Supplementary Material

A Risk Prediction Algorithm for Ovarian Cancer Incorporating 
BRCA1, BRCA2, Common Alleles and Other Familial Effects

Sarah Jervis¹, Honglin Song², Andrew Lee¹, Ed Dicks², Patricia Harrington², Caroline Baynes, Ranjit Manchanda⁴,⁵, Douglas F. Easton¹,², Ian Jacobs³,⁴, Paul P.D. Pharoah¹,², Antonis C. Antoniou¹*

1 Centre for Cancer Genetic Epidemiology, Department of Public and Primary Care, University of Cambridge
2 Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge
3 Institute of Human Development, Faculty of Medical & Human Sciences, The University of Manchester and Manchester Academic Health Science Centre
4 Institute for Women’s Health, University College London
5 Department of Gynaecological Oncology, St Bartholomew’s Hospital

*Address for correspondence:

Antonis C Antoniou

Strangeways Research Laboratory

Department of Public Health and Primary Care

Worts Causeway

Cambridge CB1 8RN

UK
Methods

Constraining the overall incidences.

The population incidences of ovarian cancer were constrained to agree with national incidence rates for England and Wales. Let \( i(t) \) denote the population incidence rate at age \( t \) and denote the value of the polygenotype \( P \) by \( p_r \) to highlight its dependence on \( R \), taking value \( r \). Then

\[
i(t) = \frac{\sum_{g, p_r} \Pr(g, p_r) f_{g, p_r}(t)}{\sum_{g, p_r} \Pr(g, p_r) S_{g, p_r}(t-1)}
\]

where \((g, p_r)\) denotes the major-locus genotype \( g \) and polygenotype \( p_r \). \( f_{g, p_r}(t) \) and \( S_{g, p_r}(t) \) are the probability of developing ovarian cancer at age \( t \) and the probability of surviving ovarian cancer by age \( t \). The major-gene genotype and polygenotype are assumed to be independent so that \( \Pr(g, p_r) = \tau_g \phi_{p_r} \), where \( \tau_g \) is the probability of major genotype \( g \) and \( \phi_{p_r} \) is the polygenotype probability, given by the binomial distribution. \( f_{g, p_r}(t) \) can be rewritten as \( \lambda_{g, p_r}(t) \) \( S_{g, p_r}(t-1) \). Thus, using the relationship

\[
S_{g, p_r}(t-1) = \exp\left[\sum_{u=0}^{t-1} \lambda_{g, p_r}(u)\right] = \exp\left[\sum_{u=0}^{t-1} \lambda_{g}(u)e^{p_r}\right] = \exp\left[\sum_{u=0}^{t-1} \lambda_{g}(u)\right]e^{p_r} = S_{g}(t-1)e^{p_r}
\]

\[
i(t) = \frac{\sum_{g, p_r} \tau_g \phi_{p_r} \lambda_{g}(t)e^{p_r} S_{g}(t-1)e^{p_r}}{\sum_{g, p_r} \tau_g \phi_{p_r} S_{g}(t-1)e^{p_r}}
\]
The above equation is then used to estimate the baseline hazard function \( \lambda_b(t) \)

### Including BRCA1 and BRCA2 ovarian cancer incidence rates

The external incidence rates of ovarian cancer in the BRCA1 and BRCA2 mutation-positive populations, denoted \( b_1(t) \) and \( b_2(t) \), are constrained by the following equation, similar to that for general population incidences. \( \lambda_b(t) \) and \( S_b(t) \) are the baseline incidence rate and survival probability of a BRCA\( i \) carrier, “free” of any polygenic effects.

\[
\begin{align*}
    b_i(t) &= \sum_{p_r} \frac{\Pr(p_r) f(t \mid BRCAi, p_r)}{\sum_{p_r} \Pr(p_r) S(t-1 \mid BRCAi, p_r)} = \sum_{p_r} \frac{\phi_{p_r} f_{b_i}(t \mid p_r)}{\sum_{p_r} \phi_{p_r} S_{b_i}(t-1 \mid p_r)} \\
    &= \frac{\sum_{p_r} \phi_{p_r} S_{b_i}(t-1 \mid p_r) \lambda_{b_i}(t) \exp(p_r)}{\sum_{p_r} \phi_{p_r} S_{b_i}(t-1 \mid p_r)} = \frac{\lambda_{b_i}(t) \sum_{p_r} \phi_{p_r} S_{b_i}(t-1)^e^{p_r} \exp(p_r)}{\sum_{p_r} \phi_{p_r} S_{b_i}(t-1)^e^{p_r}}
\end{align*}
\]

Solving this equation for the baseline incidence gives us

\[
\lambda_{b_i}(t) = \frac{b_i(t) \sum_{p_r} \phi_{p_r} S_{b_i}(t-1)^e^{p_r}}{\sum_{p_r} \phi_{p_r} S_{b_i}(t-1)^e^{p_r}},
\]

This can be solved iteratively, starting at \( S_{b_i}(0) = 1 \) and updating the baseline hazard and survival functions alternatively.

