The Genetics of Breast and Ovarian Cancer I: A Model of Family History

A. S. MACDONALD, H. R. WATERS and C. T. WEKWETE

Macdonald AS, Waters HR, Wekwete CT. The genetics of breast and ovarian cancer I: a model of family history. Scand. Actuarial J. 2003; 1: 1–27.

We present a Markov model of breast cancer (BC) and ovarian cancer (OC) and estimate its transition intensities, mainly using United Kingdom population data. In the case of BC and OC, we estimate intensities according to BRCA1 and BRCA2 genotype. We use this to estimate the probabilities that an applicant for insurance has a BRCA1 or BRCA2 mutation, given complete or incomplete knowledge of her family history of BC and OC. Life (and other) insurance underwriters typically have incomplete knowledge of family history, for example no information on the number of healthy relatives. We show how these probabilities depend strongly on estimates of the mutation frequencies and penetrances, and conclude that it may not be appropriate to apply risk estimates based on studies of high-risk families to other groups. *Key words: Breast cancer, BRCA1 gene, BRCA2 gene, family history, ovarian cancer, underwriting.*

1. INTRODUCTION

Macdonald (1997, 1999) proposed a Markov model for the study of genetics and insurance. The first application to specific genes was by Lemaire et al. (2000) and Subramanian et al. (1999) who studied life insurance underwriting in the presence of genetic tests for mutation in the BRCA1 and BRCA2 genes implicated in breast cancer (BC) and ovarian cancer (OC). The present study considers critical illness (CI) insurance in the presence of the same genetic tests. Life insurance and CI insurance present different problems, partly because of the maturity of the life insurance market compared with the CI market.

In this paper, Part I, we describe the role of the BRCA1 and BRCA2 genes in Section 2. Consideration of the underwriting of a family history of BC or OC leads us to construct a model for the life history of a relative of a woman applying for insurance. We call this the relatives' model. It is described in Section 3, and the estimation of the transition intensities in Section 4.

Given the model, we can compute conditional probabilities of the form:

P[Woman age x has a given genotype | Some observed information],

where the observed information might be:

(a) the ages and complete life-histories (with respect to BC and OC) of all the woman's relatives; or
(b) the more limited information typically available to an underwriter, such as the number of her relatives who have had BC or OC.

Sections 5 to 7 describe these calculations. These are based on penetrance estimates (that is, estimates of the probability that a woman with a gene mutation will get BC or OC during her lifetime) derived from studies of families with a very clear history of BC and/or OC. It is possible that the penetrance is lower among women who have a gene mutation but who lack a clear family history. In Section 8 we show the effect of significantly reduced penetrances. Our conclusions are in Section 9.

In Part II, we study CI insurance. We propose a model with transitions representing CI events other than the onset of BC and OC (which we call the applicant's model) and estimate its intensities. From Thiele's equations we obtain genotype-specific costs of CI insurance, and we combine these with the probabilities from the relatives' model to determine the impact of known or unknown genetic information on underwriting, and the potential for adverse selection.

2. THE GENETICS OF BREAST AND OVARIAN CANCER

For brevity, we refer the reader to other works for genetic terminology. Fischer & Berberich (1999) is a good introduction, and Strachan & Read (1999) and Pasternak (1999) are standard texts.

2.1. The BRCA1 and BRCA2 Genes

BRCA1 is a large gene (5 592 base pairs) located on chromosome 17q21. BRCA2 is even larger (10 254 base pairs) and is on chromosome 13q12-13. They were sequenced in 1994 and 1995 respectively. The functions of the proteins encoded by these genes are not known, but they are believed to be tumour suppressor genes.

Under the 'two-hit' hypothesis, both copies of one or other gene have to be disfunctional in any given cell for the protection they confer to be lost. This can happen in one of three ways:

(a) A woman can inherit two defective copies of either gene. Then tumour suppression will be switched off in every cell in the affected tissues, presumably resulting in extreme risk of cancer. However, homozygosity is so rare that this can only be supposed.
(b) A woman can inherit one defective copy of either gene. Then tumour suppression is effective until the working copy of the gene is disabled in any one or more cells, for example by a somatic mutation, and the affected cell becomes potentially cancerous.
(c) A woman can inherit two working copies of both genes. Then both copies have to be disabled in the same cell for tumour suppression to cease in that cell. This is a very rare event in any one cell.

BRCA1 and BRCA2 do not account for all BC and OC. Mutations are relatively rare in the population, and moreover there are hundreds of different mutations in each. Most of these have only ever been observed in one family. Only the most expensive form of genetic testing (full gene sequencing, currently costing about U.S. $2 400 for both genes) can detect any mutation; cheaper tests will only look for a few known mutations, or check specific regions of the gene where mutations are known to disable the gene product.

Different mutations might have different outcomes, so the mere presence of a mutation does not determine the risk, although most studies to date assume that it does (as do we). This assumption will be capable of refinement in due course, probably when the precise structures and functions of the gene products are worked out. It will then be necessary to re-assess the impact of mutations on insurance risk.

Further, it is possible that the penetrances of BRCA1 and BRCA2 mutations are affected by other genes or by the environment. That is, although the genotypes show simple Mendelian inheritance (because these are single genes) the phenotypes do not, for a variety of reasons. Early reports indicated a cumulative risk of BC of over 80% by age 70 (Easton et al., 1993, 1995) among BRCA1 mutation carriers. Discussions of the implications for insurance initially tended to assume that BRCA1 mutations would be as highly penetrant among all carriers, ignoring the fact that the gene had been discovered by studying families with a very strong family history of BC. A later study (Struewing et al., 1997) of an Ashkenazi Jewish population suggested penetrance of less than 60%, and a recent study of women affected by BC but not selected for family history or by ethnic group (Hopper et al., 1999) suggested penetrance of 40% or less. Confidence intervals were quite wide in all cases. In this study, we have fitted incidence rates to the studies based on BC and OC families, and we suppose that these are high estimates. In Section 8, we show the effect of much lower penetrances.

It is possible that other BRCA genes will be discovered, though Antoniou et al. (2000) found no evidence for another single gene associated with OC.

2.2. Gene frequencies

Population estimates of the frequencies of BRCA1 and BRCA2 mutations vary quite widely. Most studies are based on selected populations, either of women with a strong family history of BC and OC, or women who themselves have BC or OC. For example, about 2.5% of Ashkenazi Jewish women have a BRCA1 or BRCA2 mutation, but the frequency in the whole population is much lower. These women mostly have one of two specific BRCA1 mutations, or one specific BRCA2 mutation, indicating likely descent from a single ancestor; this feature is absent from the general population.

Table 1 gives some estimates of mutation frequencies from Claus et al. (1994) and Parmigiani et al. (1998). These are frequencies per allele; since every woman has two alleles of each gene the distribution of any genotype can be calculated. For example, the probability of having one BRCA1 mutation and no BRCA2 muta-

Table 1. *Estimates of the population frequency of mutations in BRCA1 and BRCA2 alleles*

Author	Frequency of mutations in	
	BRCA1 alleles	BRCA2 alleles
Claus et al. (1994)	0.0033	Not in study
Parmigiani et al. (1998) (a)	0.0006	0.00022
Parmigiani et al. (1998) (b)	0.0008	0.00030
Parmigiani et al. (1998) (c)	0.00045	0.000165

tions, using the frequencies in the second line of the table, is $2 \times 0.006 \times 0.9994 \times 0.99978^2 = 0.001199$.

For our modelling, we take estimates (b) and (c) of Parmigiani et al. (1998) as upper and lower limits.

2.3. Family history and genetic information in clinical practice

In clinical practice, when a woman presents with a family history of breast or ovarian cancer, the first task is to assess the risk that she herself runs of contracting these cancers. The relevant questions include:

(a) how many blood relatives have had breast and/or ovarian cancer;
(b) what was their precise relationship to the proband;
(c) at what ages did they contract cancer; and
(d) what were the clinical facts about each case?

The construction of the family history is not usually straightforward. Many people will not know the details, or will make mistakes, and the clinician will have to research cancer registries and medical records to get an accurate pedigree. Then the cancer risk can be assessed, relative to an ordinary member of the population (usually using proprietary computer programs). Women at high risk may then be referred to a clinical geneticist for counselling, and, if deemed appropriate, for genetic testing.

Even if a genetic test is carried out, the results need careful interpretation since:

(a) little is known about the function of the proteins encoded by BRCA1 and BRCA2;
(b) most of the hundreds of known mutations have each been observed in only one family;
(c) not all genetic tests sequence the whole gene; some only test for a number of known mutations, or just search a segment of coding region of the gene;
(d) many mutations in BRCA1 and BRCA2 surely remain to be found;
(e) other oncogenes are known to affect tumour formation, such as P53 and TPEN; and
(f) there are almost certainly other breast and ovarian cancer genes as yet undiscovered.

Given this uncertainty, genetic tests for BRCA1 and BRCA2 have not replaced conventional methods of risk assessment; they have not transformed uncertainty into certainty overnight. These factors are, of course, also relevant for the assessment of insurance risk.

3. THE RELATIVES' MODEL

We need different models for the life histories of the applicant after she is insured, and of the applicant and her relatives up until she applies for insurance:

(a) Once the applicant becomes insured, we need to model all events that lead to a claim, not just the onset of BC or OC.
(b) The relatives provide the family history of BC and OC used in underwriting. Here, we are not interested in other events, even a relative dying from some other cause.
(c) The age at onset of BC or OC is an important part of family history. We will need to incorporate this in the relatives' model, but not the applicant's model.

