STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:1111–1130 (DOI: 10.1002/sim.1668)

A breast cancer prediction model incorporating familial and personal risk factors

Jonathan Tyrer, Stephen W. Duffy and Jack Cuzick*,†

Department of Epidemiology, Mathematics and Statistics, Cancer Research U.K., Wolfson Institute of Preventive Medicine, Charterhouse Square, London EC1M 6BQ, U.K.

SUMMARY

Many factors determine a woman's risk of breast cancer. Some of them are genetic and relate to family history, others are based on personal factors such as reproductive history and medical history. While many papers have concentrated on subsets of these risk factors, no papers have incorporated personal risk factors with a detailed genetic analysis. There is a need to combine these factors to provide a better overall determinant of risk. The discovery of the BRCA1 and BRCA2 genes has explained some of the genetic determinants of breast cancer risk, but these genes alone do not explain all of the familial aggregation of breast cancer. We have developed a model incorporating the BRCA genes, a low penetrance gene and personal risk factors. For an individual woman her family history is used in conjuction with Bayes theorem to iteratively produce the likelihood of her carrying any genes predisposing to breast cancer, which in turn affects her likelihood of developing breast cancer. This risk was further refined based on the woman's personal history. The model has been incorporated into a computer program that gives a personalised risk estimate. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: breast cancer; risk factors; individual risk profile; genetics; family history

1. INTRODUCTION

A number of breast cancer risk factors have been well established for many decades, notably those related to hormonal and reproductive exposures [1]. For example, nulliparity, late age at first childbirth, early age at menarche and use of oral contraceptives or hormone replacement therapy are all associated with increased risk [2–4]. Additionally, the association of a family history of breast cancer with increased risk of the disease has long been established [5, 6]. Possibly partly due to increased awareness of family history, and continuing partly due to the establishment of the high risk, high penetrance BRCA1 and BRCA2 germline mutations [7], family history has taken on a considerable degree of importance both as a research issue and as a health concern in the public at large. Referrals of well women concerned about

*Correspondence to: Jack Cuzick, Department of Epidemiology, Mathematics and Statistics, Cancer Research U.K., Wolfson Institute of Preventive Medicine, Charterhouse Square, London EC1M 6BQ, U.K.
†E-mail: jonathan.tyrer@cancer.org.uk

Received June 2002
Accepted July 2003

their family history of breast cancer to family history clinics are common. Other risk factors pertaining to personal medical history have recently been established. These include diagnosis of atypical ductal hyperplasia and lobular carcinoma *in situ* [8].

There have been many studies detailing the relative risks of breast cancer based on these factors. Prediction of individual absolute risks as determined by such factors is less common. There is a considerable need for such prediction for purposes of management of women referred to genetic clinics and for considerations of eligibility of any prophylactic interventions in both the clinical and research settings.

Investigations to date have tended to be confined to subsets of the risk factors. In 1989 Gail *et al*. discussed the risk based on many factors including a broad classification of family history, but did not directly calculate the risk of having an adverse genotype. Other studies have investigated the genetic risks of breast cancer without taking hormonal and reproductive characteristics into account [9, 10]. This paper describes a method of combining these two approaches, by incorporating both familial and personal characteristics.

Ideally to build up a model of risk of breast cancer, one would prefer to have substantial numbers of subjects prospectively tested for BRCA1 and BRCA2 genes, all with other risk factor data, or at least case control data on an unselected group of carriers. Such data is not available and we have relied on a range of different published data sets to build up the model and employed segregation analysis techniques based on Bayes theorem to predict genetic risk from family history factors. The use of large scale published work also reduces sampling error in the prediction.

2. GENERAL APPROACH TO GENETIC RISK

For several years, it has been known that BRCA1 and BRCA2 are high risk germline mutations for breast cancer. However these high penetrance genes can only account for at most 5 per cent of the cases of breast cancer and other lower risk ‘susceptibility’ genes must be present to explain the observed familial aggregation of breast cancer. It seems likely that many genes influence breast cancer risk, but it is not feasible to create a model that could incorporate all these genes, even if they were known.

There are many possible ways in which we could model the risk seen from family history. There has not been convincing evidence for shared environmental factors, although the generation could be significant as breast cancer rates have been increasing in modern times. This factor could account in part for the higher relative risks of sisters than mothers observed in some case-control studies. On the whole however, the relative risk for mothers and sisters have been found to be similar, which is why we did not use a recessive model, which would cause the relative risk for the sister to be much higher. We also tested an approximation of the polygenic model (the hypergeometric polygenic model [11] which gave a similar fit to the dominant model we eventually used. The dominant model was eventually chosen for reasons of simplicity and because it did not overestimate the risk caused by two affected first-degree relatives.

The approach taken was to use a two locus genetic model with one of the genes based on BRCA1 and BRCA2, and the other (the low penetrance gene) chosen to give as good a fit as possible to the observed risk from family history. The ‘low penetrance’ gene was assumed to be dominant, so that the risks associated with this gene were the same for women with one or two copies. To do this we used information from the paper by Anderson *et al*. [12], in which the authors followed-up daughters of breast cancer patients. We thus created a hypothetical gene to act as a surrogate for the effect of all the other ‘unknown’ genes. This additional gene could be considered analogous to curve fitting between points. In other words, it is not necessarily believed that a single gene accounts for the familial risk not associated with BRCA1 and BRCA2, but that the model of a single remaining gene has the potential to capture any residual effects as observed in relation to family history.

Of course the proportion with a given genotype is only one ingredient in the risk model, the other being the effect of the genotype on breast cancer risk. The absolute effect depends heavily on age and can be modelled using survival analysis. None of the typical parametric survival distributions (Weibull, log-logistic, etc.) fit the association between age and breast cancer incidence. However national statistics for breast cancer can be used to determine the basic survival function. Following Antoniou *et al.* [13], BRCA1 and BRCA2 were coded into a single locus with three alleles: BRCA1 positive, BRCA2 positive and a normal allele. This gave a good approximation because the probabilities of a BRCA mutation were very small. The age specific incidences for breast and ovarian cancer for the BRCA1 and BRCA2 genes given by Narod *et al.* [14] and Ford *et al.* [7] were used. For the hypothetical gene a proportional hazards model was assumed, with the risks based around the basic survival function. It is important to ensure that the total hazard rate for the population matches the average estimated risks. The details of this are provided in Section 7.

3. METHODS

The genetic part of the model is that there are two autosomal loci which contain genes predisposing to breast cancer. The first locus contains information about the BRCA genes and may either contain the normal allele, a BRCA1 allele or a BRCA2 allele. The second locus contains a hypothetical susceptibility gene (the ‘low penetrance gene’) which causes an increase in the relative hazard of breast cancer. This low penetrance gene is dominant so that a woman with two copies will have the same phenotype as a person with one copy. Thus the phenotype of a person can be modelled by

$$\text{phenotype} = \begin{pmatrix} \text{no} \\ \text{BRCA1} \\ \text{BRCA2} \end{pmatrix} \begin{pmatrix} \text{no} \\ \text{yes} \end{pmatrix}$$

where the first column is the BRCA locus and the second column is the low penetrance gene locus.

