Evaluation of widely used models for predicting \textit{BRCA1} and \textit{BRCA2} mutations


\textbf{LETTER TO JMG}

Deleterious mutations of the \textit{BRCA1} and \textit{BRCA2} genes are a major risk factor for the development of breast and ovarian cancers.\textsuperscript{1,4} Mutation tests for these two genes are now offered in specialised clinics.\textsuperscript{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 \textit{BRCA1} or \textit{BRCA2} 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”.\textsuperscript{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 \textit{BRCA1} and \textit{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 \textit{BRCA1} gene and 53 in the \textit{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,\textsuperscript{18,19} 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 parameter values of Brcapro (with minor modifications) and is implemented in the Mlink program of the Fastlink package.\textsuperscript{18,19}

Mutations of the two genes are associated with differences in familial presentations. \textit{BRCA1} is mutated preferentially in families with breast and ovarian cancer and more rarely in families with male breast cancer.\textsuperscript{10–21} \textit{BRCA2} was mapped primarily through families with male breast cancer.\textsuperscript{22}

\begin{itemize}
  \item Performances of eight models for predicting mutations were evaluated in 568 families screened for \textit{BRCA1} and \textit{BRCA2} mutations and stratified by risk level and by clustering of cancer type
  \item Each model showed its own performance deficits, often underestimating the likelihood of a mutation in some types of families, while overestimating it for others
  \item 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
  \item Overall, two of the Mendelian models (Brcapro and a novel model developed for this study) performed better than the others
  \item Models that evaluated probabilities separately for each gene (Mendelian models only) attributed an excess of families to \textit{BRCA1} compared with \textit{BRCA2}; this effect was more pronounced for families with hereditary breast cancer
  \item 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)
\end{itemize}

Risk of breast cancer is higher for carriers of \textit{BRCA1} mutations at younger ages (<45 years), although this may not be the case at older ages.\textsuperscript{23} This shows that sufficient information may exist to assign specific mutation probabilities to each of the two genes.\textsuperscript{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.\textsuperscript{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.\textsuperscript{15–27} A recent analysis compared performances of several models, although Mendelian models were not considered.\textsuperscript{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, but these errors also are likely to occur in the practical use of predictive models. 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. 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 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. 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. Neither model can be applied to families with the same risk.

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. 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. Brcapro is another Mendelian model that is distributed as a part of the Mendelian counselling package CaGene; 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 of the Brcapro parameter files 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%. 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. 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 $\chi^2$ test to evaluate calibration. In addition, we computed the Cox and Snell $U_0$ and $U_1$ test statistics, 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 $\chi^2$ 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

<table>
<thead>
<tr>
<th>Pedigree characteristics</th>
<th>Family history of cancer</th>
<th>Mutation found</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proband Cancer</td>
<td>Age (years)</td>
<td>Number of families</td>
</tr>
<tr>
<td>Breast</td>
<td>&lt;40</td>
<td>132</td>
</tr>
<tr>
<td>40–55</td>
<td>&gt;55</td>
<td>143</td>
</tr>
<tr>
<td>Bilateral breast</td>
<td></td>
<td>64</td>
</tr>
<tr>
<td>Ovary</td>
<td>&lt;50</td>
<td>71</td>
</tr>
<tr>
<td>&gt;50</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>Breast and ovarian</td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>Male breast</td>
<td></td>
<td>49</td>
</tr>
<tr>
<td>Unaffected</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>568</td>
</tr>
</tbody>
</table>

*Average number of members per family (proband 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 \( \ln(L) = \sum a \ln(p_i) + b \ln(1-p_i) \), 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 (Bracapo and IC only), the likelihood function was modified accordingly, that is \( \ln(L) = \sum a \ln(p_i) + b \ln(1-p_i) + c \ln(1-p_i-p_2) \), where \( p_1 \) and \( p_2 \) 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 Bracapo model. 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.