Underwriters usually group age at onset into broad bands; for example one large reinsurer uses the groups < 50, $50–65$ and > 65, and so do we. This leads to the model shown in Fig. 1.

Suppose there are M relatives, including the applicant, and the ith relative has genotype g_i. By convention, the first relative ($i = 1$) is the applicant. The history of each relative is represented by the model in Fig. 1, with states labelled $0, 1, \ldots, 7$ and transition intensity $^{g_i}v_x^{0j}$ between states 0 and j. Further, define $^{g_i}v_x = \Sigma_{j=1}^{j=7} {}^{g_i}v_x^{0j}$. The probability that the ith relative is in state j at age $x + t$, given that she was healthy at age x, will be denoted $^{g_i}p_{x,t}^{0j}$ and we have:

$$^{g_i}p_{x,t}^{00} = \exp\left(-\int_0^t {}^{g_i}v_{x+s}\, ds \right), \tag{1}$$

$$^{g_i}p_{x,t}^{0j} = \int_0^t {}^{g_i}p_{x,s}^{00}\, {}^{g_i}v_{x+s}^{0j}\, ds \quad (j = 1, \ldots, 7). \tag{2}$$

The 'dead' state excludes deaths from BC or OC, since onset of these is represented by other states (the future life history is irrelevant once BC or OC has been contracted). Also, the 'healthy' state merely indicates freedom from BC and OC; a relative with heart disease or another cancer, for example, would be 'healthy' for our purposes.

4. ESTIMATION OF THE TRANSITION INTENSITIES FOR THE RELATIVES' MODEL

In this section we estimate the intensities that appear in the relatives' model, namely:

(a) incidence rates of BC and OC in the population, taken to be the rates applying to women with no BRCA1 or BRCA2 mutations;
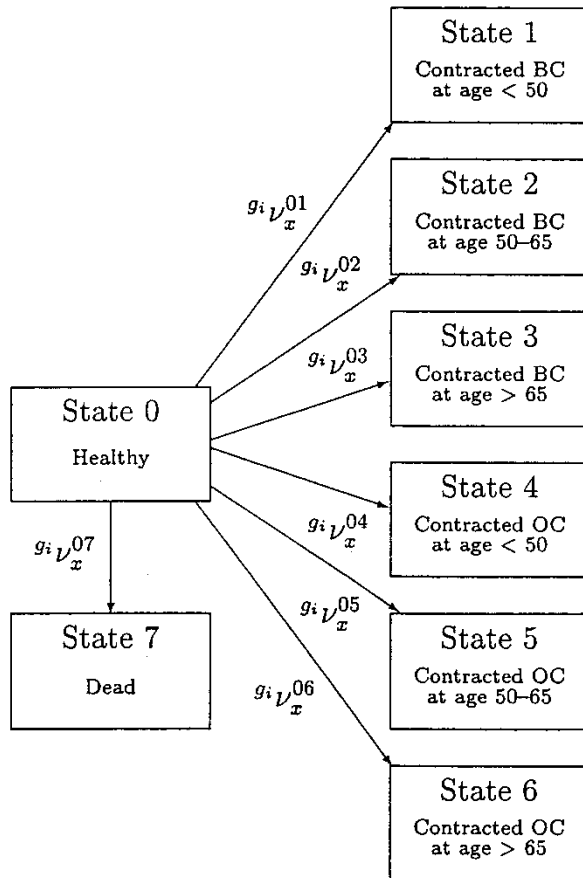
Fig. 1. A model for the *i*th relative ($i = 1, 2, 3, \ldots, M$) of the insured woman, with genotype g_i. Relative No. 1 is the woman herself.

(b) incidence rates of BC and OC for women with BRCA1 or BRCA2 mutations, or both; and

(c) rates of mortality from all causes except BC and OC.

The other intensities needed for the applicant's model will be estimated in Part II.

4.1. Population risk of breast and ovarian cancer

Fig. 2 shows crude and graduated incidence rates of BC and OC in the U.K. based on cancer registrations in England and Wales in 1990–92 (O.N.S., 1999), and mid-year population estimates in 1989–93. The latter were not adjusted to remove women with BC or OC, respectively, as doing so had very little effect. Feuer et al. (1993) measured the anomaly using a multiple decrement approach in a hypothetical cohort of women, and found that the effect was small even at ages as high as 85.
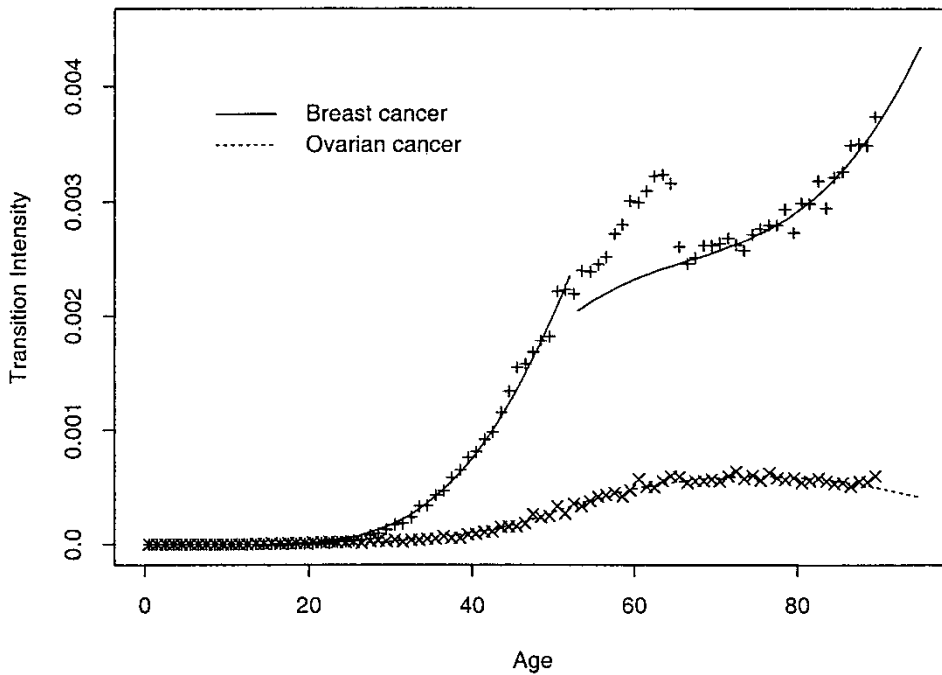
Fig. 2. Modelled population incidence rates (transition intensities) of breast and ovarian cancer.

The crude BC incidence rates are unusual at ages 50–64. The reason is the introduction of a BC screening program in the U.K. in 1988, covering women in that age group. Screening does not alter the incidence of tumour formation, but it results in tumours being discovered sooner, before they are palpable, and the 'BC incidence rates' are, of course, rates of diagnosis of BC. The crude rates in Fig. 2 show the effect of a recently-introduced program, increasing BC incidence at all the screened ages. However, screening is repeated at 3-year intervals, and it is expected that as the program matures the earlier detection will only affect ages 50–52 (Quinn & Allen, 1995) and that is what is shown in the graduated rates. Rates at ages 0–52 and 53–95 were graduated separately, by Gamma and cubic functions respectively, as follows:

$$\mu_x^{\text{BC,POP}} = \frac{1}{\Gamma(8.7305)} \, 0.0742^{8.7305} \, e^{-0.0742x} x^{7.7305} \quad \text{for } 0 \leqslant x < 53, \tag{3}$$

$$\mu_x^{\text{BC,POP}} = 0.00012 + 0.00018(x - 35) - 0.000005(x - 35)^2 + 0.0000000529(x - 35)^3$$

$$\text{for } x \geqslant 53. \tag{4}$$

The OC incidence rates were graduated by different formulae at ages under 45 and over 55, as follows:

$$\mu_x^{\text{OC,POP}} = \frac{1}{\Gamma(6.92)} (0.035^{6.92} \, e^{-0.035x} x^{5.92}) \quad \text{for } 0 \leqslant x < 45, \tag{5}$$

$$\mu_x^{\text{OC,POP}} = 0.0001554 + 0.000029(x-45) - 0.00000048(x-45)^2 \quad \text{for } x \geqslant 55, \tag{6}$$

and at ages 45–55 these formulae were blended by linear interpolation.

4.2. Breast cancer risks associated with the BRCA1 and BRCA2 genes

Ford et al. (1998) gave the estimated incidence rates of BC among mutation carriers shown in Table 2. These cover ages 20–69, but for our purposes we need rates at higher ages. There is some evidence that the excess risk falls at higher ages, so we graduated and extrapolated these crude rates by fitting truncated Gamma functions:

$$\mu_x^{\text{BC, BRCA1}} = \frac{1.25}{\Gamma(22)} (0.45^{22} \, e^{-0.45x} x^{21}), \tag{7}$$

$$\mu_x^{\text{BC, BRCA2}} = \frac{3.60}{\Gamma(30)} (0.45^{30} \, e^{-0.45x} x^{29}). \tag{8}$$

Fig. 3 shows these fitted rates, extrapolated to age 95. It is not possible to show confidence intervals; Ford et al. (1998) gave confidence intervals for cumulative risks but not for incidence rates. The cumulative risk at age x is the probability of having contracted BC before age x, assuming onset of BC to be the only decrement. For example:

$$\text{Cumulative risk at age } x, \text{ with BRCA1 mutation} = 1 - \exp\left(-\int_0^x \mu_t^{\text{BC, BRCA1}} dt \right). \tag{9}$$

We prefer to graduate incidence rates, because we have more intuition about these than about cumulative risks. In Table 3, we compare the cumulative risks calculated from our graduated incidence rates with those published, and their confidence intervals, with satisfactory results. If we consider an applicant for insurance at ages 20–50, the cumulative risks in respect of her mother and aunts (that is, at ages about 50–80) matter most.