BRCA1 and BRCA2 were assumed to be at the same locus to simplify the calculations. For a woman with a copy of a BRCA1 and BRCA2 gene the woman was assumed to have a BRCA1 phenotype. As this was a very rare event this assumption is not very important. Thus the possible states for the phenotypes at this locus are

No BRCA gene = no BRCA gene.
BRCA1 phenotype = at least one BRCA1 gene.
BRCA2 phenotype = at least one BRCA2 gene with no BRCA1 gene.

Thus there are six possible phenotypes comprising three possible BRCA phenotypes by two possible low penetrance phenotypes.

Table I. The possible phenotypes and their risks of developing breast cancer.

Phenotype	Description	Probability of getting breast cancer from ages t_i to t_j
1	No BRCA gene, no low penetrance gene	$S_0(t_i) - S_0(t_j)$
2	No BRCA gene, at least one low penetrance gene	$S_0(t_i)^\theta - S_0(t_j)^\theta$
3	BRCA1 gene, no low penetrance gene	$S_1(t_i) - S_1(t_j)$
4	BRCA1 gene, at least one low penetrance gene	$S_1(t_i)^\theta - S_1(t_j)^\theta$
5	BRCA2 gene, no low penetrance gene	$S_2(t_i) - S_2(t_j)$
6	BRCA2 gene, at least one low penetrance gene	$S_2(t_i)^\theta - S_2(t_j)^\theta$

For each phenotype this is a risk distribution for getting breast cancer by a certain age. How this risk is calculated is discussed in Section 7. Table I gives the possible phenotypes and expressions for the risks of getting breast cancer where S_i is the base survivor function for the BRCA genotype calculated in Section 7 and θ is the relative hazard caused by the low penetrance gene.

For a woman the risk of developing breast cancer between ages t_1 and t_2 is given by

$$\Pr(\text{cancer}) = 1 - \left(1 - \sum_{i=1}^{6} p_i F_i(t_1, t_2)\right)^{\alpha}$$

The terms in this expression and how they are calculated are as follows.

p_i is the probability of the woman having the relevant phenotype in Table I. The family history of the woman is used to calculate the distribution of her genotype probabilities and from this the phenotypic probabilities are calculated. This is discussed further in Section 8.

$F_i(t_1, t_2)$ is the probability of getting breast cancer between ages t_1 and t_2 given the woman's phenotype i. The formulae for this expression are given in Table I.

α is the relative risk due to personal factors and is calculated by the method of Sections 13 and 14.

4. ESTIMATION STRATEGY

The approach was in two parts. First the model needed to be developed to fit the observed risks from personal and familial factors as accurately as possible. After this was achieved risks for individual women could be determined. Figure 1 gives a schematic diagram of the approach taken to develop the model.

To determine the optimum frequency and relative hazard of the hypothetical gene in step 4 the approach given in Figure 2 was taken.

These were the steps taken to develop the model. Once the model has been determined, it will be used to calculate the risk for women wishing to find out their personal risk. The steps given in Figure 3 are used.

The rest of the paper describes these steps in more detail.

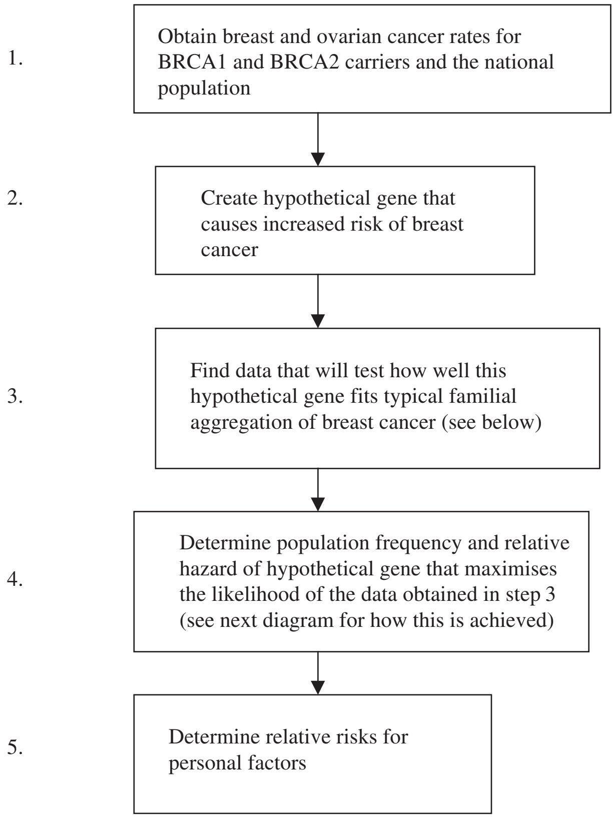
1.
Obtain breast and ovarian cancer rates for BRCA1 and BRCA2 carriers and the national population
2.
Create hypothetical gene that causes increased risk of breast cancer
3.
Find data that will test how well this hypothetical gene fits typical familial aggregation of breast cancer (see below)
4.
Determine population frequency and relative hazard of hypothetical gene that maximises the likelihood of the data obtained in step 3 (see next diagram for how this is achieved)
5.
Determine relative risks for personal factors

Figure 1. Steps taken to develop the model.

5. THE RISK FOR THE NATIONAL POPULATION

To calculate the risk of developing breast cancer by age, we used incidence rates in the general population taken from U.K. national statistics [15]. Any model that is used needs to match the average population risk of developing breast cancer to national breast cancer risks. To do this a two stage process was used.

5.1. Using the BRCA genes

The risks for 10 year periods have been published for the BRCA1 and BRCA2 gene mutations in Ford *et al.* [7] and are shown in Table II. The population frequencies of carriers of these genes were estimated to be 0.11 and 0.12 per cent for BRCA1 and BRCA2, respectively, in Peto *et al.* [16].

Information about England and Wales breast cancer risks is given in Reference [15] and is summarized in Table III.

The BRCA data were given in 10 yearly periods rather than the five yearly periods of the national data. To correspond with the national data, each BRCA period was split into the two periods proportionally relative to the national incidence data.

