LETTER TO JMG

Evaluation of widely used models for predicting *BRCA1* and BRCA2 mutations

F Marroni, P Aretini, E D'Andrea, M A Caligo, L Cortesi, A Viel, E Ricevuto, M Montagna, G Cipollini, S Ferrari, M Santarosa, R Bisegna, J E Bailey-Wilson, G Bevilacqua, G Parmigiani, S Presciuttini

J Med Genet 2004;**41**:278–285. doi: 10.1136/jmg.2003.013623

Deleterious mutations of the *BRCA1* and *BRCA2* genes are a major risk factor for the development of breast and ovarian cancers.[1–4] Mutation tests for these two genes commonly are now offered in specialised clinics.[5 6] As a result, a large number of women with personal or family histories of breast or ovarian cancer seek genetic counselling. Accurate evaluation of the probability that a woman carries a germline pathogenic mutation at *BRCA1* or *BRCA2* therefore is essential to help counsellors and those being counselled to decide whether testing is appropriate. In this context, the questions of practical interest are: Given the pedigree, what is the chance of a mutation being present? and What is the chance of the DNA laboratory finding a mutation?

After testing became available, several models were developed to assess the pre-test probability of identifying carriers of mutations. Broadly speaking, two different approaches have been used to develop predictive models: the ''empirical approach'' and the ''Mendelian approach''.[7] In empirical models, families are stratified according to variables that describe their family history; regression or other approaches are used to predict the results of Mendelian testing. In some cases, this approach simply consists of observing the proportion of mutations found in different strata. Mendelian models, in contrast, address the probability that a proband is a mutation carrier on the basis of explicit assumptions about the genetic parameters (allele frequencies and cancer penetrances in carriers and non-carriers) and the Mendelian rules of gene transmission. A consequence of the two different strategies is that the Mendelian models evaluate the probability that a proband is a gene carrier, whereas the empirical models evaluate the probability of identifying a mutation.

The main purpose of this study was to compare the performances of published models in predicting mutation test results in a large dataset. We collected pedigrees of probands investigated for *BRCA1* and *BRCA2* mutations in five clinical centres included in the Italian Consortium for Hereditary Breast and Ovarian Cancer. The combined sample included 568 families. Among those, 80 pathogenic mutations were identified in the *BRCA1* gene and 53 in the *BRCA2* gene. Eight models were investigated: the University of Pennsylvania (Penn) model, the Myriad-1 model, the Myriad Tables, the Spanish model, the Finnish model, the Yale model, the Brcapro model,[8–17] and a novel model that we refer to as the Italian Consortium (IC) model, intended to be used as a research tool. The latter is based on the parameter values of Brcapro (with minor modifications) and is implemented in the Mlink program of the Fastlink package.[18]

Mutations of the two genes are associated with differences in familial presentations. *BRCA1* is mutated preferentially in families with breast and ovarian cancer and more rarely is mutated in families with male breast cancer.[19–21] *BRCA2* was mapped primarily through families with male breast cancer.[22]

Key points

- Performances of eight models for predicting mutations were evaluated in 568 families screened for *BRCA1* and *BRCA2* mutations and stratified by risk level and by clustering of cancer type
- Each model showed its own performance deficits, often underestimating the likelihood of a mutation in some types of families, while overestimating it for others
- All models underestimated mutation probability in the low risk (<10%) group and most underestimated it for the moderate risk group (10–40%). In contrast, all models except the Myriad Tables overestimated mutation probabilities in the highest risk group
- Overall, two of the Mendelian models (Brcapro and a novel model developed for this study) performed better than the others
- Models that evaluated probabilities separately for each gene (Mendelian models only) attributed an excess of families to *BRCA1* compared with *BRCA2*; this effect was more pronounced for families with hereditary breast cancer
- This paper shows prospects for substantial improvement of performance, which could be achieved by adjusting the values of the relevant genetic parameters (allele frequencies and cancer penetrances in carriers and non-carriers)

Risk of breast cancer is higher for carriers of *BRCA1* mutations at younger ages (<45 years), although this may not be the case at older ages.[23] This shows that sufficient information may exist to assign specific mutation probabilities to each of the two genes.[24] In contrast, the models developed so far calculate joint probability of mutation, with the notable exception of the Brcapro and IC models. Brcapro's authors suggest, however, that its ability to discriminate between genes is limited.[25] In the last section of this paper, we address this problem by contrasting the probabilities calculated with the Brcapro and IC models with actual results of mutation tests, separately by gene and by family profile.

Previous validation studies considered one or two methods only or compared several methods without contrasting predictions with genetic test results.[25–27] A recent analysis compared performances of several models, although Mendelian models were not considered.[28] Our study is the first comprehensive attempt to evaluate model performances in a large series of families stratified according to family history and to consider the two genes separately.

PATIENTS AND METHODS

Data collection

Five cancer genetic clinics provided complete series of families screened for mutations in *BRCA1* and *BRCA2*. Because the clinics used different screening strategies, 458 families were screened for both genes, 104 for *BRCA1* only, and eight for *BRCA2* only. In mutation analysis, three centres used direct automatic sequencing and a combination of protein truncation test (PTT) plus single strand conformational polymorphism (SSCP), one used PTT-SSCP and fluorescence-assisted mutational analysis (FAMA), and one used PTT-SSCP only. Pedigree data included information about breast and ovarian cancer of the first degree and second degree relatives of probands. Information on family history was reported to genetic counsellors by family members. Errors in reporting are possible, particularly for second degree relatives,[29] but these errors also are likely to occur in the practical use of predictive models.[25] Eligibility criteria varied across centres, but families with multiple cases of breast and ovarian cancer or cases of early onset cancer were selected preferentially. The resulting sample consisted of 570 families; two families were of Ashkenazi ancestry (one harboured a *BRCA2* mutation) and were excluded from analysis. Among the 568 families that were included in this study, 151 had breast cancer and ovarian cancer in a single individual or in different relatives (HBOC), 357 had patients with breast cancer only (HBC), 31 had patients with ovarian cancer only (HOC), and 29 had at least one case of male breast cancer (MBC). Most of the probands (97%) were affected by breast or ovarian cancer, or both.