Iversen et al. (2000) obtained parametric estimates of penetrance curves from the same data in a study of survival with BC.

Table 2. *Incidence rates of breast cancer for carriers of BRCA1 and BRCA2 mutations. (Source: Ford et al. (1998))*

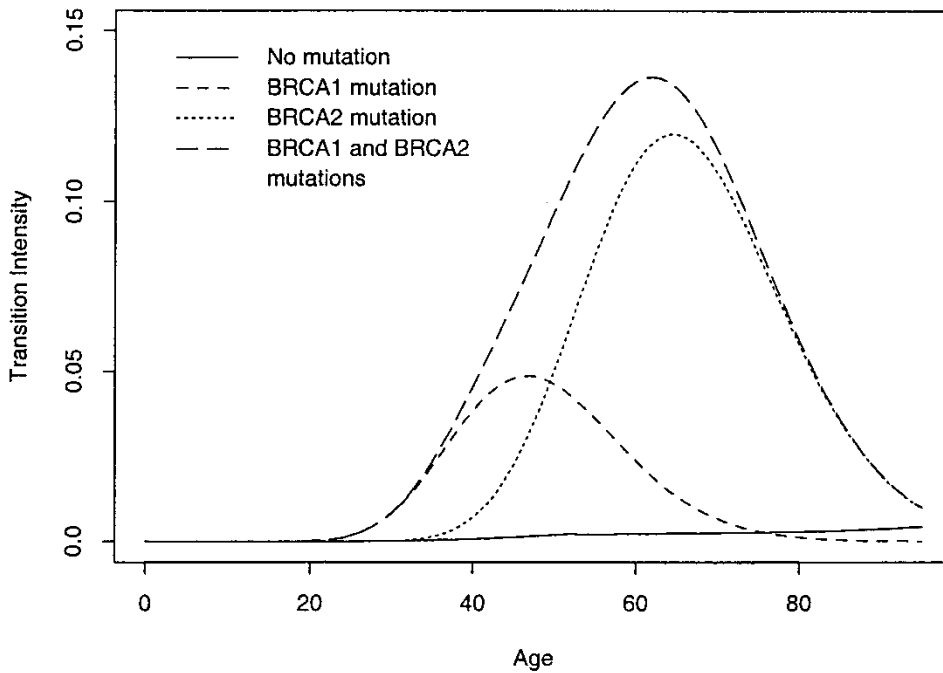|	Age group				
	20–29	30–39	40–49	50–59	60–69
BRCA1	0.003660	0.016180	0.047490	0.034800	0.021620
BRCA2	0.000633	0.0118	0.0210	0.0318	0.118

Fig. 3. Modelled incidence rates (transition intensities) of breast cancer, depending on the presence of BRCA1 and/or BRCA2 mutations.

Table 3. *Modelled cumulative risks of breast cancer, compared with observed rates and 95% confidence intervals. (Source: Ford et al. (1998))*

Age x	BRCA1 mutation			BRCA2 mutation		
	Observed cumulative risk	Confidence interval	Modelled cumulative risk	Observed cumulative risk	Confidence interval	Modelled cumulative risk
30	0.036	0.00–0.14	0.0252	0.006	0.00–0.019	0.0003
40	0.18	0.00–0.35	0.2221	0.12	0.00–0.240	0.0212
50	0.49	0.28–0.64	0.5097	0.28	0.09–0.440	0.2359
60	0.64	0.43–0.77	0.6572	0.48	0.22–0.650	0.6683
70	0.71	0.53–0.82	0.7020	0.84	0.43–0.950	0.8962

4.3. Ovarian cancer risks associated with the BRCA1 and BRCA2 gene

OC is rarer than BC, and data on genetic risk are much scarcer. Table 4 shows estimated incidence rates from Ford et al. (1998). These were also graduated and extrapolated using truncated Gamma functions:

$$\mu_x^{OC, BRCA1} = \frac{0.60}{\Gamma(21)} (0.37^{21} \, e^{-0.37x} x^{20}), \tag{10}$$

Table 4. *Incidence rates of ovarian cancer for carriers of BRCA1 and BRCA2 mutations. (Source: Ford et al. (1998))*

	Age group				
	20–29	30–39	40–49	50–59	60–69
BRCA1	0	0	0.0169600	0.0194000	0.0168900
BRCA2	0	0	0.000425	0.00722	0.0236

$$\mu_x^{\text{OC, BRCA2}} = \frac{0.86}{\Gamma(27)} (0.355^{27}\, e^{-0.355x} x^{26}). \tag{11}$$

Fig. 4 shows these fitted rates, extrapolated to age 95 (note that Figs. 3 and 4 are not to the same scale). Ford et al. (1998) do not give confidence intervals for the cumulative BRCA1 risk of OC alone, but for either BC or OC. In Table 5, we compare these with our modelled cumulative risks, again with satisfactory results.

4.4. Treatment of multiple mutations

We need to decide how to treat multiple mutations: either two mutated alleles of the same gene, or mutations in both the BRCA1 and BRCA2 genes.
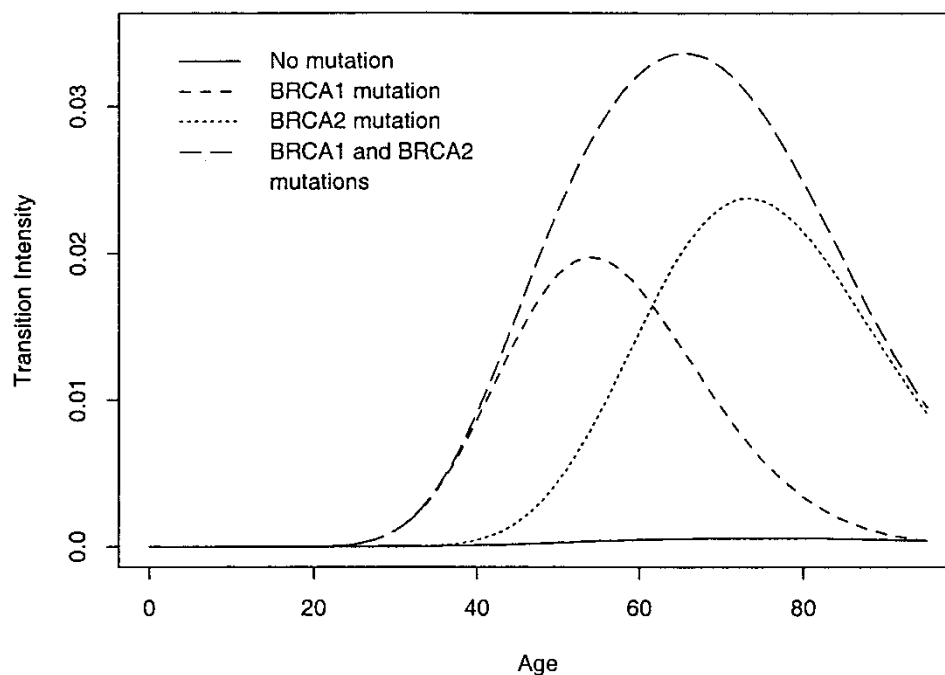

Fig. 4. Modelled incidence rates (transition intensities) of ovarian cancer, depending on the presence of BRCA1 and/or BRCA2 mutations.

Table 5. *Modelled cumulative risks of breast or ovarian cancer (BRCA1) and ovarian cancer (BRCA2), compared with observed rates and 95% confidence intervals. (Source: Ford et al. (1998))*

Age x	Breast or ovarian cancer BRCA1 mutation			Ovarian cancer BRCA2 mutation		
	Observed cumulative risk	Confidence interval	Modelled cumulative risk	Observed cumulative risk	Confidence interval	Modelled cumulative risk
30	0.036	0.00–0.14	0.0282	0.000	n/a	<0.0001
40	0.18	0.00–0.36	0.2562	0.000	n/a	0.0014
50	0.57	0.33–0.73	0.5930	0.004	0.00–0.011	0.0207
60	0.75	0.53–0.87	0.7650	0.074	0.00–0.150	0.1068
70	0.83	0.65–0.92	0.8218	0.27	0.00–0.470	0.2657

(a) Women with two mutated copies of either gene are so rare that there is no real evidence. It is extremely unlikely that a woman should inherit two copies of the same mutation, unless within a population descended from a single ancestor. We have assumed that the risks would be the same as for a carrier of one mutated allele.

(b) We suppose that the incidence rates for women with mutations in both BRCA1 and BRCA2 are obtained by adding the individual incidence rates, which seems reasonable under the 'two-hit' model. This assumption is reflected in Figs. 3 and 4, and in Table 5.

