Genetic epidemiology of early onset breast cancer

Diana Eccles, Angela Marlow, Gavin Royle, Andrew Collins, Newton E Morton

Abstract

Risks for breast cancer when there is a family history of the disease are usually calculated using data from segregation analyses which favour a single dominant gene with high penetrance. There are, however, at least three loci known to be associated with familial breast cancer (p53, BRCA1, and an as yet unpublished locus) and the frequencies and penetrances of these genes are not likely to be the same. We have attempted to address the problem of which genetic parameters should be used to calculate risks for different patterns of familial breast cancer.

Data from 384 nuclear families ascertained through a proband selected for early onset breast cancer were subjected to complex segregation analysis, correcting for ascertainment bias resulting from selection for severe phenotype. Age of onset of breast cancer, incorporated as severity, provides additional information to the segregation model over and above that given by assigning liability classes on the basis of age at observation. The use of this additional parameter in the analysis is described. There is fair agreement between estimates from this sample and previous predictions from consecutive probands and consultands. The differences suggest more than one rare dominant gene for susceptibility to breast cancer, with different penetrances. Although refinements of segregation analysis will help to delineate these different genes, perfect resolution will require identification of the mutant alleles. Methods to estimate genetic parameters under genotype specific mortality need to be developed. Meanwhile, we suggest that high and low estimates of penetrance be used in risk estimation for genetic counselling, and as a guide to candidates for entry into clinical trials of screening and chemoprevention in breast cancer.

(*J Med Genet* 1994;31:944-949)

Breast cancer affects about 1 in 12 women in the UK and 1 in 9 women in the USA, but the majority of women are affected in the peri- or postmenopausal years. It is on the basis of a higher incidence in these years that the UK National Breast Screening Programme invites women between 50 and 64 years of age for screening mammograms. A small proportion of women are affected at younger ages. Previous

segregation studies have clearly shown that rare dominant genes with high penetrance are responsible for a proportion of breast cancer cases¹⁻⁴ and that the risks to unaffected close relatives increase with decreasing age of the affected relative(s).²⁻⁴

Although the role of rare dominant genes in familial breast cancer is well established, most of the relevant genes have not been isolated and it is not yet possible to look for mutations in all the relevant genes for any individual person. Therefore segregation analysis is still required for risk assessment. Segregation analysis is a statistical method which seeks to define the mode of inheritance of a disease in a population and to define genetic parameters for diseases inherited in a mendelian fashion. A number of genetic hypotheses are considered during the analysis of a set of data and the most likely hypothesis or model is fitted. Where pedigrees are ascertained through clearly defined selection criteria, ascertainment bias can be fully corrected for⁵ and the results can be applied to the population from which the sample was taken. Such analyses for breast cancer have consistently found evidence of a dominantly inherited genetic predisposition, where the putative gene has a low frequency and a high penetrance in the population. The choice of model^{3,4} for risk calculations makes some difference to the actual risk figure arrived at but the differences are usually small. The effect of genotype specific mortality on gene frequencies at particular ages is poorly known. One would expect the population gene frequency to fall with increasing age owing to disease specific mortality. These changing gene frequencies are not incorporated into any computer program for genetic analysis, and so estimates of gene frequencies at older ages are imprecise. Nevertheless, segregation analysis still forms the basis of risk assessment which is indispensable for genetic counselling. Some sort of reproducible risk assessment should be used before admission of women to screening procedures and to clinical trials aimed at prevention, such as the tamoxifen chemoprevention study.

The working model in cancer genetics is that early onset is an indication of genetic predisposition.^{6,13} Here we address two issues with respect to early onset breast cancer.