Empirical models

The Penn model was the first predictive tool developed after the cloning of the *BRCA1* gene.[8] It is based on logistic regression results of *BRCA1* testing on five variables that represent different family histories; tables that reported probabilities of mutation detection for 28 family groups were published (different tables were produced for Ashkenazi and non-Ashkenazi families). This model is applicable to the *BRCA1* gene only, and it does not deal with families in which ovarian cancer only is present (HOC families). The Myriad-1 model is also a logistic regression model, in which 10 variables pertaining to age, ethnicity, and family history of cancer were included.[9] This model also was built on *BRCA1* data only. Two other logistic regression models were published recently and predict probabilities of mutation detection in either *BRCA1* or *BRCA2*; we refer to them in this paper as the Spanish model and the Finnish model.[11 12] Neither model can be applied to HOC families. Finally, the Myriad mutation prevalence tables display the proportion of probands, stratified in 42 possible groups, with identified mutations in *BRCA1* or *BRCA2* in the analyses performed at Myriad; we used the August 2002 update, which included more than 10 000 tests (http://www.myriadtests.com).[10 30]

Mendelian models

The Yale model was developed before the cloning of the *BRCA1* gene; it originated the Claus model for predicting risk of breast cancer.[14 15] On the basis of segregation analysis, the maximum likelihood model assumed a dominant gene with population frequency of 0.0033 and mean ages of onset of breast cancer in gene carriers and non-carriers of 55.4 (SD 15.4) years and 69.0 (15.4) years, respectively.[13] Brcapro is another Mendelian model that is distributed as a part of the Mendelian counselling package CaGene[17 25]; it incorporates mutated allele frequencies and cancer specific penetrances derived from published results and uses Bayesian updating methods to compute carrier probabilities in pedigrees. Population frequencies of mutated *BRCA1* and *BRCA2* alleles are 0.0006 and 0.00022, respectively. Penetrance files are updated regularly; in our study, we used the version available in August 2002. The Brcapro software was also used to evaluate the Yale model by replacing default penetrances with the above values. The last model investigated, the IC model, was developed specifically for this study. In this, five age groups were defined for each of five mutually exclusive phenotypes of women, and two liability classes were defined for men (with and without breast cancer, respectively), which led to a total of 27 liability classes. Incidence ratios between *BRCA1* and *BRCA2* carriers and non-carriers in each class were the mean values of the corresponding ratios in the Brcapro parameter file and were set prior to data analysis. The main difference between the Brcapro and IC models is with respect to calculation of penetrances for patients with multiple tumours (the Brcapro model multiplies probabilities of each cancer, whereas the IC model assigns specific liability classes to patients with bilateral breast cancer or breast cancer plus ovarian cancer).

Sensitivity of molecular techniques

Importantly, empirical models evaluate the probability of finding a mutation in a proband, whereas Mendelian models evaluate the probability that the proband is a gene carrier. If the sensitivity of the molecular techniques was 100%, the two values would be directly comparable across different models. As a proportion of true gene carriers yield negative tests, however, the results of Mendelian models must be converted, as described below, before any comparison can be carried out. A direct estimate of sensitivity can be obtained by examining the proportion of families negative in a test that were linked to either locus: with this approach, Ford *et al*. found a value of 64%; on the basis of their results, a more recent work assumed a sensitivity of 70%.[2 31] Molecular techniques used to detect mutations varied across contributing centres and over time within centres; however, the most frequently used technique was PTT-SSCP, for which a blinded test showed sensitivity of 72–76% for abnormal migration detection and 60–65% for sequence analysis confirmation.[32] We therefore assumed a sensitivity value of 70% and converted the probability values calculated by Mendelian models by this factor. In addition, we explored the effect of changing the above assumption by recalculating mutation detection probabilities by using sensitivities of 60% and 80%. We refer to this probability as the ''mutation detection probability.''

Data analysis

Mutation detection probabilities were computed in each family, and three analyses were performed for each model: comparison of observed and expected number of mutations, computation of the likelihood of the observed test results given the calculated probabilities, and receiver operating characteristic curve analysis.

Expected number of mutations was calculated by summing mutation detection probabilities over all families or over given subsets of families in a stratified analysis; these values were compared with the observed number of mutations by the χ^2 test to evaluate calibration.[33] In addition, we computed the Cox and Snell U_0 and U_1 test statistics,[34] and we transformed them into the standardised z distribution to obtain appropriate confidence limits. The first index examines whether the predicted probabilities are systematically too high or too low (and is analogous to the χ^2 test above), and the second index examines whether the distributions of individual assigned probabilities are too variable within families with the same risk.

Table 1 Distribution of families by proband age and cancer type

Pedigree characteristics				Family history of cancer						Mutation found	
Proband		Number	Number			Bilateral	Breast and	Male breast	Mean affected		
Cancer	Age (years)	of families	of relatives	Breast	Ovarian	breast	ovarian	cancer	members*	*BRCA1*	*BRCA2*
Breast	<40	132	1351	129	17	17	10	5	2.35	26	12
	40–55	143	1998	241	19	25	5	7	3.08	6	12
	>55	64	1010	127	9	13	3	2	3.41	4	5
Bilateral breast		71	890	90	7	20	2	0	2.68	6	10
Ovary	<50	40	469	28	31	1	3	0	2.58	14	0
	⩾50	39	582	45	24	3	2	1	2.92	11	5
Breast and ovarian		49	609	38	10	3	1	1	2.08	12	4
Male breast		15	241	27	3	1	0	0	3.07	0	5
Unaffected		15	134	24	2	3	1	0	2.00	1	0
Total		568	7284	749	122	86	27	16	2.68	80	53

*Average number of members per family (probands included) affected by one or more breast or ovarian cancers.