4.5. Rate of mortality excluding breast and ovarian cancer

The force of mortality $^{gi}v_x^{07}$ in Fig. 1 is ELT15F adjusted to remove deaths from BC and OC. The ratio ϕ_x of the number of deaths from BC and OC in 1990–92, to the total number of deaths (O.P.C.S., 1991, 1993a,b; O.N.S., 1999), was graduated below age 54 by:

$$\phi_x = \frac{8.63}{\Gamma(14.05)}(0.28^{14.05}\, e^{-0.28x}x^{13.05}) \tag{12}$$

and above age 65 by:

$$\phi_x = 1.30144 - 0.02850194x + 0.0001588314x^2 \tag{13}$$

with linear interpolation between these ages. The results are shown in Fig. 5.

5. CALCULATION OF CONDITIONAL PROBABILITIES IN THE MODEL

When a woman applies for insurance, the underwriter may learn her family history but not her genotype. Within the model, however, we are omniscient. In particular, we may assume that we know:
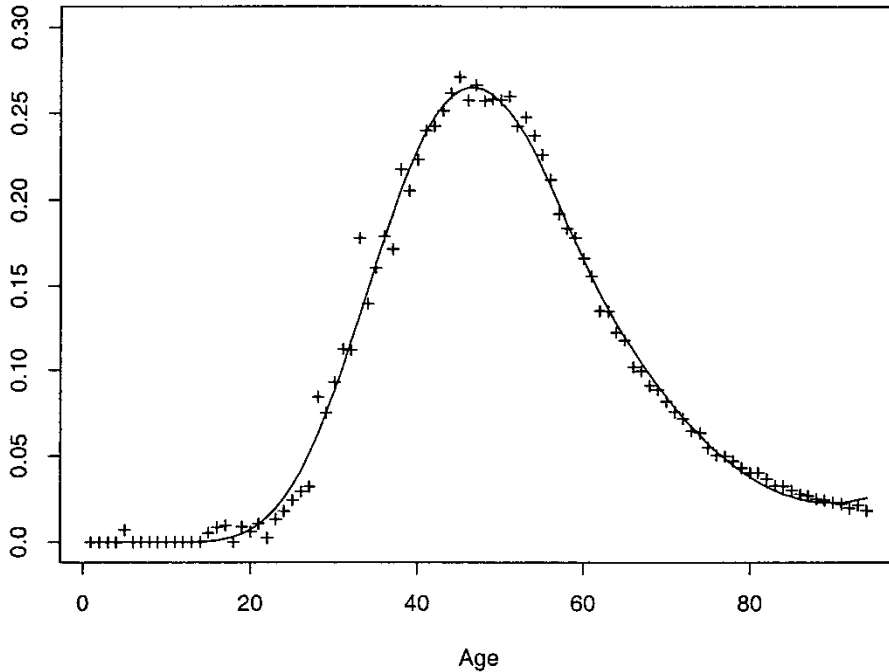
Fig. 5. Crude and graduated proportion of total deaths that are due to breast cancer and ovarian cancer, for females.

(a) the *family structure*: the numbers of sisters and aunts, their ages, and the mother's age; and

(b) the *family genotype*: the genotypes of the applicant and her mother, sisters and aunts.

From this, we can find the probability of any given family history at the time the applicant applies for insurance. Then, given the probabilities of different family structures, and the mutation frequencies, Bayes' Theorem gives the probabilities that an applicant with a given family history has one or more mutations, and hence estimates of insurance costs, given only the information available to the underwriter.

The family structure itself is relevant, but not generally considered in underwriting. For example, a woman with no sisters or aunts cannot have two affected relatives (ignoring male BC), even if she carried multiple mutations, whereas a woman with a large family is more likely to present a family history by chance.

Similar methods are often used in linkage analysis (Sham, 1998, Chapter 3). Iversen et al. (2000) used a similar procedure to estimate carrier probabilities of women in a large sample of affected families.

5.1. Family structure

We define 'family size' as the number of female relatives including the applicant, her mother, her sisters and her aunts, and denote it $M \geqslant 2$. Both maternal and paternal aunts are included, but we omit cousins, females of the grandparents' generation and beyond, and all male relatives.

For simplicity, we assume that all the applicant's sisters are the same age as she is, and the mother and all aunts are 30 years older. The number of possible family structures, in a family of size M, is then $M - 1$.

We label the family members $i = 1, 2, \ldots, M$, and adopt the convention that $i = 1$ is the applicant and $i = 2$ is her mother. We denote the age of the ith member at the birth of the applicant x_i ($x_1 = 0$ and $x_2 = 30$ always), so the family structure is represented by M and the vector:

$$X = (x_1, x_2, \ldots, x_M) \tag{14}$$

or the pair (M, X) The omission of male relatives means that we ignore male BC (which is rare, but linkage studies suggest a strong association with BRCA2 (Ford et al., 1998)).

5.2. The genotypes of family members

Given a family structure, we assume known the BRCA1 and BRCA2 genotypes of all its members. Each may have 0, 1 or 2 mutated copies of either gene, so there are up to 9^M family genotypes (actually slightly fewer, since some combinations are impossible.) However, we assume that one or two mutated copies of either gene confer the same risk (Section 4.4), so there are no more than 4^M family genotypes. This brings the computations within the reach of a fast computer.

We denote the four genotypes (0, 0), (0, 1), (1, 0) and (1, 1) where '1' in the first place indicates a BRCA1 mutation, and in the second place a BRCA2 mutation. Table 6 shows the frequencies of these genotypes given the 'low' and 'high' mutation frequencies of Parmigiani et al. (1998) ((c) and (b) respectively in Table 1). We denote the genotype of the ith member g_i, so the family genotype is the vector:

$$G = (g_1, g_2, \ldots, g_M). \tag{15}$$

Table 6. *The population frequencies of the four genotypes (0, 0), (0, 1), (1, 0) and (1, 1), given the low and high estimates of mutation frequencies from Parmigiani et al. (1998)*

Allele mutation frequencies	Population frequency of genotype			
	(0, 0)	(1, 0)	(0, 1)	(1, 1)
Low	0.998770525	0.000899501	0.000329675	0.000000299
High	0.997801689	0.001598401	0.000598951	0.000000959

This simple model does not attempt to distinguish different BRCA1 and BRCA2 mutations; a woman either has a mutation or not, and if she has, the effect does not depend on the precise mutation. This assumption is hard to avoid at this stage; it is usually made in linkage studies also (for example, Ford et al., 1998).

5.3. The definition of a family history

Any one of seven events can befall each relative, labelled $1, 2, \ldots, 7$, as in Fig. 1. Hence the life history of the ith relative at time t can be defined by the number $c_i(t) = 0, 1, \ldots, 7$ The family history is then the vector function of time $C(t) = (c_1(t), c_2(t), \ldots, c_M(t))$, or equivalently of the applicant's age x, $C(x) = (c_1(x), c_2(x), \ldots, c_M(x))$.

5.4. The distribution of family structures

Let $P[M, X]$ be the probability that a woman chosen at random has a family with size M and structure X. The estimation of a representative set of probabilities $P[M, X]$ is described in Section 7.1.

5.5. The conditional distribution of family genotype

$P[G \mid M, X]$ is the probability that the family genotype is G, given M and X. We can compute $P[G \mid M, X]$ under three assumptions:

(a) the applicant's four grandparents have mutually independent genotypes, sampled randomly from the gene frequencies;
(b) each aunt is equally likely to be a maternal or paternal aunt; and
(c) the usual rules of Mendelian inheritance apply to each gene.

Note that we must distinguish between homozygous and heterozygous mutation carriers here, in order to apply Mendel's laws, but once we have found the probabilities of all the family genotypes we can combine them, as in Section 5.2.

These calculations are computationally intensive, and determined the largest family size with which we could work ($M = 10$). Tabulating $P[G \mid M, X]$ for $M \leqslant 10$, and for high and low estimates of mutation frequencies, took almost a month on our fastest computer. Partly this is because we do not distinguish between maternal and paternal aunts in the family structure X, but we need to do so here by considering all possible combinations, given the total number of aunts. However, the simpler family structure makes the calculation of other probabilities faster.

5.6. The conditional distribution of family history

$P[C(t) \mid M, X, G]$ is the probability that, in a family of size M with ages X and genotypes G, the family history is $C(t)$ at time t after the applicant's birth.

Let the future lifetime of the ith family member be a random variable T_i. $T_1, T_2, \ldots, T_M$ are not in general unconditionally independent, because their distributions may be related through shared genes. Onset of BC or OC in one member reveals information that alters the conditional distributions of the others' lifetimes.