- (1) When age of onset and genotype specific mortality are incorporated into the segregation analysis, what are the parameters that should be used for risk calculation in families of British probands where breast cancer occurs at an unusually early age?
- (2) How do these parameters compare with

CRC Research Group
in Genetic
Epidemiology,
Department of Human
Genetics, Princess
Anne Hospital,
Southampton SO9
4HA, UK
D Eccles
A Marlow
A Collins
NE Morton

Royal South Hants
Hospital, Brintons
Terrace,
Southampton, UK
G Royle

Correspondence to
Dr Eccles.

Received 4 February 1994
Revised version accepted for
publication 21 July 1994

Table 1 Population parameters used in segregation analysis

Liability class (LI)	Liability class frequencies				Age interval
	Morbid risk R	Cumulative risk I	Cumulative mortality M		
1	0.00004	0.000045	0.000008		20–24
2	0.00027	0.000345	0.000078		25–29
3	0.00102	0.001370	0.000351		30–34
4	0.00268	0.003715	0.001039		35–39
5	0.00871	0.012462	0.003787		40–49
6	0.01602	0.027067	0.011224		50–59
7	0.02396	0.045182	0.021748		60–69
8	0.03223	0.066232	0.035132		70–79
9	0.03887	0.091373	0.054624		80–85 +

R=cumulative population risk of having breast cancer dead or alive.
I=cumulative population incidence estimated from breast cancer cases in England and Wales from 1979–1982 (IARC, 1987).
M=cumulative population mortality from breast cancer in 1980 reported in the Registrar General’s Statistical Review of England and Wales.

those estimated from British probands not selected for early onset? previously, except that isolated cases were assigned to single selection.

Material and methods

One hundred and twenty eight patients were ascertained through the breast care service in three districts in Wessex. Selection of probands was made on the basis of age at breast cancer diagnosis of less than 40 years. There was no selection for positive family history at any stage. All probands were affected with breast cancer. Patients were approached at their routine follow up clinic appointment and only one patient declined interview. For the remaining 127 probands, family history including age and cause of death were documented in first and second degree relatives and wherever possible reported diagnoses of breast cancer were confirmed from medical records or death certificates. A total of 176/200 (88%) of reported breast cancers were confirmed from records or death certificates. The remainder were reported by a first degree relative to have had a mastectomy or radiotherapy or both. The age at last observation for all family members was recorded, and age at onset of breast cancer, where known, was also documented. Nuclear families in this dataset were coded for single selection, which in the COMDS computer program is conditional on severity.⁷ In two pedigrees an affected third degree relative was identified and coded for truncate selection as a secondary proband within their nuclear family with the proband as the pointer. This dataset will be called the Wessex sample.

Iselius *et al*² reported on two British series (one selected through consecutive probands with breast cancer and the other through normal consultands). This dataset will be called the London sample and was coded as described

LIABILITY CLASS

The Wessex and London datasets were analysed in parallel and as a pooled set. The data were handled in the same way after the initial coding differences, which take account of the well defined ascertainment biases. Each person whether affected or not was placed in a liability class based on sex and age at last observation (table 1). Males were assigned to the lowest liability class (LI = 1).

SEVERITY

The POINTER program⁸ uses age at last observation to assign subjects to liability classes for segregation analysis. In a disease such as breast cancer, the earlier the age at onset, the more likely a person is to carry the high risk genotype. The 10 year survival after a diagnosis of breast cancer is about 50% in the UK for all stages, so a substantial number of those women diagnosed in one liability class would be observed in a later liability class and the information about a much earlier age at onset would be lost. In order to retrieve the information about age at onset we have used the COMDS program for analysis incorporating age at onset as severity through a parameter “S”. Where the interval between age at onset and age at observation is large, and age at onset is young, the person is assigned the more severe class; conversely where age at onset is within the same range as age at observation, the liability class assigned takes account of the age at onset to a large extent and the least severe class is assigned. In the final analysis the magnitude of the parameter S gives an indication of the extent to which the information contained in S contributes towards distinguishing alternative genetic models. Where S tends towards 1.0, a large proportion of information has been retrieved using severity classes, where S tends towards 0 then very little additional information comes from assigning severity class to individual persons.