Log likelihood

The logarithm of the likelihood of a set of mutation testing results was defined as $\ln(L) = \Sigma_i\, a \ln(p_i)+b \ln(1-p_i)$, where p_i is the mutation detection probability for family i, a is 1 if a mutation has been detected in the family and 0 otherwise, and b is 1 if no mutation has been detected and 0 otherwise. In computing probabilities separately for *BRCA1* and *BRCA2* (Brcapro and IC only), the likelihood function was modified accordingly, that is $\ln(L) = \Sigma a*\ln(p_{i1})+b*\ln(p_{i2})+c*\ln(1-p_{i1}-p_{i2})$, where p_{i1} and p_{i2} are mutation detection probabilities for *BRCA1* and *BRCA2*, respectively, and a, b, and c are binary variables (a is 1 only when a *BRCA1* mutation is detected, b is 1 only when a *BRCA2* mutation has been detected, and c is 1 only when no mutations have been detected). This assumes that the probability of testing positive for both genes is negligible. Log likelihood differences between pairs of models were tested by bootstrapping and by the paired sign test. In bootstraps, 10 000 samples were generated for each pairwise comparison, and the resulting series of values, each being a difference in total log likelihood, was ordered to obtain appropriate confidence intervals. The sign test was used to check that the median of the differences between individual likelihoods computed by any two models was different from zero.

Receiver operating characteristic curves

Receiver operating characteristic curves are used often in diagnostic test evaluations to determine the cut off value that provides the best discrimination between normal and abnormal patients. Receiver operating characteristic curve analysis was previously applied in validation studies of the Brcapro model;[26 35] here, we applied this analysis to compare the performance of the eight models. Receiver operating characteristic curve analysis is based on sensitivity and specificity of each particular predictive model; therefore, the definition of sensitivity is different from that used for molecular techniques (above). In this case, sensitivity represents the fraction of participants with mutations with detection probability higher than a given value, and specificity is the fraction of participants without mutations with probability lower than that value. Receiver operating characteristic curves are constructed by plotting sensitivity against (1 – specificity) for all possible values of the mutation detection probability; the area under the receiver operating characteristic curve is the fraction of all probands with identified mutations that have detection probabilities higher than probands with no mutation. An important threshold value for sensitivity is 10%; this is the probability threshold above which a person being counselled often is considered eligible for genetic testing.[26 36]

RESULTS

Sample characteristics

Our sample included 568 families of Caucasian ancestry. The total number of relatives was 7284, 1000 (13.7%) of whom were affected by breast cancer or ovarian cancer, or both. Cancers in probands were distributed as follows: 60% unilateral breast, 14% ovarian only, 13% bilateral breast, 9% breast and ovarian, and 3% male breast; 3% of the probands were unaffected. The total number of mutations identified was 133: 80 in *BRCA1* and 53 in *BRCA2*. Table 1 shows summary statistics of family histories of breast and ovarian cancer, as well as the number of identified mutations stratified by proband's cancer and age. Results indicate that the probability of finding mutations in *BRCA1* rather than in *BRCA2* (last two columns) varied among different groups of families. The ratio of *BRCA1* to *BRCA2* was larger than 1 in probands aged <40 years with breast cancer (row 1) but lower than 1 in probands aged >40 years (rows 2 and 3) (26:12 *v* 10:17; odds ratio 3.7 (95% confidence interval 1.3–10.4)). Similarly, *BRCA2* mutations were only found in probands with ovarian cancer when they were aged >50 years. Presence of familial correlations for the type of cancer was also apparent: for example, prevalence of ovarian cancer was higher among relatives of probands with ovarian cancer than among relatives of all other probands. In addition, the proportion of *BRCA1* and *BRCA2* mutations varied by family profile.

Comparative performance of the eight models

The subset of the total sample that could be analysed by all models consisted of 428 families (only families screened for both genes were taken into account, and 30 HOC families were excluded). The total number of identified mutations was 54 in *BRCA1* and 51 in *BRCA2*. Penn and Myriad-1 models were developed before the discovery of the *BRCA2* gene and considered mutation data in the *BRCA1* gene only; therefore, only mutations identified in this gene ($N = 54$) were counted as positive observations. Thus, results from the first two models could not be compared directly with results from the others. For the other six models, a positive observation was defined as the occurrence of a mutation in either gene ($N = 105$). Table 2 shows an overall evaluation of the predictions by all eight models. The first section (columns 2–4) shows the observed and expected statistics in the total sample; the second section shows the Cox and Snell U_0 and U_1 z transforms (columns 5 and 6), the third section (column 7) shows total log likelihoods, and the last section (columns 8–10) shows three statistics from the receiver operating characteristic curve analysis, namely the area under the curve (AUC) and the values of sensitivity and specificity in the

Table 2 Predictions of the eight models evaluated by several statistics

Model	Number of mutations		χ^2	z transform		Total log likelihood	Sensitivity (%)	Specificity (%)	AUC
	Observed	Expected		U_0	U_1				
Penn	54	64.7	1.7	−1.73	−1.49	−144.5	74	69	0.787
Myriad−1	54	60.0	0.5	−0.94	−0.54	−143.2	81	61	0.778
Myriad Tables	105	82.0	8.0	3.07	−3.53	−218.5	78	50	0.717
Spanish	105	97.9	0.7	0.95	−4.75	−240.0	80	35	0.651
Finnish	105	79.2	10.3	3.98	−9.66	−247.7	70	62	0.72
Yale	105	98.2	0.6	0.96	−17.06	−320.1	67	50	0.61
Brcapro	105	100.0	0.3	0.70	−7.42	−226.7	80	56	0.757
IC	105	109.0	0.2	−0.54	−4.57	−213.6	84	51	0.768

particular case when the threshold value of mutation detection probability was set to 10%.