An extreme example would be identical twins. At birth, their future lifetimes T_1 and T_2 are not independent, since if the first were to develop BC (say), the conditional distribution of T_2 would be changed. However, if their common genotype were known, then conditioned on that, T_1 and T_2 might be independent. In our model, we assume that G is known, and that conditioned on G, $T_1, T_2, \ldots, T_M$ are independent. We also assume that the applicant's mother is free of BC and OC at age 30. Then the calculation of $P[C(t) \mid M, X, G]$ splits into two cases:

$$P[C(t) \mid M, X, G]$$

$$= \frac{{}^{g_2}p_{0,30+t}^{00} + {}^{g_2}p_{0,30+t}^{07} - {}^{g_2}p_{0,30}^{07}}{{}^{g_2}p_{0,30}^{00}} \prod_{\substack{i \neq 2 \\ c_i(t)=0}} ({}^{g_i}p_{0,x_i+t}^{00} + {}^{g_i}p_{0,x_i+t}^{07}) \prod_{\substack{i \neq 2 \\ c_i(t)>0}} {}^{g_i}p_{0,x_i+t}^{0c_i(t)} \qquad (16)$$

if $c_2(t) = 0$, and:

$$P[C(t) \mid M, X, G] = \frac{{}^{g_2}p_{0,30+t}^{0c_2(t)} - {}^{g_2}p_{0,30}^{0c_2(t)}}{{}^{g_2}p_{0,30}^{00}} \prod_{\substack{i \neq 2 \\ c_i(t)=0}} ({}^{g_i}p_{0,x_i+t}^{00} + {}^{g_i}p_{0,x_i+t}^{07}) \prod_{\substack{i \neq 2 \\ c_i(t)>0}} {}^{g_i}p_{0,x_i+t}^{0c_i(t)}$$
$$(17)$$

if $c_2(t) > 0$. In each case, the first product is over those family members (including the applicant) who have not had BC or OC, and the second product is over those family members who have.

5.7. The distribution of the applicant's genotype, given family history

Finally, we reach the probabilities relevant to underwriting, namely those of the form:

P[Applicant age x has genotype g | Some observed information].

First assume that M, X and $C(x)$ are observed. Then let $\mathcal{G}$ be the set of all possible family genotypes and $\mathcal{G}(g)$ be the subset of $\mathcal{G}$ such that $g_1 = g$. By Bayes' Theorem:

$$P[\text{Applicant has genotype } g \mid M, X, C(x)] = \frac{\sum_{G \in \mathcal{G}(g)} P[C(x) \mid M, X, G] P[G \mid M, X]}{\sum_{G \in \mathcal{G}} P[C(x) \mid M, X, G] P[G \mid M, X]}.$$
$$(18)$$

However, usually less is observed, for example the family structure may be unknown. (The reason for assuming G known up to this point is not because that is likely, but to obtain the conditional independence in Section 5.6.) Then the conditional distribution of the applicant's genotype must be built out of the elementary probabilities $P[M, X]$, $P[G \mid M, X]$ and $P[C(t) \mid M, X, G]$, using Bayes' Theorem. We give one example. Suppose M and X are unknown, and that the underwriter only knows that two or more of the applicant's relatives have had BC

before age 50. Define $\mathscr{C}(M, X)$ to be the set of all family histories, in families of size M and with structure X, in which this has occurred. In some cases, $\mathscr{C}(M, X) = \varnothing$; define $\mathscr{H} = \{(M, X): \mathscr{C}(M, X) \neq \varnothing\}$. Then the probability that the applicant, age x, has genotype g is:

$$\sum_{(M, X)\in\mathscr{H}} \left[\frac{\displaystyle\sum_{G\in\mathscr{G}(g)} \left(\sum_{C(x)\in\mathscr{C}(M, X)} P[C(x)\,|\,M, X, G] \right) P[G\,|\,M, X]}{\displaystyle\sum_{G\in\mathscr{G}} \left(\sum_{C(x)\in\mathscr{C}(M, X)} P[C(x)\,|\,M, X, G] \right) P[G\,|\,M, X]} \right] P[M, X\,|\,(M, X)\in\mathscr{H}].$$

(19)

6. THE APPLICANT'S GENOTYPE PROBABILITIES, KNOWN FAMILY STRUCTURE

Here we give some examples to illustrate important features. Since we are mainly interested in the application of these probabilities in Part II, we limit ourselves to a few observations.

(a) In Table 7 the applicant is age 30 with 4 sisters and no aunts. Onset at age 65 or over can be ignored. We suppose that 0, 1, 2, 3, 4 or 5 relatives are affected in each of the four remaining BC and OC states. We use high estimates of mutation frequencies (Parmigiani et al. (1998) (b) in Table 1). Points to note are:
   (1) BC or OC under age 50 presents a significant risk of a BRCA1 mutation, rising steeply once more than one relative is affected.
   (2) BC or OC over age 50 can only be the mother, in these families. The mutation probabilities are much lower, but in the case of OC, not negligible.
   (3) BRCA2 mutations are much less likely than BRCA1 mutations, in these families. This is because of the higher ages at which BRCA2 presents the greatest risk; strong histories of BC at ages over 50, for example, do lead to higher BRCA2 mutation probabilities.

Table 7. *Applicant's genotype probabilities with four sisters and no aunts. Applicant age 30. Uses high estimates (Parmigiani et al. (1998) (b) in Table 1) of mutation frequencies*

| Cases of | | Genotype | | | | Cases of | | Genotype | | | |
| BC | BC | | | | | OC | OC | | | | |
<50	50–65	(0, 0)	(1, 0)	(0, 1)	(1, 1)	<50	50–65	(0, 0)	(1, 0)	(0, 1)	(1, 1)
0	0	0.999	0.001	0.000	0.000	0	0	0.999	0.001	0.000	0.000
1	0	0.977	0.019	0.004	0.000	1	0	0.967	0.031	0.002	0.000
0	1	0.989	0.005	0.006	0.000	0	1	0.984	0.012	0.004	0.000
2	0	0.791	0.207	0.002	0.000	2	0	0.768	0.231	0.001	0.000
3	0	0.541	0.458	0.001	0.000	3	0	0.546	0.454	0.000	0.000
4	0	0.506	0.493	0.000	0.000	4	0	0.508	0.492	0.000	0.000
5	0	0.502	0.497	0.000	0.000	5	0	0.502	0.497	0.000	0.000

    (4) We find (not shown in the table) that the probabilities depend hardly at all on the number of sisters; given the rarity of BC and OC at these ages, the fact that one relative is affected is much more significant than the numbers who are unaffected.

(b) In Tables 8 and 9 we suppose that the family size M is 4 or 6, and consider all possible family structures (sisters and aunts). In Table 8 either one relative has BC before age 50, or two or more have had BC before age 50; in each case, it is not known which are the affected relatives. The applicant is age 30 and we assume high estimates of the mutation frequencies. In Table 9 we assume two relatives have had BC before age 50, and we also show the effect of lower mutation frequencies (Parmigiani et al. (1998) (c) in Table 1). Points to note are:

    (1) If the applicant has sisters, the probability that she has a mutation is significantly increased.

Table 8. *The applicant's genotype probabilities, given one relative, or at least two relatives, with BC before age 50, for $M = 4$ or 6. Applicant age 30. High estimates of mutation frequencies*

Number of		1 Case of BC < 50 genotype				2 or More cases of BC < 50 genotype			
Sisters	Aunts	(0, 0)	(1, 0)	(0, 1)	(1, 1)	(0, 0)	(1, 0)	(0, 1)	(1, 1)
2	0	0.976	0.020	0.004	0.000	0.782	0.217	0.002	0.000
1	1	0.986	0.012	0.002	0.000	0.865	0.123	0.012	0.000
0	2	0.990	0.008	0.002	0.000	0.903	0.087	0.009	0.000
4	0	0.977	0.019	0.004	0.000	0.784	0.214	0.002	0.000
3	1	0.986	0.012	0.002	0.000	0.863	0.127	0.010	0.000
2	2	0.990	0.008	0.002	0.000	0.902	0.089	0.009	0.000
1	3	0.992	0.006	0.001	0.000	0.924	0.069	0.007	0.000
0	4	0.994	0.005	0.001	0.000	0.938	0.056	0.006	0.000

Table 9. *The applicant's genotype probabilities, given two relatives with BC before age 50, for $M = 4$ or 6. Applicant age 30. High and low estimates of mutation frequencies*

Number of		High mutation frequencies genotype				Low mutation frequencies genotype			
Sisters	Aunts	(0, 0)	(1, 0)	(0, 1)	(1, 1)	(0, 0)	(1, 0)	(0, 1)	(1, 1)
2	0	0.787	0.211	0.002	0.000	0.853	0.146	0.001	0.000
1	1	0.869	0.119	0.012	0.000	0.916	0.076	0.007	0.000
0	2	0.910	0.081	0.009	0.000	0.944	0.051	0.006	0.000
4	0	0.791	0.207	0.002	0.000	0.856	0.142	0.001	0.000
3	1	0.869	0.121	0.010	0.000	0.916	0.077	0.007	0.000
2	2	0.910	0.082	0.009	0.000	0.944	0.051	0.005	0.000
1	3	0.933	0.061	0.007	0.000	0.959	0.037	0.004	0.000
0	4	0.947	0.047	0.006	0.000	0.968	0.028	0.003	0.000

(2) The mutation probabilities depend almost entirely on the number of aunts.

(3) Knowing that two or more relatives are affected, rather than that exactly two are affected, increases the probability of a mutation, but only a little. Most of the information in the family history, it seems, is conveyed by the first few cases. Note that this is quite consistent with Table 7; there the probability that the applicant has a gene mutation is very much higher as the number of affected relatives increases, but in equation (19) the family histories are weighted by the probabilities $P[C(x) \mid M, X, G]$.

(4) The BRCA2 mutation probabilities show some odd features, because of the dominance of BRCA1 risk below age 50. They are *lower* if two relatives are affected than if one relative is affected, and if the applicant has only sisters.

(5) Lower mutation frequencies give significantly lower mutation probabilities. Thus, good estimates of mutation frequencies are important; unfortunately, estimating the frequencies of rare mutations is very difficult.