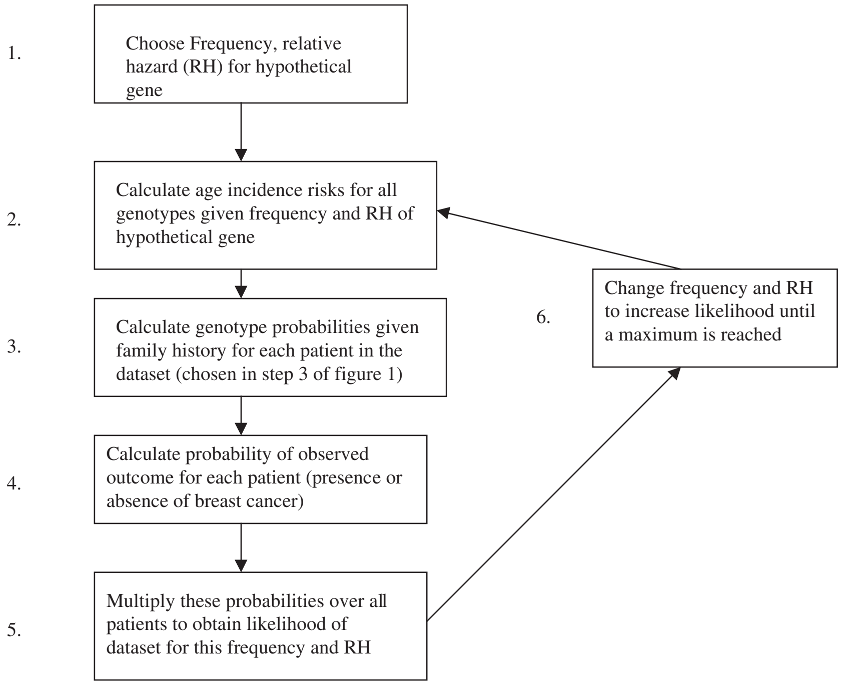
1.
Choose Frequency, relative hazard (RH) for hypothetical gene
2.
Calculate age incidence risks for all genotypes given frequency and RH of hypothetical gene
3.
Calculate genotype probabilities given family history for each patient in the dataset (chosen in step 3 of figure 1)
4.
Calculate probability of observed outcome for each patient (presence or absence of breast cancer)
5.
Multiply these probabilities over all patients to obtain likelihood of dataset for this frequency and RH
6.
Change frequency and RH to increase likelihood until a maximum is reached

Figure 2. How the frequency and relative hazard for the hypothetical gene was chosen.

For example the incidence of breast cancer for years 40−44 was 0.0056, 45−49 was 0.0090 and the probability (not incidence) of a BRCA1 gene carrier getting breast cancer between ages 40−49 was 0.31 (calculated by 0.49−0.18 from Table II). Thus the probability of getting breast cancer between ages 40−44 for a BRCA1 carrier was estimated at $0.31 * 0.0056/(0.0056 + 0.0090)$ and between ages 45−49 at $0.31 * 0.0090/(0.0056 + 0.0090)$.

Once the risk for the BRCA carriers had been calculated, the risk for a non-BRCA carrier was computed so that the population risk would be the same as the national population using the frequencies given above.

5.2. *Ovarian cancer risks for BRCA carriers*

It is known that BRCA carriers [8, 14] also have a predisposition to developing ovarian cancer compared to national rates [15]. If we assume that the development of breast and ovarian cancer is independent given genotype then both of these can be used to determine the risk of being a BRCA carrier. The risks of ovarian cancer for these cases are given in Table IV.

6. MODELLING THE RISK WITH RESPECT TO FAMILY HISTORY

To be able to predict breast cancer risk accurately, we need reliable data that gives an indication of risk on the basis of family history. To do this we looked for papers giving the

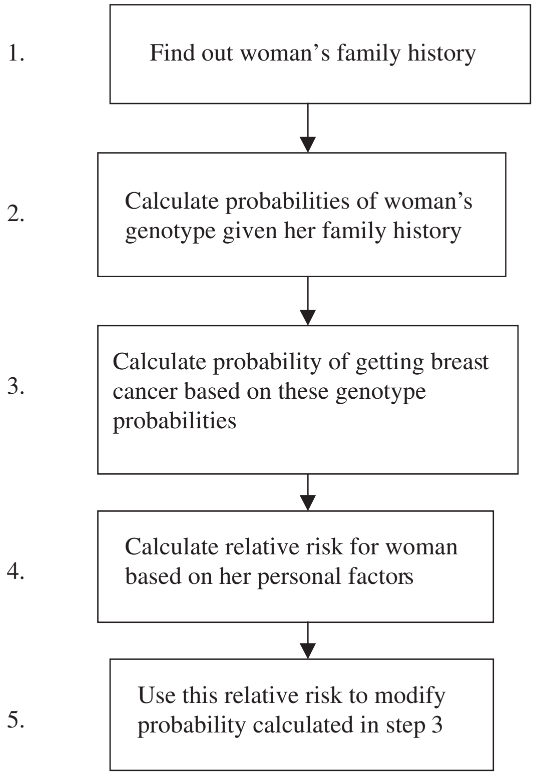
1.
Find out woman's family history
2.
Calculate probabilities of woman's genotype given her family history
3.
Calculate probability of getting breast cancer based on these genotype probabilities
4.
Calculate relative risk for woman based on her personal factors
5.
Use this relative risk to modify probability calculated in step 3

Figure 3. Calculating an individual woman's risk of developing breast cancer.

Table II. Cumulative risk of breast cancer for BRCA1 and BRCA2 carriers for different ages.

Age (yr)	Cumulative risk from BRCA1 gene $(1 - S_1(t))$	Cumulative risk from BRCA2 gene $(1 - S_2(t))$
30	0.036	0.006
40	0.18	0.12
50	0.49	0.28
60	0.64	0.48
70	0.71	0.84

relative risks of breast cancer for different types of family history. This was a difficult task, as individualized family histories for each person did not tend to be presented in most papers. A wide range of family histories with 0, 1, 2 or more relatives affected would be ideal, as this would give greater power in discriminating between the parameters in the model.

The paper by Anderson *et al.* [12] describes a study in which daughters of mothers who had breast cancer were followed up. This paper was chosen because the number of people

Table III. U.K. population incidence of breast cancer in 5-year age groups [15].

Age	Breast cancer incidence in in 5 year period
20–24	0.00006
25–29	0.0004
30–34	0.00133
35–39	0.003035
40–44	0.00563
45–49	0.009015
50–54	0.01221
55–59	0.012825
60–64	0.013855
65–69	0.012215
70–74	0.014155
75–79	0.016435
80–84	0.01786

Table IV. Ovarian cancer risks for BRCA carriers and comparison with national data.

Age	Cumulative risk from BRCA1 gene	Cumulative risk from BRCA2 gene	Cumulative risk for the population
30	0.001	0.00036	0.000205
40	0.005	0.00126	0.00072
50	0.16	0.004	0.00228
60	0.30	0.074	0.00587
70	0.42	0.27	0.01133

Table V. The incidence of breast cancer in Anderson *et al.* [12].