Overall performances of Penn and Myriad-1 models were similar; both slightly overestimated the probability of detecting mutations, their total log likelihoods were close, and their AUCs were almost identical. When the number of expected mutations was considered, the Myriad Tables and Finnish models performed worst, underestimating the overall detection probability (predicting 78% and 75% of the number of mutations actually found, respectively). The remaining four models showed a better agreement between observed and expected values, but they differed in their total likelihoods, probably indicating error compensation between different family strata. This hypothesis was supported by the value of the second Cox and Snell index (column 6), which showed highly significant values for all models but the Penn and Myriad-1 models.

When the total log likelihood was considered, the IC model attained the maximum value, followed by the Myriad Tables and the Brcapro model; the others (Spanish, Finnish, and Yale) were distant. Bootstrap tests showed that the difference between the IC model and Myriad Tables was not significant, whereas all other comparisons were below the significance level of 0.05. On the other hand, the sign test was significant for the IC and Myriad model comparison, as well as all other cases. The receiver operator characteristic curve analysis (fig 1) also showed some differences among models. The Brcapro and IC models ranked first (AUC 76% and77%), although the difference between those and the Myriad Tables and Finnish model was small (AUC 72%); the Spanish and Yale models performed worst (61% and 65%).

Performances of the Mendelian models assumed a value of 70% for the sensitivity of molecular techniques. To explore the consequences of modifying this value, we recalculated observed and expected χ^2 and total log likelihoods for the Brcapro and IC models with sensitivities of 60% and 80%. Resulting χ^2 values were higher in both cases for both models and log likelihoods also indicated a poorer fit (they were lower by about 5 log units with sensitivity 80%); an exception was the IC model when a sensitivity of 60% was used, in which case a small log likelihood increase was observed (0.58 log units).

Table 3 shows the log likelihoods stratified by proband's type of cancer and age (as in table 1), for the six models that evaluated mutation probabilities in either gene. Prediction ability varied considerably among models across the various categories of families. The largest difference concerned probands aged <40 years with breast cancer, where the likelihood of the Myriad Tables was 8–10 units lower than that of the two Mendelian models; this was apparently caused by a large underestimation of mutation detection probability by this model compared with the other two (17.2 *v* 22.5 and 23.3 predicted when 29 mutations were observed). On the other hand, the Myriad Tables performed better than the Mendelian models in families of probands aged >55 years with breast cancer and in those with bilateral breast cancer. In this category, the Mendelian models predicted a twofold excess of mutations (24.5 and 22.9 expected mutations in the Brcapro and the IC models, respectively, compared with 8.8 in the Myriad Tables with 14 observed mutations). Another important outcome concerned the families of probands aged >50 years with ovarian cancer, for which the two Mendelian models gave likelihoods that differed by about 6 log units.

To further investigate differences in performances, families were stratified separately by risk in three groups (<10%, 10–40%, and >40%) for each model (table 4). The proportion of families with probabilities <10% varied among models from 31% in the Spanish model to 54% in the Finnish model. All models underestimated detection probability in the <10% risk group; the largest discrepancy was observed for the Yale model (35 observed mutations *v* 5.5 expected) and the smallest for the Myriad Tables (23 *v* 13.0). In the intermediate risk group (10–40%), a general excess of identified mutations was also observed, although the three Mendelian models produced predictions close to actual observations. The proportion of families in the highest risk group was the most variable among models, ranging from 10% in the Myriad Tables to 30% in the IC model. With the exception of the Myriad Tables, which almost exactly predicted the correct number of mutations, all other models overestimated the detection probability.

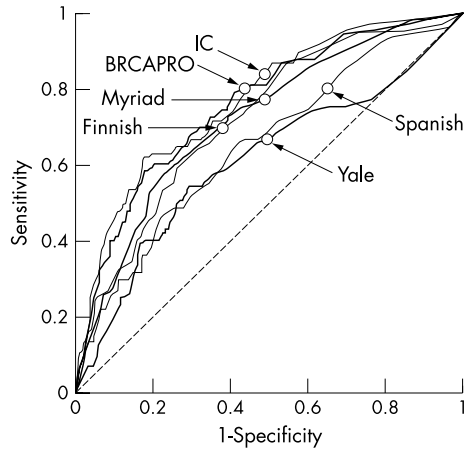

Figure 1 Receiving operator characteristic curves of six models that evaluated mutation detection probability in either *BRCA1* or *BRCA2*. ○ value of sensitivity and specificity with threshold value of mutation detection probability set to 10%.

Table 3 Comparison of log likelihoods of six predictive model by probands' characteristics

Pedigree characteristics				Model					
Proband									
Cancer	Age (years)	Number of families	Number of mutations	Myriad	Spanish	Finnish	Yale	Brcapro	IC
Breast	<40	94	29	−60.1	−60.0	−60.9	−54.6	−51.8	−50.7
	40–55	120	17	−37.9	−40.4	−34.7	−40.7	−34.4	−37.6
	>55	49	8	−23.1	−22.6	−25.6	−29.1	−28.4	−27.6
Bilateral breast		60	14	−32.0	−32.7	−34.2	−44.5	−37.7	−37.0
Ovary	<50	21	9	−15.2	−17.0	−24.7	−37.6	−13.4	−12.6
	⩾50	23	11	−21.5	−28.3	−31.5	−53.8	−28.4	−22.2
Breast and ovarian		38	12	−22.4	−26.8	−25.4	−41.2	−22.7	−19.0
Male breast		9	4	−3.7	−7.7	−8.8	−10.5	−6.1	−4.3
Unaffected		14	1	−2.6	−4.4	−1.8	−8.1	−3.8	−2.7
Total		428	105	−218.5	−240.0	−247.7	−320.1	−226.7	−213.6
χ^2				25.29	22.73	28.66	260.78	26.45	16.93

Differentiating probabilities between *BRCA1* and *BRCA2*

The Brcapro and IC models compute mutation probabilities separately for the two genes and cover all possible configurations of breast and ovarian cancers; this allowed us to examine their performances with respect to both genes in all the 568 families stratified by the four typical profiles (HBC, HOC, HBOC, and MBC). Table 5 shows the results of this analysis, in terms of χ^2 and log likelihood statistics considering the two genes jointly and then separately by gene.