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 the BRCA2 gene only; \( N = 105 \). 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 Nagelkerke \( U_1 \) statistics (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
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% and 77%), 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 \( \chi^2 \) and total log likelihoods for the Brcapro and IC models with sensitivities of 60% and 80%. Resulting \( \chi^2 \) 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 vs 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 vs 5.5 expected) and the smallest for the Myriad Tables (23 vs 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of mutations</th>
<th>Observed</th>
<th>Expected</th>
<th>( \chi^2 )</th>
<th>Sensitivity</th>
<th>Specificity (%)</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penn</td>
<td>54</td>
<td>64.7</td>
<td>1.7</td>
<td>(-1.73)</td>
<td>74</td>
<td>69</td>
<td>0.78</td>
</tr>
</tbody>
</table>
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 $\chi^2$ 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 $\chi^2$ values (4.2 vs 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 vs 27 observed in BRCA1 and about 12.5 predicted vs 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 vs 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 3
Comparison of log likelihoods of six predictive model by probands’ characteristics

<table>
<thead>
<tr>
<th>Pedigree characteristics</th>
<th>Model</th>
<th>Myriad</th>
<th>Spanish</th>
<th>Finnish</th>
<th>Yale</th>
<th>Brcapro</th>
<th>IC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proband Cancer Age (years)</td>
<td>Number of families</td>
<td>Number of mutations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast &lt;40</td>
<td>94</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast 40-55</td>
<td>120</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast &gt;55</td>
<td>49</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilateral breast &lt;50</td>
<td>60</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilateral breast &gt;50</td>
<td>23</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ovary &lt;50</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ovary &gt;50</td>
<td>23</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast and ovarian</td>
<td>38</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male breast</td>
<td>9</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unaffected</td>
<td>14</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>428</td>
<td>105</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\chi^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

### Table 4
Expected versus observed number of mutations by risk.

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of families</th>
<th>Observed</th>
<th>Expected</th>
<th>Observed/expected $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P&lt;0.01</td>
<td>Myriad Tables</td>
<td>186 (43)</td>
<td>23</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>134 (31)</td>
<td>21</td>
<td>7.8</td>
</tr>
<tr>
<td></td>
<td>Finnish</td>
<td>232 (54)</td>
<td>32</td>
<td>9.8</td>
</tr>
<tr>
<td></td>
<td>Yale</td>
<td>198 (46)</td>
<td>35</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td>Brcapro</td>
<td>203 (47)</td>
<td>21</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td>IC</td>
<td>192 (43)</td>
<td>17</td>
<td>5.9</td>
</tr>
<tr>
<td>0.1&lt;P&lt;0.4</td>
<td>Myriad Tables</td>
<td>200 (47)</td>
<td>60</td>
<td>46.4</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>213 (50)</td>
<td>51</td>
<td>45.8</td>
</tr>
<tr>
<td></td>
<td>Finnish</td>
<td>133 (31)</td>
<td>42</td>
<td>28.2</td>
</tr>
<tr>
<td></td>
<td>Yale</td>
<td>114 (27)</td>
<td>26</td>
<td>23.9</td>
</tr>
<tr>
<td></td>
<td>Brcapro</td>
<td>101 (24)</td>
<td>22</td>
<td>21.9</td>
</tr>
<tr>
<td></td>
<td>IC</td>
<td>116 (27)</td>
<td>22</td>
<td>26.1</td>
</tr>
<tr>
<td>P&lt;0.4</td>
<td>Myriad Tables</td>
<td>42 (10)</td>
<td>22</td>
<td>22.6</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>81 (19)</td>
<td>33</td>
<td>44.2</td>
</tr>
<tr>
<td></td>
<td>Finnish</td>
<td>63 (15)</td>
<td>31</td>
<td>41.1</td>
</tr>
<tr>
<td></td>
<td>Yale</td>
<td>116 (27)</td>
<td>44</td>
<td>68.7</td>
</tr>
<tr>
<td></td>
<td>Brcapro</td>
<td>124 (29)</td>
<td>62</td>
<td>72.6</td>
</tr>
<tr>
<td></td>
<td>IC</td>
<td>130 (30)</td>
<td>66</td>
<td>77.0</td>
</tr>
</tbody>
</table>

*Probabilities stratified in three groups separately by model, so that families in each group differ across models.
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 $\chi^2$ 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 $\chi^2$ 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.

Analysis of Mendelian models for accuracy in discriminat-
ing 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 HBOC 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 rate in different familial groups. 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.