	Breast cancer at follow-up age 0–39		Breast cancer at follow-up age 40–53	
Mother's age (years) at breast cancer	Observed	Expected	Observed	Expected
$\leqslant 39$	19	3.30	9	2.33
40–49	69	23.34	59	29.69
50–59	89	39.17	129	61.94
$\geqslant 60$	131	76.21	296	169.72

observed in the study was high, giving estimates with a relatively low standard error. The data reported forms Table V.

The number of patients followed-up was not given, however this can be estimated by the equation

$$\text{Expected number} = \text{proportion expected from population} \times \text{Number of patients}$$

This implies that number of patients = expected number/proportion expected.

6.1. How well does the model fit?

For a set of candidate parameters (see next section) the expected risk can be calculated for the entries in Table V based on both the mother's and daughter's age. Define group ij to be based on Table V so that group 32 refers to the follow-up age from 40–53 for someone whose mother was affected at age 50–59. Define p_{ij}, obs_{ij} and exp_{ij} to be the risks, observed entries and expected entries for group ij. For example p_{32} is the risk from ages 40–53 for someone whose mother was diagnosed at age 50–59. Also define pop_1 to be the population risk of getting breast cancer from age 0–39 and similarly pop_2 to be the population risk from age 40–53. Define n_{ij} to be the estimate of the number at risk so that $n_{ij} = \text{exp}_{ij}/\text{pop}_j$. Then the likelihood for the data is

$$L = \prod_{ij} p_{ij}^{\text{obs}_{ij}}(1 - p_{ij})^{(n_{ij} - \text{obs}_{ij})}$$

It is easier to work with the log-likelihood which is

$$L = \sum \text{obs}_{ij} \log(p_{ij}) + (n_{ij} - \text{obs}_{ij}) \log(1 - p_{ij}) \qquad (1)$$

The maximization of this expression will be discussed later.

7. CALCULATING THE RISKS FOR A PARTICULAR AGE AND GENOTYPE

A proportional hazards model to the development of breast cancer was taken for the hypothetical susceptibility gene. It was assumed that this gene increased the risk of breast cancer by the same relative rate for women with and without a BRCA gene, but that it had no influence on ovarian cancer risk. It was also assumed that the susceptibility gene was dominant, so that the risks for a heterozygous woman would be the same as for a homozygous woman with the susceptibility gene.

In the section that follows we refer to the genotype of a woman as her genotype at the risk gene locus (i.e. we ignore her genotype at the BRCA locus). We also define the term 'absolute risk' to mean the chance of getting breast cancer by a certain age, ignoring competing causes of mortality.

We defined the baseline hazard and survivor functions for women with no copies of the risk gene. Define $\lambda(t)$ to be the baseline hazard function and $S(t)$ to be the baseline survivor function. The hazard for a woman with genotype g is defined to be $\theta_g\lambda(t)$ where θ_g is the relative rate for a woman with genotype g compared to someone with no copies of the risk gene. The survivor function for this woman would then be $S(t)^{\theta_g}$. If we define the relative hazard caused by the risk gene as θ then $\theta_g = 1$ for someone with no copies and $\theta_g = \theta$ someone with at least one copy of this risk gene.

To calculate the absolute risks over age for the different genotypes it is necessary to calculate either the baseline hazard or survivor functions. The approach taken was to calculate the survivor function $S(t)$, as equations in $S(t)$ are easier to solve. Assuming no interaction between these genes and the BRCA status, we calculate the base survivor functions separately for women with no BRCA genes, a BRCA1 gene or a BRCA2 gene. We define these survivor functions as S_0, S_1 and S_2, respectively. This procedure is illustrated for women without a BRCA gene. For women with a BRCA gene, the survivor function $S_{\text{non}}(t)$ was replaced with the appropriate BRCA survivor function.

First we estimate the survivor function. Define $S_{\text{non}}(t)$, $S_{\text{pop}}(t)$ as the survivor functions for women with no BRCA genes and for the population, respectively. Then $S_{\text{non}}(t) = (S_{\text{pop}}(t) - p_1 S_{\text{BRCA1}}(t) - p_2 S_{\text{BRCA2}}(t))/(1 - p_1 - p_2)$, where p_1 and p_2 are the proportion of BRCA1 and BRCA2 carriers in the population.

We define p_g to be the proportion of the population with the g genotype and choose the baseline survivor function $S_o(t)$ such that

$$\sum_g p_g S_0(t)^{\theta_g} = S_{\text{non}}(t)$$

After solving for $S_0(t)$, the risk of developing breast cancer between ages t_i and t_j for someone of genotype g would be calculated as $S_0(t_i)^{\theta_g} - S_0(t_j)^{\theta_g}$. Thus for someone with no copies of the risk gene the risk would be $S_0(t_i) - S_0(t_j)$ and for someone with at least one copy the risk would be $S_0(t_i)^{\theta} - S_0(t_j)^{\theta}$.

For BRCA1 carriers the risk of developing breast cancer between ages t_i and t_j for someone of genotype g is $S_1(t_i)^{\theta_g} - S_1(t_j)^{\theta_g}$ after solving for $S_1(t)$, and similarly for BRCA2 carriers the risk is $S_2(t_i)^{\theta_g} - S_2(t_j)^{\theta_g}$.

This equation was calculated for each of the five yearly periods. However there exists no analytic solution and so the equation must be solved numerically. This was done using a Newton–Raphson method [17].

8. CALCULATION OF GENETIC PROBABILITIES

The approach taken was similar to the segregation analysis method discussed in Ott [18] and Parmigiani *et al.* [10]. Bayes theorem was used to calculate the genetic probability based on the family history and estimated population frequency of genes.

8.1. Calculating the probabilities

The treatment follows that of Ott [18]. An example pedigree is shown and the method used can be generalized to other pedigrees.

Consider for example the pedigree in Figure 4.

For this pedigree assign a number $i = 1, 2, \ldots, m$ to each individual. Let x_i denote the phenotype of individual i. For our case the phenotype denotes the person's age, sex and presence or absence of breast cancer and ovarian cancer. Denote by $P(g_i)$ the probability that individual i has genotype g_i given the population gene frequencies, $P(x_i \mid g_i)$ to be the probability of individual i's phenotype given the genotype g_i and $P(g_i \mid g_j g_k)$ the probability that individual i has genotype g_i given parental genotypes g_j and g_k.

For our particular model $P(x_i \mid g_i)$ has already been calculated by the method of Section 7. $P(g_i)$ is calculated using the population frequencies for the genes and assuming Hardy–Weinburg equilibrium and $P(g_i \mid g_j g_k)$ is calculated under the assumption that the genes are unlinked and autosomal.