Total log likelihoods calculated over the 568 families were −381.7 for the IC model and −396.2 for Brcapro: with a difference of 14.5 log units. Most of this difference (10.7 log units) was because of the HBOC profile, in which Brcapro predicted 41.1 mutations and the IC model 47.7 (57 were observed). This discrepancy was also responsible for most of the difference between total χ^2 values (4.2 *v* 14.8). When we examined the predictions separately by gene, we still found a difference of total likelihoods between the two models (7.1 log unit difference for *BRCA1* and 6.3 for *BRCA2*, both in favour of the IC model). The most striking feature of this analysis, however, was the large excess of *BRCA1* mutations predicted by both models for the HBC group, with a corresponding large deficit of predicted *BRCA2* mutations (about 48 mutations predicted by both models *v* 27 observed in *BRCA1* and about 12.5 predicted *v* 33 observed in *BRCA2*). For the other profiles, predictions were more accurate, although both models underestimated the number of mutations detected in *BRCA2* for the HBOC profile (about six predicted *v* 12 observed).

DISCUSSION

Determination of the probability that a proband carries a *BRCA1* or *BRCA2* mutation by using family history is important and challenging. It requires weighing the possibility that a given cluster of cases among relatives is because of chance against the possibility of a predisposing gene. A simple approach—the ''empirical approach''—involves collecting families tested for the genes, searching for variables that best discriminate between positive and negative families, and then building a model based on these. An alternative approach is to estimate the allele frequencies in the population and the cancer penetrances in both gene carriers and non-carriers and then applying a Mendelian model to each family (the Mendelian approach). A disadvantage of the empirical approach is that it needs large samples to provide reliable predictions; in addition, empirical models often refer to ''number of cases per family'' without clearly defining what a family is, which implies that this variable could mean

Table 4 Expected versus observed number of mutations by risk.* Values are numbers (percentages)

Model	Number of families	Observed	Expected	Observed/expected	χ^2
P<0.01					
Myriad Tables	186 (43)	23	13.0	1.8	8.2
Spanish	134 (31)	21	7.8	2.7	23.6
Finnish	232 (54)	32	9.8	3.3	52.1
Yale	198 (46)	35	5.5	6.4	162.6
Brcapro	203 (47)	21	5.5	3.8	44.3
IC	182 (43)	17	5.9	2.9	21.5
0.1<P<0.4					
Myriad Tables	200 (47)	60	46.4	1.3	5.2
Spanish	213 (50)	51	45.8	1.1	0.7
Finnish	133 (31)	42	28.2	1.5	8.6
Yale	114 (27)	26	23.9	1.1	0.2
Brcapro	101 (24)	22	21.9	1.0	0.0
IC	116 (27)	22	26.1	0.8	0.8
P>0.4					
Myriad Tables	42 (10)	22	22.6	1.0	0.0
Spanish	81 (19)	33	44.2	0.7	6.3
Finnish	63 (15)	31	41.1	0.8	7.2
Yale	116 (27)	44	68.7	0.6	21.8
Brcapro	124 (29)	62	72.6	0.9	3.7
IC	130 (30)	66	77.0	0.9	3.9

*Probabilities stratified in three groups separately by model, so that families in each group differ across models.

Table 5 Comparison between the Brcapro and IC models in the entire dataset and by gene

Gene	Number of families	Number of mutations	IC model Mutations expected	IC model χ^2	IC model Log likelihood	Brcapro model Mutations expected	Brcapro model χ^2	Brcapro model Log likelihood
BRCA1* and *BRCA2								
HBC	357	60	62.7	0.14	−204.2	59.3	0.01	−204.8
HBOC	151	57	47.7	2.63	−138.8	41.2	8.31	−149.5
HOC	31	8	6.7	0.31	−16.1	4.0	4.70	−15.7
MBC	29	8	10.7	1.12	−22.7	11.6	1.83	−26.2
Total	568	133	127.9	4.20	−381.7	116.0	14.84	−396.2
BRCA1								
HBC	357	27	49.0	11.4	−97.2	47.4	10.1	−94.3
HBOC	151	45	41.2	0.5	−102.7	35.8	3.1	−110.4
HOC	31	7	6.3	0.1	−13.2	3.7	3.3	−13.4
MBC	21	1	3.4	2.0	−6.6	3.5	2.2	−8.6
Total	560	80	99.9	14.1	−219.6	90.5	18.8	−226.7
BRCA2								
HBC	276	33	13.7	28.7	−117.4	11.8	39.5	−121.4
HBOC	131	12	6.6	4.7	−43.1	5.4	8.3	−45.6
HOC	28	1	0.4	1.1	−3.9	0.3	2.0	−2.9
MBC	29	7	5.8	0.3	−18.1	5.9	0.2	−19.0
Total	464	53	26.4	34.8	−182.5	23.5	50.0	−188.9

very different things in different families. A disadvantage of the Mendelian approach is that accurate estimates of penetrances and allele frequencies may be difficult to obtain; in addition, all existing empirical and Mendelian models currently assume that all mutant alleles at each gene have the same penetrance.

We compared relative performances of five empirical and three Mendelian models. We evaluated calibration of models with χ^2 analysis, refinement of models with receiver operator characteristic curve analysis, and overall goodness of fit of models with log likelihood. Three of these eight models (Penn, Myriad-1, and Yale) were developed before the discovery of *BRCA2* (Yale was proposed before the cloning of *BRCA1*) and were investigated here for completeness. Penn and Myriad-1 could include observations on *BRCA1* only, and HOC families necessarily were excluded from analysis; within these limits, they performed relatively well, considering both the observed and expected χ^2 statistics and the receiver operator characteristic curve analysis. The Yale model performed worse than all other models, although it must be acknowledged that the original analysis had the purpose of estimating the genetic parameters of a gene predisposing to breast cancer rather than predicting mutation risks. Among the other five models, Mendelian models provided higher resolution, as indicated by analysis of the receiver operator characteristic curve results. This is probably the consequence of calculating individualised probabilities—a major advantage of this approach compared with methods that tabulate probability values for a discrete number of familial groups. In addition, Mendelian models were more accurate for estimating the overall number of mutations. Considering log likelihood analysis, the Myriad Tables provided a value between those of the two Mendelian models.