To include this parameter, we used the scheme set out in table 2. A severity class from SV1 to SV5 was assigned to all affected cases. Where the age at onset was 40 years or greater, cases were assigned to the least severe class 1

Table 2 Definition of severity for persons with known age of onset. Severity frequencies taken from age specific breast cancer incidence for UK population (incident cases 1979–1982 in England and Wales (IARC, 1987))

Age at onset	Liability class (age of observation)								
	20–24	25–29	30–34	35–39	40–49	50–59	60–69	70–79	80+
	1	2	3	4	5	6	7	8	9
<25	SV1	SV5	SV5	SV5	SV5	SV5	SV5	SV5	SV5
25–29		SV1	SV4	SV4	SV4	SV4	SV4	SV4	SV4
30–34			SV1	SV3	SV3	SV3	SV3	SV3	SV3
35–39				SV1	SV2	SV2	SV2	SV2	SV2
40 +					SV1	SV1	SV1	SV1	SV1

Severity frequencies (SV1, ..., SV5): 0.9617, 0.0241, 0.0106, 0.00311, 0.0005.

Table 3 Segregation analysis of the Wessex data (joint likelihood)

Model	Joint				
	D	T	Q	S	In L
Dominant, S=0	(1)	2.31107	0.00362	(0)	722.780572
Dominant, S	(1)	2.17555	0.00243	0.57536	690.412189
Additive, S=0	(0.5)	4.54100	0.00386	(0)	722.997913
Additive, S	(0.5)	4.44544	0.00213	0.52675	691.636779
Recessive, S=0	(0)	6.98289	0.09911	(0)	738.495829
Recessive, S	(0)	4.51093	0.08123	0.86596	719.664364
General model, S=0	0.66099	3.45815	0.00380	(0)	722.687587
General model, S	0.82289	2.64137	0.00244	0.57483	690.399051

Table 4 Segregation analysis of the Wessex data (conditional likelihood)

Model	Conditional				
	D	T	Q	S	In L
Dominant S=0	(1)	2.40769	0.00525	(0)	455.589889
Dominant S	(1)	2.23437	0.00318	0.75685	431.446894
Additive S=0	(0.5)	4.63389	0.00643	(0)	455.520754
Additive S	(0.5)	4.58532	0.00269	0.63551	432.599051
Recessive S=0	(0)	16.55377	0.05001	(0)	493.707590
Recessive S	(0)	19.70944	0.05054	0.00076	493.323756
General model S=0	0.59788	3.92478	0.00616	(0)	455.228293
General S	0.73773	3.01501	0.00325	0.76089	431.349074

(SV1). For age at onset under 40 years where the age at onset and the age at observation fell within the same liability class, the severity class assigned was SV1 and where the age at onset was younger than the liability class age range, increasing severity scores were assigned as SV2, SV3, SV4, and SV5 according to the age at onset of breast cancer. In this way the maximum allowance for onset age falling outside the liability class in which a person was observed was made for those with a very early age at onset (<25 years) with decreasing severity classes for those with increasing ages at onset.

CORRECTION FOR ASCERTAINMENT

Ascertainment

Ascertainment was defined as single selection for families with a single proband, as truncate selection for sibships ascertained through a secondary proband and a pointer, and otherwise as incomplete selection with a small ascertainment probability π (0.001). In the London sample this last group included multiplex selection through at least two affected sibs as described by Iselius *et al.*² In COMDS single selection implies conditioning on severity and liability class of the proband and pointer as an ascertainment event. Other types of ascertainment neglect severity and condition on the measure of ascertainment for the family assuming that probands are drawn at random from affecteds in their liability class.⁹

Analysis

The parameters of the genetic model were estimated by maximising the likelihood of the

observed phenotypes in families using joint and conditional likelihoods. Joint likelihood conditions on parent and offspring phenotypes whereas conditional likelihood considers offspring phenotypes conditional on parental phenotypes. The conditional likelihood method is more robust with respect to ascertainment bias but extracts less information from the data than the joint likelihood approach.