The BRCA1 gene, with a higher frequency of mutations and higher risk at younger ages, is the more important risk factor. Moreover, for a given number of affected relatives, the risk varies quite widely across different family structures. We could look at more examples, but as these are only steps along the way to estimating insurance costs, we press on.

7. THE APPLICANT'S GENOTYPE PROBABILITIES, UNKNOWN FAMILY STRUCTURE

7.1. Estimating the distribution of family structures M, X

Often, the only information will be the number of affected relatives, with M and X unknown. Then we need estimates of $P[M, X]$ in order to use Equation (19).

Define:

D^m = Number of daughters borne by the applicant's mother
D^{mm} = Number of daughters borne by the applicant's maternal grandmother
D^{fm} = Number of daughters borne by the applicant's paternal grandmother
S^{fm} = Number of sons borne by the applicant's paternal grandmother.

We model these as random variables. In general, let D be the total number of daughters and S the total number of sons born to a given woman, and $C = D + S$ the total number of her children. We suppose the probability that a child is male is about $1.06/2.06$ (see for example Coleman & Salt (1992)), so:

$$P[D = d] = \sum_{c=d}^{c=\infty} P[C = c]\binom{c}{d}\left(\frac{1}{2.06}\right)^d\left(\frac{1.06}{2.06}\right)^{c-d}, \tag{20}$$

$$P[D = d, S > 0] = \sum_{c=d+1}^{c=\infty} P[C = c]\binom{c}{d}\left(\frac{1}{2.06}\right)^d\left(\frac{1.06}{2.06}\right)^{c-d}, \tag{21}$$

$$P[S > 0] = 1 - \sum_{c=0}^{c=\infty} P[C = c]\left(\frac{1}{2.06}\right)^c, \tag{22}$$

where, in practice, the summation will stop at some number of children depending on the available data (which means that D and S will not be independent). The applicant has s sisters if $D^m = s + 1$, a^m maternal aunts if $D^{mm} = a^m + 1$, and a^f paternal aunts if $D^{fm} = a^f$ and $S^{fm} > 0$, so the probability that the applicant has s sisters and a aunts is:

$$\frac{P[D^m = s + 1]}{P[D^m > 0]}\left(\sum_{a^m=0}^{a^m=a} \frac{P[D^{mm} = a^m + 1]}{P[D^{mm} > 0]}\frac{P[D^{fm} = a - a^m, S^{fm} > 0]}{P[S^{fm} > 0]}\right) \tag{23}$$

and we can compute $P[M, X]$ given estimates of $P[C = c]$ $(c \geqslant 0)$ appropriate for the applicant's mother and grandmothers. These will not be the same; family size has changed significantly over time. Since our examples are mostly of applicants age 30, we should consider mothers born in about 1940 and grandmothers born in about 1910. Table 10 shows the numbers of children (or, in the case of the 1940–44 birth cohorts, the expected final numbers of children) based on interviews in 1987 (Shaw, 1990). From these, and the average family sizes, the average number of children, conditional on there being four or more, is 4.87 for the 1930–34 cohort, 4.80 for the 1935–39 cohort, and 4.72 for the 1940–44 cohort.

Table 11 shows the numbers of children according to year of marriage, based on the 1971 census; we have chosen years 1931–35 and 1961–65 as representative of the cohorts that interest us. The latter cohort was still having children in 1971. For our purposes:

Table 10. *Distribution of final or expected numbers of children born to women born in England and Wales in 1930–44. (Source: Shaw (1990))*

Birth cohort	Number of children					Average family size
	0	1	2	3	$\geqslant 4$	
	%	%	%	%	%	
1930–34	13	13	29	22	23	2.49
1935–39	10	11	35	24	20	2.49
1940–44	9	13	40	20	18	2.38

Table 11. *Distribution of numbers of children according to year of mother's marriage, England and Wales. (Source: O.P.C.S. (1983))*

Year of marriage	Number of children							
	0	1	2	3	4	5	6	$\geqslant 7$
	%	%	%	%	%	%	%	%
1931–35	17.5	27.0	26.6	14.3	7.0	3.50	1.80	2.30
1961–65	11.3	20.5	44.4	17.5	4.8	1.10	0.30	0.10

(a) we assume that $C \leqslant 7$;
(b) we take the first line of Table 11 as the distribution of C for the applicant's grandmothers; and
(c) we take the third line of Table 10 as the distribution of C for the applicant's mother, and distribute the 18% with four or more children in proportion to the numbers given in the Fertility Survey from the 1971 Census (O.P.C.S., 1983), resulting in 13.7%, 3.14%, 0.86% and 0.29% of women with 4, 5, 6 and 7 children respectively.

The resulting distribution of family structures is given in Table 12. In practice, for computational reasons we are limited to $M \leqslant 10$, that is, no more than 8 sisters and aunts. These families account for 99.86% of the total probability.

7.2. Probabilities of gene mutations with unknown M and X

We can now calculate probabilities of the applicant's genotype given a family history, but in ignorance of the family structure (M, X), using Equation (19) and Table 12. Our calculations are limited to $M \leqslant 10$, so we necessarily truncate the

Table 12. *Distribution of the number of the applicant's sisters and aunts*

Sisters	Aunts	Probability	Sisters	Aunts	Probability	Sisters	Aunts	Probability	Sisters	Aunts	Probability
0	0	0.12452712	2	0	0.02217054	4	0	0.00064970	6	0	0.00000599
0	1	0.17178831	2	1	0.03058482	4	1	0.00089628	6	1	0.00000827
0	2	0.12460245	2	2	0.02218395	4	2	0.00065010	6	2	0.00000600
0	3	0.06975886	2	3	0.01241972	4	3	0.00036396	6	3	0.00000336
0	4	0.03418573	2	4	0.00608635	4	4	0.00017836	6	4	0.00000165
0	5	0.01485461	2	5	0.00264468	4	5	0.00007750	6	5	0.00000071
0	6	0.00555446	2	6	0.00098890	4	6	0.00002898	6	6	0.00000027
0	7	0.00171508	2	7	0.00030535	4	7	0.00000895	6	7	0.00000008
0	8	0.00047017	2	8	0.00008371	4	8	0.00000245	6	8	0.00000002
0	9	0.00011468	2	9	0.00002042	4	9	0.00000060	6	9	0.00000001
0	10	0.00002303	2	10	0.00000410	4	10	0.00000012	6	10	0.00000000
0	11	0.00000328	2	11	0.00000058	4	11	0.00000002	6	11	0.00000000
0	12	0.00000024	2	12	0.00000004	4	12	0.00000000	6	12	0.00000000
1	0	0.07516976	3	0	0.00480188	5	0	0.00008109			
1	1	0.10369858	3	1	0.00662432	5	1	0.00011186			
1	2	0.07521523	3	2	0.00480479	5	2	0.00008114			
1	3	0.04210935	3	3	0.00268997	5	3	0.00004543			
1	4	0.02063593	3	4	0.00131823	5	4	0.00002226			
1	5	0.00896686	3	5	0.00057281	5	5	0.00000967			
1	6	0.00335290	3	6	0.00021418	5	6	0.00000362			
1	7	0.00103530	3	7	0.00006614	5	7	0.00000112			
1	8	0.00028382	3	8	0.00001813	5	8	0.00000031			
1	9	0.00006922	3	9	0.00000442	5	9	0.00000007			
1	10	0.00001390	3	10	0.00000089	5	10	0.00000001			
1	11	0.00000198	3	11	0.00000013	5	11	0.00000000			
1	12	0.00000015	3	12	0.00000001	5	12	0.00000000			

distribution in Table 12. We denote the maximum family size used here $\hat{M}$. Table 13 shows that we lose little by truncation; it shows the probabilities in respect of an applicant age 30, who has two relatives with BC before age 50, truncating the distribution in Table 13 at $\hat{M} = 3, 4, \ldots, 10$. Also shown are:

(a) the proportion of the full probability of Table 12 included in the truncated distributions; and
(b) the proportion $P[M \leqslant \hat{M}, (M, X) \in \mathcal{H}]$ of the full probability of Table 12 included in the conditional truncated distributions $P[M, X \mid (M, X) \in \mathcal{H}]$ used in Equation (19) (that is, conditioning on families in which the given history is possible).

So, for example, if $\hat{M} = 10$, we include family structures that account for almost 99.9% of the total probability in Table 12, but about 12.6% of that probability attaches to family structures for which the event 'two relatives with BC before age 50' is impossible. In this example, $\hat{M} = 9$ or $\hat{M} = 8$ would give quite acceptable genotype probabilities, bearing in mind the uncertainty of most of the components of the model. Using smaller values of $\hat{M}$ gives a significant speed increase (the last line of Table 13 took more than a day to compute) and doing so will always overstate the mutation probabilities; the probability that a given history has a genetic basis is higher in smaller families. In what follows, we use $\hat{M} = 9$ only, checks with $\hat{M} = 7$ or 8 confirming that this was satisfactory.