A novel feature of our study is the analysis of predicted probabilities in the families stratified by probands' characteristics. Different approaches make different types of errors, so the possible similarity of results at the level of the total sample may be the consequence of error compensation in different family strata. For example, the Myriad Tables predicted little more than half of the observed mutations for families of probands aged <40 years with breast cancer compared with a better prediction by the Mendelian models, but this error was compensated for by the Myriad Tables' better prediction for families of probands aged >55 years with breast cancer and those with bilateral breast cancer. Further analyses of this type may help to identify the categories of families for which adjustments of the parameter values that influence probability calculation are most needed.

Another interesting result concerned the number of observed and expected mutations in the families stratified by risk according to each model. All models underestimated the probability of detecting mutations in the families in the lowest risk class (⩽10%). The model performing best in this analysis was the Myriad Tables, but the predicted number of mutations was only about half the observed number. As the proportion of families included in this group was large for all models (about 45% on average), the number of ''missed'' mutations was large on an absolute scale. This result may have important consequences. On one hand, the number of actual mutations in low risk families may be higher than previously thought; on the other hand, the risk conferred by these mutations may be lower than anticipated.[1 2 37] This lack of fit may be specific to the Italian population, although data available so far about *BRCA* mutations in Italy does not suggest this.[38] An alternative explanation would be penetrance heterogeneity among mutations; in this case, an excess of mutations that confer lower cancer risk would be identified in participants with relatively mild family histories.

Analysis of Mendelian models for accuracy in discriminating between the two genes showed an area of study in which further investigation could increase performance substantially. Our data confirmed the existence of different patterns of clinical expression between the two genes, as shown by the different *BRCA1*:*BRCA2* mutation detection rate in different profiles. Both models predicted an overall excess of *BRCA1* mutations, and this excess was particularly large for HBC families (which were preferentially mutated in *BRCA2*); this suggests that current parameterisation of the models still is inadequate to attribute correct probabilities to each gene and that margins for improvement exist.

Conclusion

Whereas present Mendelian models perform generally better than empirical models (and in addition provide individualised probabilities that cover all possible familial configurations) adjustment of genetic parameters in two main areas could substantially improve their performance. These areas concern the families at low risk, who are likely to constitute a large fraction of future people being counselled and for whom the models underestimate the mutation detection

probability, and the ability to discriminate between the two genes. Experience gained during our analysis suggests that a promising strategy is to re-estimate parameters from the data by maximum likelihood. As our data represent the genetic condition existing in Italy, this work may lead to a version of the Brcapro software customised for this country—an example that later could be extended to other populations.

Authors' affiliations

F Marroni, P Aretini, M A Caligo, G Cipollini, G Bevilacqua, Department of Oncology, Transplants and New Technologies in Medicine, Section of Pathology, University of Pisa, Pisa, Italy
F Marroni, S Presciuttini, Center of Statistical Genetics, University of Pisa, Pisa, Italy
E D'Andrea, M Montagna, Department of Oncology and Surgical Sciences, Section of Oncology (E.D'A.); IST, Section of Viral and Molecular Oncology (M.M.); University of Padua, Padua, Italy
L Cortesi, S Ferrari, Department of Oncology and Hematology (L.C); Department of Biomedical Science, Section of Biological Chemistry (S.F.); University of Modena and Reggio Emilia, Modena, Italy
A Viel, M Santarosa, Experimental Oncology 1, Oncology Referral Center, IRCCS, Aviano (PN), Italy
E Ricevuto, R Bisegna, Department of Experimental Medicine, University of L'Aquila, L'Aquila, Italy
J E Bailey-Wilson, S Presciuttini, Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, USA
G Parmigiani, Departments of Oncology and Biostatistics, Johns Hopkins University, Baltimore, MD, USA

Correspondence to: Dr Presciuttini, Genetica Statistica, c/o Centro Retrovirus, SS Abetone e Brennero 2, 56127 Pisa, Italy; sprex@biomed.unipi.it

Conflicts of interest: None declared.

Funding: The Italian Consortium for Hereditary Breast and Ovarian Cancer is funded by the Italian Association for Cancer Research (AIRC). This work was in part supported by the National Research Council (CNR) and Italian Ministry of University and Research (MIUR).

