KM, Garber JE, Weeks JC (1997). Decision analysis – effects of prophylactic mastectomy and oophorectomy on life expectancy among women with BRCA1 or BRCA2 mutations [published erratum appears in *N Engl J Med* 1997 337: 434]. *N Engl J Med* **336**: 1465–1471.

Struewing JP, Hartge P, Wacholder S, *et al.* (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *N Engl J Med* **336**: 1401–1408.

Talamini R, Franceschi S, La Vecchia C, *et al.* (1996). The role of reproductive and menstrual factors in cancer of the breast before and after menopause. *Eur J Cancer* **32A**: 303–310.

Tartter PI, Gajdos C, Rosenbaum Smith S, Estabrook A, Rademaker AW (2002). The prognostic significance of Gail model risk factors for women with breast cancer. *Am J Surg* **184**: 11–15.

Thorlacius S, Struewing JP, Hartge P, *et al.* (1998). Population-based study of risk of breast cancer in carriers of BRCA2 mutation. *Lancet* **352**: 1337–1339.

Veronesi U, Maisonneuve P, Rotmensz N, *et al.* (2003). Italian randomized trial among women with hysterectomy: tamoxifen and hormone-dependent breast cancer in high-risk women. *J Natl Cancer Inst* **95**: 160–165.

Veronesi U, Maisonneuve P, Sacchini V, Rotmensz N, Boyle P, Italian Tamoxifen Study Group (2002). Tamoxifen for breast cancer among hysterectomised women. *Lancet* **359**: 1122–1124.

Waterhouse J, Muir CS, Correa P, Powell J (1977). *Cancer Incidence in Five Continents*, Vol III. Lyon: IARC Scientific Publications.

Waterhouse J, Muir CS, Shanmugaratnam K, Powell J (1987). *Cancer Incidence in Five Continents*. In: Muir CS, Waterhouse J, Mack T, Powell J, Whelan SL (editors): Vol V. Lyon: IARC Scientific Publications.

Willett W, Stampfer MJ (1986). Total energy intake: implications for epidemiologic analyses. *Am J Epidemiol* **124**: 17–27.

Wynder EL (1952a). Some practical aspects of cancer prevention. *N Engl J Med* **246**: 492–503.

Wynder RL (1952b). Some practical aspects of cancer prevention (continued). *N Engl J Med* **246**: 539–46.

Wynder EL. (1952c). Some practical aspects of cancer prevention (concluded). *N Engl J Med* **246**: 573–84.

Zanetti R, Crosignani PX, Rosso S (1997). *Cancer in Italy 1988–1992*. Roma: Il Pensiero Scientifico Editore.

AUTHOR QUERY FORM

LIPPINCOTT WILLIAMS AND WILKINS

JOURNAL NAME **CEJ** **4/23/04**

ARTICLE NO: **2004004**

QUERIES AND / OR REMARKS

Query No	Details Required	Authors Response
AQ1	Waterhouse et al., 1982 - dated 1987 in ref list?	
AQ2	Muir et al., 1987 - please supply reference.	
AQ3	Higginson and Muir 1977 - please supply reference.	
AQ4	Clarify - “We used occupational physical” - meaning?	