Results

Table 3 shows the results of the analysis on the Wessex dataset under joint likelihood and under conditional likelihood (table 4). The data favour the dominant model for both joint and conditional likelihood with gene frequencies of 0.0036 and 0.0053 respectively. The introduction of severity corrects the gene frequency to 0.0024 (joint) and 0.0032 (conditional). The severity parameter contributes more information using conditional likelihood ($S=0.7569$ for conditional likelihood versus $S=0.5754$ for joint likelihood).

Tables 5 and 6 illustrate the comparison of Wessex and London data with and without the severity parameter. For both joint and conditional likelihood where severity=0, there is significant evidence of heterogeneity between the London and Wessex samples (table 5). However, where severity is incorporated for conditional likelihood (table 6) there is no significant heterogeneity ($\chi^2=6.3$, $p>0.05$, NS). Evidence for heterogeneity remains, however, using joint likelihood ($\chi^2=8.4$, $p<0.05$).

Penetrance estimates differ quite markedly between the London and the Wessex datasets

Table 5 Comparison of two samples with S=0

	Joint			Conditional		
	-2 In L+C	t	q	-2 In L+C	t	q
London	465.839287	1.83656	0.00384	126.144401	2.05382	0.00255
Wessex	722.780572	2.31107	0.00362	455.589889	2.40769	0.00525
London + Wessex	1188.619859			581.734290		
Pooled	1201.915258	2.02196	0.00340	591.789074	2.17637	0.00284

$\chi^2=13.295399$ (2 df) 0.01< p<0.05.

$\chi^2=10.054784$ (2 df) 0.01< p<0.05.

Table 6 Comparison of two samples with $S \neq 0$

	Joint likelihood			Conditional likelihood		
	$-2 \ln L + C$	t	q	$-2 \ln L + C$	t	q
London S = 0.44882	452.547613	1.78165	0.00397	111.452756 S = 0.45607	2.00238	0.00258
Wessex S = 0.57536	690.412189	2.17555	0.00243	431.446894 S = 0.75685	2.23437	0.00318
London + Wessex Pooled S = 0.55237	1142.959802 1151.401697	1.94111	0.00308	542.899650 549.205803 S = 0.56430	2.09492	0.00265
$\chi^2 = 8.441895$ (3df) 0.01 < p < 0.05.			$\chi^2 = 6.306153$ (3df) p > 0.05 (NS).			

Table 7 Comparison of penetrance values under different genetic models

Liability class	Age	Heterogeneous under joint likelihood				Homogeneous under conditional likelihood	
		London		Wessex		Pooled	
		G/G	G'/-	G/G	G'/-	G/G	G'/-
1	20-24	0.00000	0.00555	0.00000	0.00979	0.00000	0.00888
2	25-29	0.0001	0.03137	0.00007	0.05883	0.00008	0.05250
3	30-34	0.00066	0.09093	0.00055	0.16925	0.00058	0.15081
4	35-39	0.00232	0.17932	0.00221	0.31469	0.00224	0.28340
5	40-49	0.00967	0.36177	0.00976	0.56544	0.00976	0.52106
6	50-59	0.02269	0.57641	0.02310	0.84096	0.02303	0.78552
7	60-69	0.03939	0.77029	0.04052	1.00000	0.04009	1.00000
8	70-79	0.05918	0.95011	0.06167	1.00000	0.06126	1.00000
9	80-85 +	0.08413	1.0000	0.08693	1.00000	0.08653	1.00000
		$q = 0.00397$		$q = 0.00243$		$q = 0.00265$	

(table 7, fig 1), with penetrance reaching 100% by 69 years of age in the early onset (Wessex) dataset compared to 77% at the same age in the London dataset. Fig 1 illustrates these higher and lower estimates of penetrance.