Received 22 August 2003
Accepted for publication 16 October 2003

REFERENCES

1 **Easton DF**, Ford D, Bishop DT. Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *Am J Hum Genet* 1995;**56**:265–71.
2 **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, and the Breast Cancer Linkage Consortium. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* 1998;**62**:676–89.
3 **Anglian Breast Cancer Study Group**. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. *Br J Cancer* 2000;**83**:1301–8.
4 **Bishop DT**. BRCA1 and BRCA2 and breast cancer incidence: a review. *Ann Oncol* 1999;**10**(suppl 6):113–9.
5 **Steel M**, Smyth E, Vasen H, Eccles D, Evans G, Moller P, Hodgson S, Stoppa-Lyonnet D, Chang-Claude J, Caligo M, Morrison P, Haites N. Ethical, social and economic issues in familial breast cancer: a compilation of views from the EC Biomed II Demonstration Project. *Dis Markers* 1999;**15**:125–31.
6 **Edwards RT**. Steering a course around the genetic iceberg. *J Public Health Med* 2001;**23**:3–4.
7 **Parmigiani G**. *Modeling in medical decision making. A bayesian approach*. New York: Wiley, 2002.
8 **Couch FJ**, DeShano ML, Blackwood MA, Calzone K, Stopfer J, Campeau L, Ganguly A, Rebbeck T, Weber BL. BRCA1 mutations in women attending clinics that evaluate the risk of breast cancer. *N Engl J Med* 1997;**336**:1409–15.
9 **Shattuck-Eidens D**, Oliphant A, McClure M, McBride C, Gupte J, Rubano T, Pruss D, Tavtiian S, Teng DHF, Adey N, Staebell M, Gumpper K, Lundstrom R, Hulick M, Kelly M, Holmen J, Lingenfelter B, Manley S, Fujumura F, Luce M, Ward B, Cannon-Albright L, Steele L, Offit K, Gilewski T, Norton L, Brown K, Schulz C, Hampel H, Schluger A, Giulotto E, Zoli W, Ravaioli A, Nevanlinna H, Pyrhonen S, Rowley P, Scalia J, Michaelson R, Scottt R, Radice P, Pierotti M, Garber J, Isaacs C, Peshkin B, Lippman M, Dosik M, Caligo M, Greenstein R, Pilarski R, Weber B, Burgemeister R, Frank T, Skolnick M, Thomas A. BRCA1 sequence analysis in women at high risk for susceptibility mutations. Risk factor analysis and implications for genetic testing. *JAMA* 1997;**278**:1242–50.
10 **Frank TS**, Deffenbaugh AM, Reid JE, Hulick M, Ward BE, Lingenfelter B, Gumpper KL, Scholl T, Tavtigian SV, Pruss DR, Critchfield GC. Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: analysis of 10,000 individuals. *J Clin Oncol* 2002;**20**:1480–90.
11 **de la Hoya M**, Osorio A, Godino J, Sulleiro S, Tosar A, Perez-Segura P, Fernandez C, Rodriguez R, Diaz-Rubio E, Benitez J, Devilee P, Caldes T. Association between BRCA1 and BRCA2 mutations and cancer phenotype in Spanish breast/ovarian cancer families: implications for genetic testing. *Int J Cancer* 2002;**97**:466–71.
12 **Vahteristo P**, Eerola H, Tamminen A, Blomqvist C, Nevanlinna H. A probability model for predicting BRCA1 and BRCA2 mutations in breast and breast-ovarian cancer families. *Br J Cancer* 2001;**84**:704–8.
13 **Claus EB**, Risch N, Thompson WD. Genetic analysis of breast cancer in the cancer and steroid hormone study. *Am J Hum Genet* 1991;**48**:232–42.
14 **Claus EB**, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer* 1994;**73**:643–51.
15 **Claus EB**, Risch N, Thompson WD. The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. *Breast Cancer Res Treat* 1993;**28**:115–20.
16 **Berry DA**, Parmigiani G, Sanchez J, Schildkraut J, Winer E. Probability of carrying a mutation of breast-ovarian cancer gene BRCA1 based on family history. *J Natl Cancer Inst* 1997;**89**:227–38.
17 **Parmigiani G**, Berry D, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *Am J Hum Genet* 1998;**62**:145–58.
18 **Schaffer AA**. Faster linkage analysis computations for pedigrees with loops or unused alleles. *Hum Hered* 1996;**46**:226–35.
19 **Osorio A**, Barroso A, Martinez B, Cebrian A, San Roman JM, Lobo F, Robledo M, Benitez J. Molecular analysis of the BRCA1 and BRCA2 genes in 32 breast and/or ovarian cancer Spanish families. *Br J Cancer* 2000;**82**:1266–70.
20 **Martin AM**, Blackwood MA, Antin-Ozerkis D, Shih HA, Calzone K, Colligon TA, Seal S, Collins N, Stratton MR, Weber BL, Nathanson KL. Germline mutations in BRCA1 and BRCA2 in breast-ovarian families from a breast cancer risk evaluation clinic. *J Clin Oncol* 2001;**19**:2247–53.
21 **Shih HA**, Couch FJ, Nathanson KL, Blackwood MA, Rebbeck TR, Armstrong KA, Calzone K, Stopfer J, Seal S, Stratton MR, Weber BL. BRCA1 and BRCA2 mutation frequency in women evaluated in a breast cancer risk evaluation clinic. *J Clin Oncol* 2002;**20**:994–9.
22 **Wooster R**, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Sea S, Tran T, Averill D, Fields P, Marshall G, Narod S, Lenoir GM, Lynch H, Feunteun J, Devilee P, Cornelisse CJ, Menko FH, Daly PA, Ormiston W, McManus R, Pye C, Lewis CM, Cannon Albright L, Peto J, Ponder BAJ, Skolnick MH, Easton DF, Goldgar DE, Stratton MR. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12–13. *Science* 1994;**265**:2088–90.
23 **Malone KE**, Daling JR, Neal C, Suter NM, O'Brien C, Cushing-Haugen K, Jonasdottir TJ, Thompson JD, Ostrander EA. Frequency of BRCA1/BRCA2 mutations in a population-based sample of young breast carcinoma cases. *Cancer* 2000;**88**:1393–402.
24 **Aretini P**, D'Andrea E, Pasini B, Viel A, Costantini RM, Cortesi L, Ricevuto E, Agata S, Bisegna R, Boiocchi M, Caligo MA, Chieco-Bianchi L, Cipollini G, Crucianelli R, D'Amico C, Federico M, Ghimenti C, De Giacomi C, De Nicolo A, Della Puppa L, Ferrari S, Ficorella C, Iandolo D, Manoukian S, Marchetti P, Marroni F, Menin C, Montagna M, Ottini L, Pensotti V, Pierotti M, Radice P, Santarosa M, Silingardi V, Turchetti D, Bevilacqua G, Presciuttini S. Different expressivity of BRCA1 and BRCA2: analysis of 179 Italian pedigrees with identified mutation. *Breast Cancer Res Treat* 2003;**81**:71–9.
25 BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. *J Clin Oncol* 2002;**20**:2701–12.
26 **Euhus DM**, Smith KC, Robinson L, Stucky A, Olopade OI, Cummings S, Garber JE, Chittenden A, Mills GB, Rieger P, Esserman L, Crawford B, Hughes KS, Roche CA, Ganz PA, Seldon J, Fabian CJ, Klemp J, Tomlinson G. Pretest prediction of BRCA1 or BRCA2 mutation by risk counselors and the computer model BRCAPRO. *J Natl Cancer Inst* 2002;**94**:844–51.
27 **Shannon KM**, Lubratovich ML, Finkelstein DM, Smith BL, Powell SN, Seiden MV. Model-based predictions of BRCA1/2 mutation status in breast carcinoma patients treated at an academic medical center. *Cancer* 2002;**94**:305–13.
28 **De la Hoya M**, Diez O, Perez-Segura P, Godino J, Fernandez JM, Sanz J, Alonso C, Baiget M, Diaz-Rubio E, Caldes T. Pre-test prediction models of BRCA1 or BRCA2 mutation in breast/ovarian families attending familial cancer clinics. *J Med Genet* 2003;**40**:503–10.
29 **Love RR**, Evans AM, Josten DM. The accuracy of patient reports of a family history of cancer. *J Chronic Dis* 1985;**38**:289–93.
30 **Frank TS**, Manley SA, Olopade OI, Cummings S, Garber JE, Bernhardt B, Antman K, Russo D, Wood ME, Mullineau L, Isaacs C, Peshkin B, Buys S, Venne V, Rowley PT, Loader S, Offit K, Hampel H, Brener D, Winer EP, Clark S, Weber B, Strong LC, Reiger P, McClure M, Ward B, Shattuck-Eidens D, Oliphant A, Skolnick MH, Thomas A. Sequence analysis of BRCA1 and BRCA2: correlation of mutations with family history and ovarian cancer risk. *J Clin Oncol* 1998;**16**:2417–25.
31 **Basham VM**, Lipscombe JM, Ward JM, Gayther SA, Ponder BA, Easton DF, Pharoah PD. BRCA1 and BRCA2 mutations in a population-based study of male breast cancer. *Breast Cancer Res* 2002;**4**:R2.