Define $P(x_i \mid \mathbf{g_i})$ to be a vector of probabilities for the phenotype of i for each possible genotype. Similarly define $P(\mathbf{g_i} \mid g_j g_k)$ to be the vector of probabilities for each genotype given parental genotypes g_j and g_k and also $P(x_i \mid \mathbf{g_i}) P(\mathbf{g_i} \mid g_j g_k)$ to be the vector with individual entries $P(x_i \mid g_{i_n}) P(g_{i_n} \mid g_j g_k)$ where g_{i_n} is the nth possible genotype for individual i.

To calculate the likelihood of the whole pedigree [18, 19].

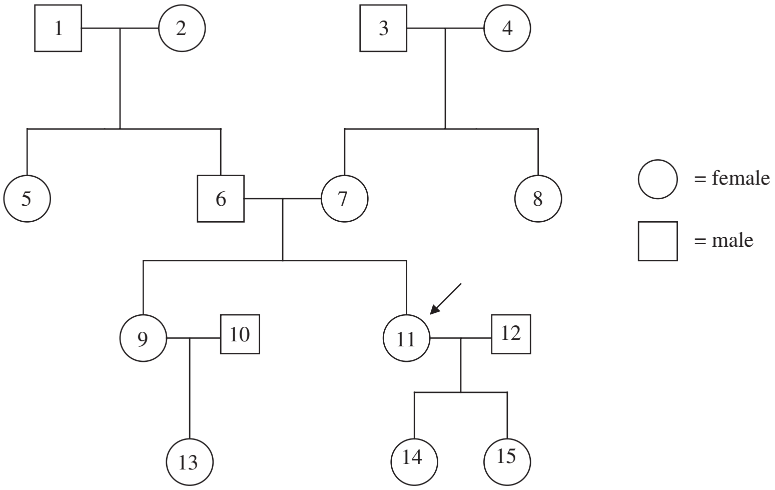

Figure 4. An example pedigree.

$L = \sum_{g_1} \cdots \sum_{g_m} \prod_{i=1}^{m} P(x_i \mid g_i) P(g_i \mid ..)$ where $P(g_i \mid ..)$ is either the probability that individual i has genotype g_i given the parental genotypes or the population gene frequencies if the parents aren't known.

To calculate the expression we need to derive the probabilities for subsections of the pedigree and substitute the probabilities for these subsections.

For example consider individuals 11, 12, 14 and 15 in the above pedigree.

We want to calculate probabilities conditional on the genotypes of individual 11.

$$P(x_{11}, x_{12}, x_{14}, x_{15} \mid g_{11}) = P(x_{11} \mid g_{11}) \left[\sum_{g_{12}} P(x_{12} \mid g_{12}) P(g_{12}) \sum_{g_{14}} P(x_{14} \mid g_{14}) P(g_{14} \mid g_{11} g_{12}) \right.$$
$$\left. \times \sum_{g_{15}} P(x_{15} \mid g_{15}) P(g_{15} \mid g_{11} g_{12}) \right]$$

These probabilities can be stored to obtain a modified vector $P(x_{11}^*, g_{11})$ (see Ott [18]) with the effect that the individuals 12, 14 and 15 will not be needed for further calculations.

Consider the upper left corner of the example pedigree, with individuals 1, 2, 5 and 6.

We want to calculate the distribution of genotypes for individual 6 so that the individuals 1, 2 and 5 are accounted for in the calculations. We could calculate them individually by

$$P(x_1, x_2, x_5, x_6 \mid g_6) = \sum_{g_1} P(x_1 \mid g_1) P(g_1) \sum_{g_2} P(x_2 \mid g_2) P(g_2)$$
$$\times \sum_{g_5} P(x_5 \mid g_5) P(g_5 \mid g_1 g_2) P(x_6 \mid g_6) P(g_6 \mid g_1 g_2)$$

However as this needs to be done for each possible genotype it is quicker to use a vector-based version of this equation, which replaces the scalar g_6 with the vector $\mathbf{g_6}$ of all possible genotypes.

Thus the procedure for the example pedigree would be:

The modified probabilities of 6 are evaluated conditional on 1, 2 and 5, the probabilities of 7 conditional on 3, 4 and 8, the probabilities of 9 conditional on 10 and 13 and the probabilities of 11 conditional on 12, 14 and 15. Once this is completed the genotype probabilities of 11 are calculated conditional on the probabilities of 6, 7 and 9.

9. MAXIMIZING THE LIKELIHOOD

Using the methods of Sections 7 and 8 we can calculate the probability of getting breast cancer for the pedigree members in Section 6 for our candidate set of parameters. This will enable us to calculate the log likelihood

$$L = \sum \mathrm{obs}_{ij} \log(p_{ij}) + (n_{ij} - \mathrm{obs}_{ij}) \log(1 - p_{ij})$$

of equation (1).

We want to choose the parameters to maximize this expression. This was done by a Powell routine (see Press *et al.* [17]).

10. RESULTS OF THE GENETIC MODELLING

The parameters for the hypothetical gene that gave the best fit to the data in Anderson *et al.* [12] were:

Proportion of gene in population = 0.1139
Relative risk for at least one copy of the risk gene = 13.0377.

Figure 5 plots the cumulative risk of breast cancer against age for women based on their phenotype (excluding women carrying both a BRCA gene and the low penetrance gene).

The predicted risks from the model that includes the hypothetical gene and the BRCA genes against the risks observed in Anderson *et al.* [12] are shown in Table VI.

The observed relative risks for daughters whose mother developed breast cancer before 40 years were calculated from relatively small numbers of cases (see Table V) and so have wide confidence intervals. This helps to explain the discrepancy between the observed and predicted risks for these groups.

11. INTERPRETATION AND DISCUSSION OF RESULTS

The hypothetical gene has a high frequency with about 21 per cent of the population being a carrier. The risk for non-carriers to age 70 is only about 2 per cent while the risk to carriers is 24 per cent.

As the data in the validation exercise [12] only concerned the risk to daughters from mothers with breast cancer the difference in risk between mothers and sisters affected could not be modelled. Most studies have not been able to find any significant difference between the risks, though there does seems in general to be slightly more risk associated with an affected sister than with an affected mother of the same age. It may also be that the relative risk decreases