CLINICAL APPLICATION

An illustraton of the way the Wessex model performs in a risk calculation is given in fig 2A and B which also incorporates the results of the same calculation using the parameters derived from USA and UK data.^{4,10} The consultand is subject 6 in fig 3 and the family history is as illustrated. The calculation of the probability that this consultand, unaffected by breast cancer at 35 years of age, will develop the disease has been calculated at 10 year intervals using the three segregation models represented in fig 4. The probability of affectedness for the

consultand by a given age (j) with a particular family history is given by the likelihood of the pedigree (F) with the consultand affected at that age divided by the likelihood of the pedigree with the consultand omitted.

$$P(Aff|j,F) = P(F,Aff|j).$$

The main likelihoods are calculated from the segregation parameters in each of the models using the computer program MENDEL¹¹ and essentially provide for each pedigree member a probability that they carry a high risk genotype based on age and phenotype (affected or unaffected).

Discussion

The incidence of breast cancer increases with increasing age and is relatively rare below the age of 40 years. Data from genetic epidemiology studies of breast cancer are likely to be deficient in numbers of early onset cases. The models thus reflect an average of the parameters from all the dominant genes for breast cancer in the population. We selected cases of breast cancer specifically for early onset to try to refine a segregation model reflecting a gene or genes predisposing to unusually early onset breast cancer. We have addressed the problems of correcting for ascertainment bias in this sample and incorporating information about age of onset as an additional parameter “S” for severity which has not been included in previous analyses using POINTER.² These factors improve the accuracy of the model to be used in risk calculations.

In the Wessex dataset the severity parameter gives more information about the genetic model under conditional likelihood, whereas for the London dataset the severity parameter is very similar for joint and conditional but is less than for the Wessex dataset. This is not surprising given the composition of the Wessex dataset with selection for more “severe” early onset

Figure 1 Penetrance as cumulative incidence estimated under joint likelihood from pedigrees ascertained through consecutive and selected (London) and early onset (Wessex) breast cancer patients.

Figure 2 Risk profiles calculated from the three separate segregation models when the consultand is 35 years and unaffected (A) and when she is 60 years and unaffected (B).

probands compared with the London dataset where selection was not dependent on age at onset in probands. Under joint likelihood the severity correction may be incomplete using the method described but a more likely explanation is that there is true heterogeneity under joint likelihood and this could be explained by invoking separate alleles or loci responsible for early and late onset disease. The latter seems to be a plausible explanation in view of the observations in linkage studies on 17q21.^{8,12} Average age at onset under 45 years in a family confers a greater chance of finding linkage to BRCA1. At least some breast cancer clusters in families where age at onset

Figure 3 Typical pedigree structure for a consultand attending a cancer family clinic.

Figure 4 Penetrance estimated from USA data⁴ using a step function model along with penetrance estimated from the London data²¹³ and the Wessex data (fig 1).

is on average greater than 45 years are probably the result of one or more other dominant genes with different penetrances and frequencies. Other family clusters may result from chance or shared environmental factors as phenocopies become increasingly common in the population with increasing age.

Selecting for early onset cases should give a more representative estimate of genetic parameters when estimating risks for a family history of early onset disease. Where age at onset in the family is more typical, the London dataset may provide a better estimate. Until parameters for each separate allele can be estimated, preferably with the development of methods that allow for genotype specific mortality, the range provided here (table 7, figs 1 and 4) will be useful for risk prediction and also gives some indication of the range of risks which could be calculated for persons with differing family histories. Comparison of fig 2A and B illustrates the effect of age specific penetrance and the reduction in genetic risk with increasing age at observation. Where age at onset is unusually early, the Wessex model may be more accurate. However, a range of risks (low and high estimates) may be helpful to the clinician for use

in the clinic as well as providing a means of ranking risks for high risk women entering clinical trials of screening and prevention in breast cancer.