Table 14 shows mutation probabilities for applicants age 30 and 50 with:

(a) a *known* number of relatives (1-5) with BC before age 50;
(b) an *unknown* number of relatives ($\geqslant 2$) with BC before age 50. If two or more relatives have BC, rather than exactly two, a much larger number of family histories has to be included in $\mathcal{H}$ in Equation (19). It might be reasonable in

Table 13. *The effect of the maximum family size $\hat{M}$ on probabilities of the applicant's genotype, given two relatives with BC before age 50 and unknown (M, X). Applicant age 30. Uses high estimates (Parmigiani et al. (1998) (b) in Table 1) of mutation frequencies*

Maximum family size $\hat{M}$	$P[M \leqslant \hat{M}]$	$P[M \leqslant \hat{M}, (M, X) \in \mathcal{H}]$	Probability that applicant's genotype is			
			(0, 0)	(1, 0)	(0, 1)	(1, 1)
3	0.371485	0.246958	0.840002	0.150439	0.009416	0.000144
4	0.621957	0.497430	0.859793	0.130586	0.009498	0.000123
5	0.802318	0.677790	0.872152	0.118564	0.009174	0.000109
6	0.908071	0.783543	0.879421	0.111602	0.008876	0.000101
7	0.961763	0.837236	0.883358	0.107866	0.008679	0.000097
8	0.985829	0.861301	0.885273	0.106059	0.008573	0.000095
9	0.995313	0.870786	0.886090	0.105292	0.008524	0.000094
10	0.998610	0.874083	0.886394	0.105007	0.008505	0.000094

Table 14. *The applicant's genotype probabilities given family history (BC before age 50 only), unknown (M, X). Maximum family size $\hat{M} = 9$. Uses high estimates (Parmigiani et al. (1998) (b) in Table 1) of mutation frequencies*

Applicant's age	BC < 50	High mutation frequencies genotype				Low mutation frequencies genotype			
		(0, 0)	(1, 0)	(0, 1)	(1, 1)	(0, 0)	(1, 0)	(0, 1)	(1, 1)
30	1	0.986	0.012	0.002	0.000	0.992	0.007	0.001	0.000
30	2	0.886	0.105	0.009	0.000	0.927	0.068	0.005	0.000
30	3	0.662	0.325	0.012	0.000	0.714	0.276	0.010	0.000
30	4	0.552	0.439	0.008	0.001	0.566	0.426	0.007	0.000
30	5	0.523	0.471	0.005	0.001	0.528	0.468	0.004	0.001
30	⩾2	0.883	0.109	0.009	0.000	0.924	0.070	0.005	0.000
50	1	0.988	0.010	0.002	0.000	0.993	0.006	0.001	0.000
50	2	0.895	0.095	0.010	0.000	0.933	0.060	0.006	0.000
50	3	0.666	0.316	0.018	0.000	0.721	0.264	0.014	0.000
50	4	0.547	0.439	0.013	0.001	0.563	0.425	0.012	0.000
50	5	0.519	0.472	0.008	0.001	0.523	0.469	0.008	0.001
50	⩾2	0.888	0.101	0.010	0.000	0.929	0.065	0.006	0.000

practice to use smaller values of $\hat{M}$, even $\hat{M} = 7$ (depending on the details of the problem); and

(c) high and low estimates of the mutation frequencies.

The applicant's age has little effect. For an applicant age 30, knowing that two or more relatives had BC before age 50, rather than exactly two, increases the probability that she has a BRCA1 mutation by about 3.5%. For an applicant age 50, the increase is about 7%. However, knowing that more than two relatives had BC before age 50 makes a very great difference. Comparing Table 14 with Tables 7 to 9, we see that significant errors may be made if the family structure is unknown or is ignored.

For the sake of brevity, we omit examples of other family histories.

8. THE EFFECT OF LOWER PENETRANCE

So far we have used incidence rates of BC and OC from studies of families with strong histories of BC and/or OC. There, BRCA1 and BRCA2 are highly penetrant, for example the lifetime risk of BC, given a BRCA1 mutation, may be as high as 85%. However, it is likely that mutations are less penetrant in families not selected specifically because of their BC and/or OC history; Hopper et al. (1999) studied women with BC in Australia, and suggested penetrance of 40% or less.

The best evidence available is that used in Sections 4.2 and 4.3, so here we use this to give the 'shape' of the age-related incidence rates of BC and OC, but we reduce the excess incidence rates of BC and OC (that is, the difference between the modelled incidence rates of the at-risk and (0, 0) genotypes) to:

(a) 50% of the values in Sections 4.2 and 4.3; this corresponds to a reduction in the penetrance of BRCA1 roughly from 85% to 60%; or
(b) 25% of the values in Sections 4.2 and 4.3; this corresponds to a reduction in the penetrance of BRCA1 roughly from 85% to 40%.

Table 15 shows genotype probabilities for women age 30 with two relatives with BC before age 50, with family size $M = 4$ or 6, allowing for these reduced incidence rates. We assume high estimates of mutation frequencies (compare with Table 9). In all cases, the applicant's mutation probabilities are reduced significantly. The reduction is greatest (by a factor or 4 or 5) when the affected relatives are sisters, and the mutation probabilities greatest. Reduced incidence will also reduce insurance costs, which will magnify the effect of the assumed penetrance on the theoretical premium loadings in respect of mutation carriers. We shall consider this in Part II.

Table 16 shows an example of the effect of reduced penetrance when the family structure is unknown (a history of BC before age 50 with high estimates of mutation frequencies; other examples show similar features). Again, the applicant's mutation probabilities are greatly reduced, especially if she is older.

The results of this section indicate the importance of the penetrance in assessing genetic risk. Even reducing it from 85% to 40% (still quite high) greatly reduces the risk. Therefore, mutation probabilities based on known high-risk families probably are not applicable to mutation carriers in families without such a history; that is quite strongly suggested by the emerging genetics research. In insurance terms, the applicants at greatest risk are exactly those who would have been identified through their family history; the mutation probabilities calculated in previous sections of this paper might apply to them. In the absence of a family history, such high risks should not be assumed.

Table 15. *The applicant's genotype probilities, given two relatives with BC before age 50, for* $M = 4$ *or 6. Applicant age 30. High estimates of mutation frequencies. BC and OC excess incidence rates of at-risk genotypes at 50% and 25% of previous estimates*

Number of		Excess incidence rates 50% of observed genotype				Excess incidence rates 25% of observed genotype			
Sisters	Aunts	(0, 0)	(1, 0)	(0, 1)	(1, 1)	(0, 0)	(1, 0)	(0, 1)	(1, 1)
2	0	0.897	0.101	0.002	0.000	0.959	0.039	0.002	0.000
1	1	0.935	0.059	0.006	0.000	0.972	0.025	0.003	0.000
0	2	0.952	0.043	0.005	0.000	0.979	0.020	0.002	0.000
4	0	0.899	0.099	0.002	0.000	0.960	0.039	0.001	0.000
3	1	0.935	0.060	0.005	0.000	0.972	0.026	0.003	0.000
2	2	0.952	0.044	0.004	0.000	0.978	0.020	0.002	0.000
1	3	0.962	0.034	0.004	0.000	0.981	0.017	0.002	0.000
0	4	0.969	0.028	0.003	0.000	0.983	0.015	0.002	0.000

Table 16. *The effect of the family history (BC before age 50 only) on probabilities of the applicant's genotype, unknown (M, X). Maximum family size $\hat{M} = 9$. Uses high estimates (Parmigiani et al. (1998) (b) in Table 1) of mutation frequencies. BC and OC excess incidence rates of at-risk genotypes at 50% and 25% of previous estimates*

Applicant's age	BC < 50	Excess incidence rates 50% of observed genotype				Excess incidence rates 25% of observed genotype			
		(0, 0)	(1, 0)	(0, 1)	(1, 1)	(0, 0)	(1, 0)	(0, 1)	(1, 1)
30	1	0.990	0.009	0.002	0.000	0.993	0.006	0.001	0.000
30	2	0.943	0.053	0.004	0.000	0.975	0.023	0.002	0.000
30	3	0.795	0.197	0.007	0.000	0.920	0.076	0.004	0.000
30	4	0.627	0.366	0.007	0.001	0.806	0.189	0.005	0.000
30	5	0.551	0.443	0.004	0.001	0.672	0.323	0.004	0.001
30	⩾2	0.942	0.054	0.004	0.000	0.974	0.023	0.002	0.000
50	1	0.994	0.005	0.001	0.000	0.995	0.004	0.001	0.000
50	2	0.962	0.034	0.004	0.000	0.979	0.019	0.002	0.000
50	3	0.852	0.139	0.009	0.000	0.930	0.066	0.004	0.000
50	4	0.711	0.279	0.010	0.000	0.826	0.167	0.007	0.000
50	5	0.645	0.347	0.007	0.001	0.704	0.288	0.007	0.000
50	⩾2	0.960	0.035	0.004	0.000	0.978	0.019	0.002	0.000

If this conclusion is confirmed by future epidemiological research, we can say that genetic testing for BRCA1 and BRCA2 is not likely to create new groups of disadvantaged women, unable to obtain or afford insurance (the 'genetic underclass' often referred to by the media) as long as family history continues to be used in underwriting.

9. CONCLUSIONS

9.1. Estimation of the applicant's mutation probabilities

(a) The applicant's mutation risk depends strongly on the family structure; it is greatest if sisters are affected.
(b) Consequently, if detailed information is not obtained, the assessed risk will be an average across all possible family structures, and/or combinations of affected relatives, and could be quite inaccurate.
(c) Two pieces of epidemiological data are crucial:
  (1) the population frequencies of mutations; and
  (2) penetrance with and without a strong family history of BC and OC. Neither of these has been estimated very reliably.