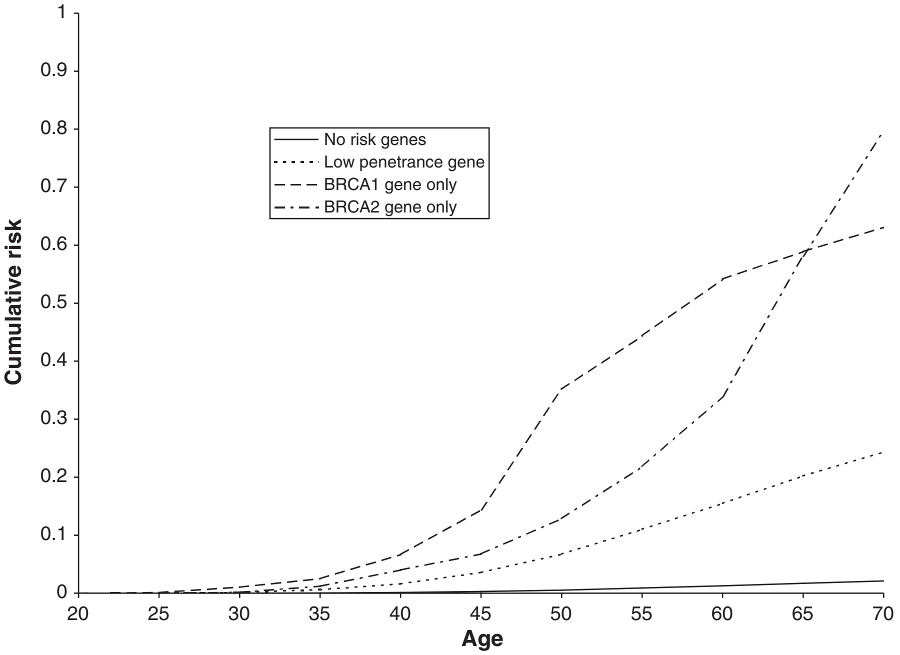

Figure 5. Cumulative risk of breast cancer by age and hypothetical gene status.

Table VI. Predicted relative risks for the hypothetical gene model compared with observed relative risks, for Anderson *et al.* [12], by subjects and maternal age at diagnosis.

	Follow-up age 0–39		Follow-up age 40–53	
Mother's age (years) at breast cancer	Predicted relative risk	Observed relative risk	Predicted relative risk	Observed relative risk
⩽39	3.62	5.76	2.25	3.86
40–49	2.48	2.96	1.98	1.99
50–59	1.94	2.27	1.82	2.08
⩾60	1.82	1.72	1.77	1.74

with age for women with a copy of the risk gene compared to women with no copies—as is the case with BRCA1 and BRCA2 and as suggested by Cui *et al.* [20]. This would help to explain the higher relative risks in the table above for women whose mother was affected below the age of 40.

12. PERSONAL FACTORS INCLUDED IN THE MODEL

There are of course other risk factors besides family history. Some of these are medical, usually concerning benign breast disease. The presence of certain benign conditions can cause

Table VII. Risk factors included in the risk calculation, with their effects as observed in the population.

Factor	Effect
Age at menarche	Risk decreases with increased age at menarche
Parity	Risk is generally less for parous women
Age at first childbirth (if parous)	Risk increases for later age at first childbirth
Age at menopause (if postmenopausal)	Risk increases for a later age at menopause
Atypical hyperplasia	A four-fold increase in risk if present
Lobular carcinoma *in situ*	An eight-fold increase in risk if present
Height	Risk increases with increased height
BMI	Risk increases for post-menopausal women with increased BMI

Table VIII. Relative risks caused by childbearing (from [6]—web Figure 1 on the lancet web site).

Age at first child (y)	Relative risk
Nulliparous	1.0
<20	0.67
20–24	0.74
25–29	0.88
$\geqslant 30$	1.04

Table IX. Relative risk caused by height (extrapolated from Reference [21]).

Height (m)	Relative risk
<1.6	1.0
1.6–1.7	$1.05 + 2 \times (\text{height} - 1.6)$
1.7+	1.24

a large increase in the chances of developing breast cancer. The other risk factors are more moderate, and include age at menarche, age at menopause, nulliparity and age at first childbirth, weight, height and HRT. The risk factors used are summarized in Tables VII–X. A proportional hazards model is assumed, and it is also assumed that risks are multiplicative.

At this stage, some risk factors have not yet been included. These include use of exogenous hormones such as HRT and ductal carcinoma *in situ*. The risk of subsequent invasive cancer associated with ductal carcinoma *in situ* is large, but it is difficult to get an accurate estimation of risk, especially with the various preventative treatments that are used when this is diagnosed (surgery, radiotherapy, etc.). It is also likely that even if not an obligate precursor of breast cancer, ductal carcinoma *in situ* represents a state of breast cancer so far advanced towards invasive carcinoma that it shares, rather than adds to, the other risk factors for breast cancer.

Table X. Relative risk caused by BMI (post-menopausal) (from Reference [21]).

BMI (kg/m^2)	Relative risk
<21	1.0
21–23	1.14
23–25	1.15
25–27	1.26
>27	1.32

We have tried to include most of the established risk factors which are readily ascertainable. HRT and oral contraceptive use involves establishing history of use and the particular preparation that was used. In addition they are risk factors of relatively recent observation. We have not therefore incorporated them into the risk model as yet.

13. THE RELATIVE RISKS USED FOR CALCULATION

The relative risks for the personal risk factors are shown below. The risk factors for childbearing were taken for women without a family history but these are similar to the risk factors for women with a family history (see Reference [6]) (see Tables VIII–X).

The relative risk associated with age at menarche was estimated to decrease by a factor of 0.95 for each year older at menarche [22].

The relative risk associated with age at menopause is estimated as an increase in relative risk by a factor of 1.028 for each year older at menopause [4].

The estimated relative risk associated with atypical hyperplasia is 4.0 (Page *et al.* [23]), and that associated with lobular carcinoma *in situ* is 8.0 (Page *et al.* [8]).

14. ADJUSTMENT TO MATCH POPULATION AVERAGE RATES

For all of the risk factors, by definition the relative risks give the ratio of risks between the presence and absence of factors. However, we need the risk relative to the general population to be able to calculate probabilities. To do this, we need the population frequencies for the categories of the risk factors.

Suppose a given factor can take forms $1,2,\ldots,n$. Define the relative risks for the factors by $f_1, f_2, \ldots, f_n$ (normally one of these will be one) and the population frequencies by $p_1, p_2, \ldots, p_n$. Then the risk for an average person will be $\sum_i p_i f_i$ and so the relative risk for a person with form j compared to the population is $f_j/\sum_i p_i f_i$. Table XI gives estimates of the population risk from our risk factors.

The relative risks given earlier should be divided by the relevant population risk to get a risk relative to the population risk.

The average population risk for atypical hyperplasia and lobular carcinoma *in situ* was not calculated, as these were rare diagnoses. However their detection is becoming more common with the advent of population screening and it may be more appropriate to decrease the risk

Table XI. Average population risks relative to the baseline risk.

Factor	Average population risk relative to baseline risk	Baseline group
Age at menarche	0.99	Menarche at age 13
Age at first child	0.78	Nulliparous
Height	1.1	<1.6 m
BMI (post-menopausal)	1.17	< 21 kg/m^2

Table XII. Frequency distribution of age at menopause.

Age of menopause	Probability
<35	0.0117
35–39	0.0302
40–44	0.0966
45–49	0.2757
50–54	0.4628
$\geqslant 55$	0.1230

for women without these conditions, particularly for older women where their prevalence is more common.