Of course molecular detection of mutant alleles would give much improved risk estimates, but it will be some time before this becomes available. In addition, if the true gene frequency were 0.0030, and if detection were perfect, it would take 6000 random persons to give a standard error of 0.0005.

This work was supported by the Cancer Research Campaign, grant number SP1935/0102. We would like to thank Mrs Bridget Dunn for her assistance with pedigree work; she is supported by a grant from the Wessex Cancer Trust.

Glossary of terms

Ascertainment. Correction for ascertainment bias is achieved through coding the pedigree data in a way which reflects the way in which a pedigree was selected.

Conditional likelihood. Approach is defined by offspring given parental phenotype.

Genotype specific mortality. Where a disease can cause premature death. The gene frequency in the ageing population will not be constant; this factor is not taken into account in any current segregation analyses.

Joint likelihood. Defined by offspring and parental phenotypes where both are analysed with no significance being attached to the phenotype in previous generations.

Liability class. Age at observation of all observed persons is used to assign each into a liability class where the incidence of a disease increases with age.

Segregation analysis. A method which seeks to define the most likely mode of inheritance of a disease in a given population and to estimate the frequency and penetrance if a genetic model is likely.

Severity. This is an additional parameter which can be estimated (using COMDS) to retrieve information about age at onset from the data, some of which is lost when age at observation (liability class) alone is used.

Single selection. Used to define families ascertained through a single affected proband taken at random from a defined popu-

lation, giving an ascertainment probability proportional to the number of affected in the family.

Truncate selection. Used to define families ascertained through at least one affected child, where the ascertainment probability for the family does not depend on the number of affected children if greater than 1.

Pointer. This is the person outside a nuclear family through whom that nuclear pedigree has been ascertained. Her relationship to the closest relative in each nuclear pedigree is reflected in the coding.

- 1 Newman B, Austin M, Lee M, King MC. Inheritance of human breast cancer: evidence for autosomal dominant transmission in high risk families. *Proc Natl Acad Sci USA* 1988;**85**:3044-8.
- 2 Iselius L, Slack J, Littler M, Morton NE. Genetic epidemiology of breast cancer in Britain. *Ann Hum Genet* 1991;**55**:151-9.
- 3 Iselius L, Littler M, Morton NE. Transmission of breast cancer: a controversy resolved. *Clin Genet* 1992;**41**:211-17.
- 4 Claus EB, Risch N, Thompson WD. Genetic analysis of breast cancer in the cancer and steroid hormone study. *Am J Hum Genet* 1992;**148**:232-42.
- 5 Cannings C, Thompson EA. Ascertainment in the sequential sampling of pedigrees. *Clin Genet* 1977;**12**:208-12.
- 6 Houlston RS, McCarter E, Parbhoo S, Scurf J, Slack J. Family history and risk of breast cancer. *J Med Genet* 1992;**29**:154-7.
- 7 Morton NE, Rao DC, Lalouel JM. Segregation analysis of familial data. In: *Methods in genetic epidemiology*. Basel: S Karger, 1983:92-7.
- 8 Hall JM, Lee MK, Newman B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 1990;**250**:1684-9.
- 9 Morton NE, Shields DC, Collins A. Genetic epidemiology of complex phenotypes. *Ann Hum Genet* 1991;**55**:301-14.
- 10 Shields DC, Marlow AJ, Houlston RS, Eccles DM, Morton NE. Prediction of genetic risks from segregation analyses of morbid risks. *Hum Hered* 1994;**44**:52-5.
- 11 Lange K, Weeks D, Boehnke M. Programs for pedigree analysis: MENDEL, FISHER and dGene. *Genet Epidemiol* 1988;**5**:471-2.
- 12 Easton DF, Bishop DT, Ford D, Crockford GP, and the Cancer Linkage Consortium. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. *Am J Hum Genet* 1993;**52**:678-701.
- 13 Knudson AG. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA* 1971;**68**:820-3.