9.2. Penetrance estimation

The story of the BRCA1 and BRCA2 genes may exemplify the course of genetical research. The existence of families with strong histories of BC and OC suggested the existence of one or more responsible genes, by implication highly penetrant *in these families*. This led to the discovery of BRCA1 and BRCA2 mutations, but it

was not possible to infer that they would be equally penetrant in other families, and it now appears that they might not be. Early studies may often select at-risk populations, for obvious reasons, but the resulting risk estimates cannot automatically be applied outside the selected population. It is conceivable that genetic test information will not be very informative outside the study population, in which case:

(a) only those at risk will be likely to be tested;
(b) the test results will only be relevant for insurance in the presence of other risk factors which are known anyway.

Neither insurers nor the media should assume that the discovery of a new, apparently predictive, gene mutation implies the creation of new insurance risk categories.

We considered penetrances lower than those observed in high-risk families (which was speculative, but consistent with the expectations of geneticists (Antoniou et al., 2000)) and found the applicant's mutation probabilities much reduced. Insurance costs will also be reduced (see Part II).

9.3. The importance of population mutation frequencies

We considered higher and lower estimates of the frequency of mutations in the population, with quite a significant effect, higher frequencies increasing the risk. They are irrelevant if a genetic test result is known, but good estimates would be needed to draw reliable conclusions about an applicant's risk, given her family history.

Studying different population frequencies, however, helps to answer a different question. It is possible that more genes will be found in which rare mutations lead to BC and OC. Suppose, for example, one were found with the same penetrance as BRCA1 mutations, on a different chromosome from BRCA1. Ignoring the tiny possibility of both mutations being present, the impact would be exactly the same as a higher population frequency of BRCA1 mutations. So the high estimates of mutation frequencies that we have used do not necessarily suggest an upper bound to genetic risk, even if low estimates should be confirmed by future research; we have not explicitly allowed for the discovery of BRCA3.

9.4. Scope for extending the methodology

(a) We had just two genes to consider, and we could assume that they acted independently. Larger numbers of genes pose two distinct problems:
   (1) the number of family genotypes grows exponentially with the number of genes; and
   (2) there may be interactions between genes, which does not worsen the combinatorics but makes the estimation of intensities much harder.
(b) Although both BRCA1 and BRCA2 have hundreds of mutations, we assumed that each had only two alleles, and further that possessing two mutations in either had the same effect as possessing one. The number of individual and

family genotypes grows explosively as the number of mutations or alleles increases.

We used brute force calculation, and even this simple case stretched our computing resources. The time-consuming aspect is lack of information; if the family structure and history are unknown, the number of cases to be included in Equation (19), increases in number exponentially with the largest family size used. For our purposes, it was appropriate to use large families (up to ten members) even though some calculations took weeks. Using smaller family sizes will always overstate the genetic risk, and our results show that, in this particular case, the error is usually small, compared with the width of typical underwriting bands.

Computers get ever faster, but the growth in genetic knowledge is keeping pace, and brute force will quickly meet its intrinsic limits. Further progress may rely on brute force plus algorithms to eliminate family histories and structures that contribute little to the genetic risk or insurance costs. One useful approach (Antoniou et al., 2000) assumes that some mutations dominate others, effectively ignoring multiple mutations; with K genes this gives $K+1$ individual genotypes, instead of 3^K allowing for homozygotes or 2^K treating homozygotes and heterozygotes alike. Assuming that BRCA1 dominates BRCA2 (as did Antoniou et al. (2000)) we found only a small change ($<1\%$) in the probability that the applicant has no mutations, but some large percentage changes to some probabilities of a BRCA2 mutation (though these were small in magnitude).

ACKNOWLEDGEMENTS

REFERENCES

Antoniou, A. C., Gayther, S. A., Stratton, J. F., Ponder, A. J. P. & Easton, D. F. (2000). Risk models for breast and ovarian cancer. *Genetic Epidemiology* **18,** 173–190.

Claus, E., Risch, N. & Thompson, W. (1994). Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer* **73** (3), 645–651.

Coleman, D. & Salt, J. (1992). *The British population*. Oxford University Press.

Easton, D. F., Bishop, D. T., Ford, D., Crockford, G. P. & The Breast Cancer Linkage Consortium (1993). Genetic linkage analysis in familial breast and ovarian cancer. *American Journal of Human Genetics* **52,** 678–701.

Easton, D. F., Ford, D., Bishop, D. T. & The Breast Cancer Linkage Consortium (1995). Breast and ovarian cancer incidence in BRCA1-mutation carriers. *American Journal of Human Genetics* **56,** 265–271.

Feuer, E., Wun, L., Boring, C., Flanders, W., Timmel, M. & Tong, T. (1993). The lifetime risk of developing breast cancer. *Journal of the National Cancer Institute* **85,** 892–897.

Fischer, E.-P., & Berberich, K. (1999). *Impact of modern genetics on insurance*. Publications of the Cologne Re, No. 42.

Ford, D., Easton, D. F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D. T., Weber, B., Lenoir, G., Chang-Claude, J., Sobol, H., Teare, M. D., Struewing, J., Arason, A., Scherneck, S., Peto, J., Rebbeck, T. R., Tonin, P., Neuhausen, S., Barkardottir, R., Eyfjord, J., Lynch, H., Ponder,

B. A. J., Gayther, S. A., Birch, J. M., Lindblom, A., Stoppa-Lyonnet, D., Bignon, Y., Borg, A., Hamann, U., Haites, N., Scott, R. J., Maugard, C. M., Vasen, H., Seitz, S., Cannon-Albright, L. A., Schofield, A., Zelada-Hedman, M. & The Breast Cancer Linkage Consortium (1998). Genetic heterogeneity and penetrance analysis of the BRC1 AND BRCA2 genes in breast cancer families. *American Journal of Human Genetics* **62,** 676–689.

Hopper, J. L., Southey, M. C., Dite, G. S., Jolley, D. J. Giles, G. G., McCredie, M. R. E., Easton, D. F., Venter, D. J. & The Australian Breast Cancer Family Study (1999). *Population-based estimates of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in BRCA1 and BRCA2.* Preprint, University of Melbourne Centre for Genetic Epidemiology.

Iversen, E. S., Parmigiani, G., Berry, D. A. & Schildkraut, J. M. (2000). Genetic susceptibility and survival: Application to breast cancer. *Journal of the American Statistical Association* **95,** 28–42.

Lemaire, J., Subramanian, K., Armstrong, K. & Asch, D. A. (2000). Pricing term insurance in the presence of a family history of breast or ovarian cancer. *North American Actuarial Journal* **4,** 75–87.

Macdonald, A. S. (1997). How will improved forecasts of individual lifetimes affect underwriting? *Philosophical Transactions of the Royal Society* B **352**, 1067–1075, and (with discussion) *British Actuarial Journal* **3**, 1009–1025 and 1044–1058.

Macdonald, A. S. (1999). Modeling the impact of genetics on insurance. *North American Actuarial Journal* **3** (1), 83–101.

O.N.S. (1999). *Cancer 1971–1997 CD-ROM.* Office for National Statistics, London.

O.P.C.S. (1983). *Fertility report from the 1971 census. Series DS No. 5.* Office of Population Censuses and Surveys.

O.P.C.S. (1991). *1990 mortality statistics: Cause. Series DH 12 No. 17.* Office of Population Censuses and Surveys.

O.P.C.S. (1993a). *1991 mortality statistics: Cause. Series DH 12 No. 18.* Office of Population Censuses and Surveys.

O.P.C.S. (1993b). *1992 mortality statistics: Cause. Series DH 12 No. 19.* Office of Population Censuses and Surveys.

Parmigiani, G., Berry, D. & Aguilar, O. (1998). Determining carrier probabilities for breast cancer susceptibility genes BRCA1 and BRCA2. *American Journal of Human Genetics* **62,** 145–158.

Pasternak, J. J. (1999). *An introduction to human molecular genetics.* Fitzgerald Science Press, Bethesda, Maryland.

Quinn, M. & Allen, E. (1995). Changes in incidence and mortality from breast cancer in England and Wales since the introduction of screening. *British Medical Journal* **311,** 1391–1395.

Sham, P. (1998). *Statistics in human genetics.* Arnold.

Shaw, C. (1990). Fertility assumptions for 1989-based population projections for England and Wales. *Population Trends* **61,** 17–23.

Strachan, T. & Read, A. P. (1999). *Human molecular genetics* (second edn). BIOS Scientific Publishers, Oxford.

Struewing, J. P., Hartge, P., Wacholder, S., Baker, S. M., Berlin, M., McAdams, M., Timmerman, M. M., Brody, L. C. & Tucker, M. A. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *The New England Journal of Medicine* **336,** 1401–1408.

Subramanian, K., Lemaire, J., Hershey, J. C., Pauly, M. V., Armstrong, K. & Asch, D. A. (1999). Estimating adverse selection costs from genetic testing for breast and ovarian cancer: The case of life insurance. *Journal of Risk and Insurance* **66,** 531–550.

Address for correspondence:
A. S. Macdonald, H. R. Waters and C. T. Wekwete
Department of Actuarial Mathematics and Statistics
Heriot-Watt University
Edinburgh EH14 4AS, UK
E-mail: A.S.Macdonald@ma.hw.ac.uk