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

<table>
<thead>
<tr>
<th>Gene</th>
<th>Number of families</th>
<th>Number of mutations</th>
<th>Brcapro model</th>
<th>IC model</th>
<th>$\chi^2$</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1 and BRCA2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBC</td>
<td>357</td>
<td>60</td>
<td>62.7</td>
<td>0.14</td>
<td>–204.2</td>
<td>59.3</td>
</tr>
<tr>
<td>HBOC</td>
<td>151</td>
<td>57</td>
<td>47.7</td>
<td>2.63</td>
<td>–138.8</td>
<td>41.2</td>
</tr>
<tr>
<td>HOC</td>
<td>31</td>
<td>8</td>
<td>6.7</td>
<td>0.31</td>
<td>–16.1</td>
<td>4.0</td>
</tr>
<tr>
<td>MBC</td>
<td>29</td>
<td>8</td>
<td>10.7</td>
<td>1.12</td>
<td>–22.7</td>
<td>11.6</td>
</tr>
<tr>
<td>Total</td>
<td>568</td>
<td>133</td>
<td>127.9</td>
<td>4.20</td>
<td>–381.7</td>
<td>116.0</td>
</tr>
<tr>
<td>BRCA1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBC</td>
<td>357</td>
<td>27</td>
<td>49.0</td>
<td>11.4</td>
<td>–97.2</td>
<td>47.4</td>
</tr>
<tr>
<td>HBOC</td>
<td>151</td>
<td>45</td>
<td>41.2</td>
<td>0.5</td>
<td>–102.7</td>
<td>35.8</td>
</tr>
<tr>
<td>HOC</td>
<td>31</td>
<td>7</td>
<td>6.3</td>
<td>0.1</td>
<td>–13.2</td>
<td>3.7</td>
</tr>
<tr>
<td>MBC</td>
<td>21</td>
<td>1</td>
<td>3.4</td>
<td>2.0</td>
<td>–6.6</td>
<td>3.5</td>
</tr>
<tr>
<td>Total</td>
<td>560</td>
<td>80</td>
<td>99.9</td>
<td>14.1</td>
<td>–219.6</td>
<td>90.5</td>
</tr>
<tr>
<td>BRCA2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBC</td>
<td>276</td>
<td>33</td>
<td>13.7</td>
<td>28.7</td>
<td>–117.4</td>
<td>11.8</td>
</tr>
<tr>
<td>HBOC</td>
<td>131</td>
<td>12</td>
<td>6.6</td>
<td>4.7</td>
<td>–43.1</td>
<td>5.4</td>
</tr>
<tr>
<td>HOC</td>
<td>28</td>
<td>1</td>
<td>0.4</td>
<td>1.1</td>
<td>–3.9</td>
<td>0.3</td>
</tr>
<tr>
<td>MBC</td>
<td>29</td>
<td>7</td>
<td>5.8</td>
<td>0.3</td>
<td>–18.1</td>
<td>5.9</td>
</tr>
<tr>
<td>Total</td>
<td>464</td>
<td>53</td>
<td>26.4</td>
<td>34.8</td>
<td>–182.5</td>
<td>23.5</td>
</tr>
</tbody>
</table>

### Notes

1. The table includes a comparison between the Brcapro and IC models in the entire dataset and by gene, with columns for gene, number of families, number of mutations, Brcapro model, IC model, $\chi^2$, and log likelihood. The data is presented in a tabular format, with rows for different genes and columns for various metrics.

2. The data shows that the Myriad Tables provided better predictions, while the Mendelian models performed worse on average, 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.

3. The table highlights the importance of considering the overall number of mutations, as well as the expected and observed $\chi^2$ statistics and the receiver operator characteristic curve analysis.

4. The 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.

5. The 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.

6. Both models predicted an overall excess of BRCA1 mutations, and this excess was particularly large for HBOC families (which were preferentially mutated in BRCA2).

7. This suggests that current parameterisation of the models still is inadequate to attribute correct probabilities to each gene and that margins for improvement exist.

8. 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.
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 this country, 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 Santoro, 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, Department of Oncology and Biostatistics, Johns Hopkins University, Baltimore, MD, USA
Correspondence to: Dr Presciuttini, Genetica Statistica, c/o Centro Retròvirus, SS Abetone e Brennero 2, 56127 Pisa, Italy; s pres@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
Transforming growth factor-$\beta_1$ 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

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 $\beta_1$ (TGF-$\beta_1$) 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$\rightarrow$C) of the TGF-$\beta_1$ gene that produces a substitution at codon 10 (Leu$\rightarrow$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-$\beta_1$ gene occurs more commonly in control subjects than in individuals with COPD. This allele is associated with increased production of TGF-$\beta_1$ which raises the possibility that TGF-$\beta_1$ has a protective role in COPD.

Evaluation of widely used models for predicting \textit{BRCA1} and \textit{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 and S Presciuttini

doi: 10.1136/jmg.2003.013623

Updated information and services can be found at:
http://jmg.bmj.com/content/41/4/278

\textit{These include:}

References
This article cites 34 articles, 11 of which you can access for free at:
http://jmg.bmj.com/content/41/4/278#BIBL

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections
Breast cancer (239)
Molecular genetics (1254)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/