For menopausal information the calculation is slightly different. If a woman is post-menopausal then the calculation is as above. However if the woman is premenopausal then the average rate is calculated given that the person is not post-menopausal by her current age. If the woman is perimenopausal then the age at menopause is taken to be her current age. The population probability distribution used for the age at menopause is shown in Table XII.

The overall risk for a person compared to the population is the product of her risks for each individual risk factor, relative to the population risk for each factor.

15. COMBINING GENETIC AND NON-GENETIC RISKS

To calculate the risk for a person based on her family history and other factors a simple procedure was used.

1. Calculate the absolute risk at current age based purely on the family history.
2. Calculate the relative risk for the person based on her personal factors.
3. Use this relative risk to alter the risk calculated in step 1. The new risk is calculated by the formula
   (a) final risk $= 1 - (1 - \text{risk based on family history})^{\text{relative hazard}}$.
   This equation follows from using the proportional hazards model.

An alternative approach is to calculate the risk for a person over all possible genotypes given her personal risk factors and use these calculations in working out the risk based on the family history. This superficially is more in keeping with the model's assumptions. However this method tends to decrease the effect of the personal risk factors (small risks become larger and large risks become smaller), because the change in high-risk individuals has less effect (they are quite likely to get breast cancer whatever their personal factors). This follows from the fact that $\sum_i S_i^\lambda > (\sum_i S_i)^\lambda$ when $\lambda > 1$ and $S_i > 0$ for all i, where the S_i could represent the survival probabilities for the different genotypes. If this approach was taken then the parameters would have to be increased so that they gave the same average change in relative risk for a woman at a hypothetical age. This approach relies on a lot of assumptions and would not make much difference in most cases where the change in relative risk is quite small. This approach also would not be suitable for LCIS or atypical hyperplasia as these cases probably indicate an underlying genetic susceptibility to breast cancer.

16. HOW THE RISK FACTORS CAN CHANGE RISK

To give an idea of how various factors can influence risk we consider a woman whose risk factors gradually increase.

1. A 30 year old woman whose menarche was at age 16, had her first child at age 22 and is of height 1.55 m (her weight is not relevant as she is premenopausal). Her family history consists of a mother and two grandmothers all of age 68 and all unaffected with breast or ovarian cancer.
2. A 30 year old woman whose menarche was at age 11, is nulliparous and is of height 1.75 m. Her family history is the same as woman 1.
3. A 30 year old woman whose personal factors and family history are the same as woman 2 excepting that her maternal grandmother had breast cancer at age 68.
4. A 30 year old woman whose personal factors and family history are the same as woman 2 excepting that her paternal grandmother had breast cancer at age 68.

The risk for woman 3 is less than that of woman 4 as her genes have been passed through the mother's side and the mother is unaffected with breast cancer.

Figure 6 plots the cumulative risk of breast cancer for these women as well as the population risk.

17. DISCUSSION AND DEMONSTRATION

The above develops a model for predicting individual risk of breast cancer from both familial and non-familial risk factors for the disease. The purpose of the paper is to demonstrate both the steps in the process of development of the predictive model and of its use. The model has been computerized and an interactive program is available on request.

For use of the method, consider the following examples:

1. A woman aged 45, premenopausal, with one child born when she was 26, whose mother had breast cancer at age 52. With no information about other family members, her age at menarche, height or weight, her 10-year risk is estimated as 5.0 per cent.
2. A woman aged 50, menopausal at age 48, nulliparous, menarche at age 12, height 1.6 m, weight 55 kg, with no family history of breast cancer but with a history of atypical ductal

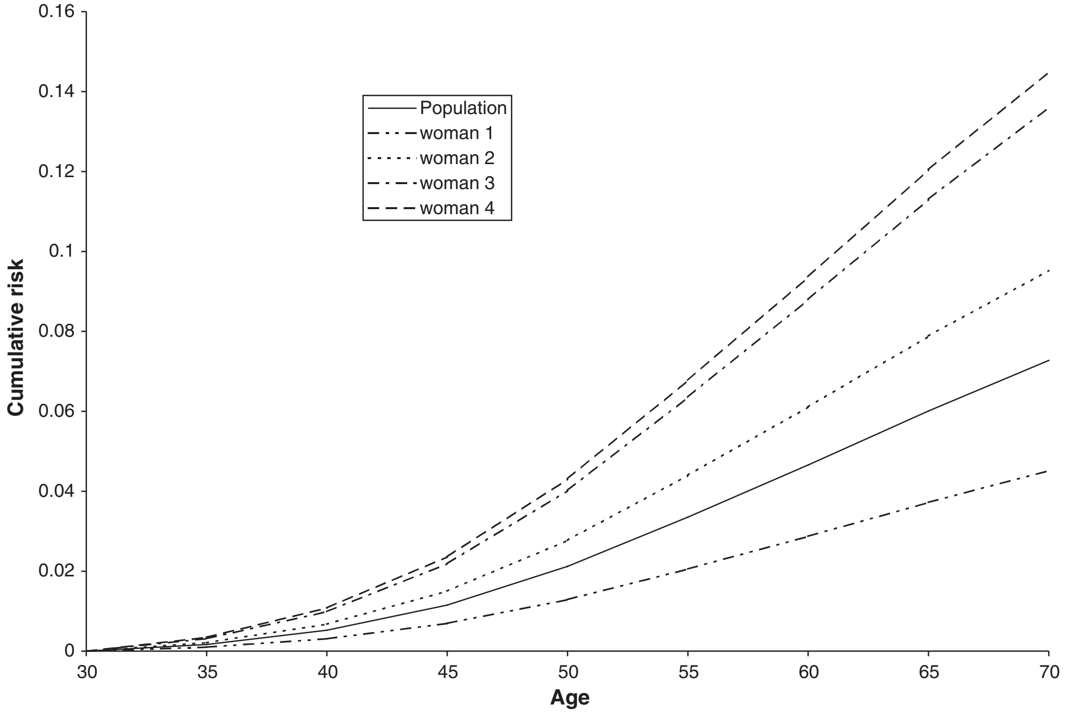

Figure 6. Cumulative risk of breast cancer for some example women.

hyperplasia. Assuming an age of 70 for her mother and two grandmothers her 10-year risk is estimated as 9.8 per cent.

3. A woman aged 35, with a sister who had breast cancer at 40, a mother who had breast cancer at 50 and a maternal grandmother who had breast cancer at 55. The paternal grandmother lived to age 75 and did not develop breast cancer. With no information about her reproductive history, height and weight her 10-year risk is estimated as 3.5 per cent.