32 **Eng C**, Brody LC, Wagner TM, Devilee P, Vijg J, Szabo C, Tavtigian SV, Nathanson KL, Ostrander E, Frank TS. Interpreting epidemiological research: blinded comparison of methods used to estimate the prevalence of inherited mutations in BRCA1. *J Med Genet* 2001;**38**:824–33.
33 **DeGroot MH FS**. Assessing probability assessors: calibration and refinement. *Statistical decision theory and related topics III* 1982;**1**:291–314.
34 **Cox DR**, Snell EJ. *Analysis of binary data*. London, New York, Tokyo, Melbourne, Madras: Chapman and Hall, 1989:41–3.
35 **Iversen E**, Parmigiani G, Berry D. Validating Bayesian prediction models: a case study in genetic susceptibility to breast cancer. *Case studies in Bayesian statistics* 1998;**IV**:321–38.
36 **American Society of Clinical Oncology**. *Policy statement: genetic testing for cancer susceptibility (approved February 20, 1996) Recommendations pertaining to clinical aspects of genetic testing for cancer susceptibility*. Alexandria: ASCO, 2002.
37 **Ford D**, Easton DF, Peto J. Estimates of the gene frequency of BRCA1 and its contribution to breast and ovarian cancer incidence. *Am J Hum Genet* 1995;**57**:1457–62.
38 **Ottini L**, D'Amico C, Noviello C, Lauro S, Lalle M, Fornarini G, Colantuoni OA, Pizzi C, Cortesi E, Carlini S, Guadagni F, Bianco AR, Frati L, Contegiacomo A, Mariani-Costantini R. BRCA1 and BRCA2 mutations in central and southern Italian patients. *Breast Cancer Res* 2000;**2**:307–10.

ECHO

Transforming growth factor-β_1 genotype and susceptibility to chronic obstructive pulmonary disease

L Wu, J Chau, R P Young, V Pokorny, G D Mills, R Hopkins, L McLean, P N Black

Please visit the *Journal of Medical Genetics* website [www.jmedgenet.com] for a link to the full text of this article.

Background: Only a few long term smokers develop symptomatic chronic obstructive pulmonary disease (COPD) and this may be due, at least in part, to genetic susceptibility to the disease. Transforming growth factor β_1 (TGF-β_1) has a number of actions that make it a candidate for a role in the pathogenesis of COPD. We have investigated a single nucleotide polymorphism at exon 1 nucleotide position 29 (T→C) of the TGF-β_1 gene that produces a substitution at codon 10 (Leu→Pro).
Methods: The frequency of this polymorphism was determined in 165 subjects with COPD, 140 healthy blood donors, and 76 smokers with normal lung function (resistant smokers) using the polymerase chain reaction and restriction enzyme fragment length polymorphism.
Results: The distribution of genotypes was Leu-Leu (41.8%), Leu-Pro (50.3%), and Pro-Pro (7.9%) for subjects with COPD, which was significantly different from the control subjects (blood donors: Leu-Leu (29.3%), Leu-Pro (52.1%) and Pro-Pro (18.6%), $p = 0.006$; resistant smokers: Leu-Leu (28.9%), Leu-Pro (51.3%) and Pro-Pro (19.7%), $p = 0.02$). The Pro^{10} allele was less common in subjects with COPD (33%) than in blood donors (45%; OR = 0.62, 95% CI 0.45 to 0.86, $p = 0.005$) and resistant smokers (45%; OR = 0.59, 95% CI 0.40 to 0.88, $p = 0.01$).
Conclusions: The proline allele at codon 10 of the TGF-β_1 gene occurs more commonly in control subjects than in individuals with COPD. This allele is associated with increased production of TGF-β_1 which raises the possibility that TGF-β_1 has a protective role in COPD.
▲ *Thorax* 2004;**59**:126–129.