It is easy to add new factors or to change the parameters that are used in the model. In particular, it would have been beneficial to have more data available on family history, in particular relative risks associated with two or more relatives with breast cancer. This type of data would give greater power to discriminate between possible parameters. It is important when collecting such data to take account of all relatives whether or not they are affected with breast cancer. For example, two affected sisters out of three might be more indicative of a genotype predisposing to cancer than two out of five.

There are two features of this model and the corresponding computer program, which contribute an advance on previous work. First the familial risk estimation uses not only segregation analysis based on the existence of the known BRCA1 and BRCA2 mutations, but also on an unknown predisposing gene. This is important because it is clear that the BRCA1 and BRCA2 mutations alone cannot account for the increased risk associated with family history of breast cancer. The second important feature is the incorporation of the non-familial, hormonal and clinical risk factors.

Our final model is not perfect. Alternative models and strategies for arriving at the model parameters could be suggested. It could be argued that our risk estimates associated with BRCA1 and BRCA2 positive status are too high [24, 25], and that the effect of factors such as age at first birth may vary according to BRCA1 and BRCA2 status [26]. However, using the other BRCA estimates gave a less good fit to our example data set and also the higher risks will at least give better risk estimates for women from high risk families.

The program is likely to be a useful tool, both in the clinical setting, for establishing the risk of the 'worried well', and the research setting. The latter may include determining risk criteria for inclusion in prevention studies, and prediction of incidence in single-arm early detection programs [27]. The two new features render it a potentially valuable advance on the current technology.

REFERENCES

1. MacMahon B, Cole P, Brown JB, Aoki K, Lin TM, Morgan RW, Woo N. Oestrogen profiles of Asian and North American women. *Lancet* 1971; **2**(7730):900–902.
2. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* 1989; **81**:1879–1886.
3. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. *Lancet* 1996; **347**(9017):1713–1727.
4. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet* 1997; **350**(9084):1047–1059.
5. Thompson WD. Genetic epidemiology of breast cancer. *Cancer* 1994; **74**:279–287.
6. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet* 2001; **358**(9291):1389–1399.
7. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, Bishop DT, Weber B, Lenoir G, Chang-Claude J, Sobol H, Teare MD, Struewing J, Arason A, Scherneck S, Peto J, Rebbeck TR, Tonin P, Neuhausen S, Barkardottir R, Eyfjord J, Lynch H, Ponder BA, Gayther SA, Zelada-Hedman M *et al*. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *American Journal of Human Genetics* 1998; **62**:676–689.
8. Page DL, Kidd Jr TE, Dupont WD, Simpson JF, Rogers LW. Lobular neoplasia of the breast: higher risk for subsequent invasive cancer predicted by more extensive disease. *Human Pathology* 1991; **22**(12):1232–1239.
9. Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer: implications for risk prediction. *Cancer* 1994; **73**:643–651.
10. Parmigiani G, Berry DA, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *American Journal of Human Genetics* 1998; **62**:145–158.
11. Lange K. *Mathematical and Statistical Methods for Genetic Analysis*. Springer: New York, Berlin, Heidelberg, 1997;131–134.
12. Anderson H, Bladstrom A, Olsson H, Moller TR. Familial breast and ovarian cancer: a Swedish population-based register study. *American Journal of Epidemiology* 2000; **152**(12):1154–1163.
13. Antoniou AC, Pharoah PD, McMullan G, Day NE, Ponder BA, Easton D. Evidence for further breast cancer susceptibility genes in addition to BRCA1 and BRCA2 in a population-based study. *Genetic Epidemiology* 2001; **21**(1):1–18.
14. Narod SA, Ford D, Devilee P, Barkardottir RB, Lynch HT, Smith SA, Ponder BA, Weber BL, Garber JE, Birch JM *et al.* An evaluation of genetic heterogeneity in 145 breast-ovarian cancer families. Breast Cancer Linkage Consortium. *American Journal of Human Genetics* 1995; **56**(1):254–264.
15. Cancer statistics registrations, The Stationary office Series MB1 no. 27, England and Wales, 1994.
16. Peto J, Collins N, Barfoot R, Seal S, Warren W, Rahman N, Easton DF, Evans C, Deacon J, Stratton MR. Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. *Journal of the National Cancer Institute* 1999; **91**(11):943–949.

17. Press WH, Teukolsky SA, Vettering WT, Flannery BP. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press: Cambridge, 1992.
18. Ott J. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *American Journal of Human Genetics* 1974; **26**(5):588–597.
19. Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. *Human Heredity* 1971; **21**(6):523–542.
20. Cui J, Antoniou AC, Dite GS, Southey MC, Venter DJ, Easton DF, Giles GG, McCredie MR, Hopper JL. After BRCA1 and BRCA2-what next? Multifactorial segregation analyses of three-generation, population-based Australian families affected by female breast cancer. *American Journal of Human Genetics* 2001; **68**(2): 420–431.
21. van den Brandt PA, Spiegelman D, Yaun SS, Adami HO, Beeson L, Folsom AR, Fraser G, Goldbohm RA, Graham S, Kushi L, Marshall JR, Miller AB, Rohan T, Smith-Warner SA, Speizer FE, Willett WC, Wolk A, Hunter DJ. Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk. *American Journal of Epidemiology* 2000; **152**(6):514–527.
22. Robbins AS, Brescianini S, Kelsey JL. Regional differences in known risk factors and the higher incidence of breast cancer in San Francisco. *Journal of the National Cancer Institute* 1997; **89**(13):960–965.
23. Page DL, Dupont WD. Anatomic indicators (histologic and cytologic) of increased breast cancer risk. *Breast Cancer Research and Treatment* 1993; **28**(2):157–166.
24. Hopper JL, Southey MC, Dite GS, Jolley DJ, Giles GG, McCredie MR, Easton DF, Venter DJ. Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in BRCA1 and BRCA2. Australian Breast Cancer Family Study. *Cancer Epidemiology Biomarkers & Prevention* 1999; **8**(9):741–747.
25. Struewing JP, Hartge P, Wacholder S, Baker SM, Berlin M, McAdams M, Timmerman MM, Brody LC, Tucker MA. The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *The New England Journal of Medicine* 1997; **336**(20):1401–1408.
26. Jernstrom H, Lerman C, Ghadirian P, Lynch HT, Weber B, Garber J, Daly M, Olopade OI, Foulkes WD, Warner E, Brunet JS, Narod SA. Pregnancy and risk of early breast cancer in carriers of BRCA1 and BRCA2. *Lancet* 1999; **354**(9193):1846–1850.
27. Mackay J, Rogers C, Fielder H, Blamey R, Macmillan D, Boggis C, Brown J, Pharoah PD, Moss S, Day NE, Myles J, Austoker J, Gray J, Cuzick J, Duffy SW. Development of a protocol for evaluation of mammographic surveillance services in women under 50 with a family history of breast cancer. *Journal of Epidemiology and Biostatistics* 2001;**6**(5):365–369.