The population incidence rate equation becomes:

\[
i(t) = \frac{\sum_{g,p_r} \Pr(g, p_r) f_{g;p_r}(t)}{\sum_{g,p_r} \Pr(g, p_r) S_{g;p_r}(t-1)} = \frac{\sum_{g,p_r} \Pr(g, p_r) f_{g;p_r}(t)}{\sum_{p_r} \phi_{p_r} \left( \sum_{k=0}^{k-1} \tau_k S_k(t-1)^e^{p_r} \right)}
\]

where \( k \) represents the major genotypes (0=non-carrier, 1=BRCA1 carrier, 2=BRCA2 carrier, 3,..K-1= carriers of
other hypothetical ovarian cancer susceptibility genetic variants). Multiplying both sides by the quotient gives:

\[ i(t) \sum_{p_r} \phi_{p_r} \left( \sum_{k=0}^{K-1} \tau_k S_k (t-1)^{r_{p_r}} \right) = \sum_{g, p_r} \Pr(g, p_r) f_{g, p_r} (t) \]

\[ = \sum_{p_r} \phi_{p_r} \left( \lambda_0 \tau_0 S_0 (t-1)^{r_{p_r}} + \sum_{k=3}^{K-1} \lambda_0 \tau_k S_k (t-1)^{r_{p_r}} r(k) \right) e^{p_r} + \sum_{p_r} \phi_{p_r} \left( \tau_1 f_{1, p_r} (t) + \tau_2 f_{2, p_r} (t) \right) \]

Rearrangement of this and use of parts of the \textit{BRCA1} and \textit{BRCA2} incidence rate equations produce the following equation for the baseline hazard function.

\[ \lambda_0 (t) = \frac{i(t) \sum_{p_r} \phi_{p_r} \left( \sum_{k=0}^{K-1} \tau_k S_k (t-1)^{r_{p_r}} \right) - \sum_{p_r} \phi_{p_r} \left( \tau_1 f_{1, p_r} (t) + \tau_2 f_{2, p_r} (t) \right)}{\sum_{p_r} \phi_{p_r} \left( \tau_0 S_0 (t-1)^{r_{p_r}} + \sum_{k=3}^{K-1} \tau_k S_k (t-1)^{r_{p_r}} r(k) \right) e^{p_r}} \]

\[ = \frac{i(t) \sum_{p_r} \phi_{p_r} \left( \sum_{k=0}^{K-1} \tau_k S_k (t-1)^{r_{p_r}} \right) - \sum_{p_r} \phi_{p_r} \left( \tau_1 \lambda_0 (t) S_0 (t-1)^{r_{p_r}} + \tau_2 \lambda_0 (t) S_0 (t-1)^{r_{p_r}} \right) \exp (p_r)}{\sum_{p_r} \phi_{p_r} \left( \tau_0 S_0 (t-1)^{r_{p_r}} + \sum_{k=3}^{K-1} \tau_k S_k (t-1)^{r_{p_r}} r(k) \right) \exp (p_r)} \]

which can be solved as part of a multi-step iterative process along with the baseline \textit{BRCA1} and \textit{BRCA2} hazard functions and the survival functions for each major genotype.

For the major gene and mixed models of inheritance the major gene component was based on three genes: \textit{BRCA1}, \textit{BRCA2}, and a third hypothetical gene. Because the probability of having both a \textit{BRCA1} and a \textit{BRCA2} mutation is very small, we coded \textit{BRCA1} and \textit{BRCA2} as a single locus with three alleles: \textit{BRCA1} positive, \textit{BRCA2} positive, and a normal allele. The third gene was diallelic with a normal and a mutant allele, and was assumed to be unlinked to \textit{BRCA1} and \textit{BRCA2}. For simplicity, \textit{BRCA1} mutations were assumed to be dominant over all other alleles and \textit{BRCA2} mutations.
were assumed to be dominant over hypothetical third locus disease alleles. There were, therefore, five potential risk categories based on the major genotype: BRCA1 carriers, BRCA2 carriers, heterozygotes for the third locus risk allele, homozygotes for the third locus allele and non-carriers.

**Incorporating breast cancer into the model**

Like ovarian cancer, the baseline hazard functions and survival functions are obtained from constraining the national incidence rates.

Under the assumption that breast cancer incidence is independent of the polygenotype in these models, the baseline hazard functions \( \lambda^b_1(t) \) and \( \lambda^b_2(t) \), for the BRCA1 and BRCA2-mutation-positive populations respectively, are just the incidence rates \( inc^b_1(t) \) and \( inc^b_2(t) \).

The breast cancer incidence rates for the general population are constrained by the equation:

\[
i^b(t) = \frac{\sum_{g,p_r} \Pr(g,p_r) f^b_g(t) S^o_g(t) (t-1)}{\sum_{g,p_r} \Pr(g,p_r) S^o_g(t-1) S^b_g(t-1)} = \frac{\sum_g \tau_g f^b_g(t) \sum_{p_r} \phi_{p_r} S^o_g(t-1) (t-1)}{\sum_g \tau_g S^b_g(t-1) \sum_{p_r} \phi_{p_r} S^o_g(t-1) (t-1)},
\]

Multiplying by the right-hand quotient and utilising information on the ovarian cancer survival functions gives the equation

\[
i^b(t) \sum_g \tau_g S^b_g(t-1) \sum_{p_r} \phi_{p_r} S^o_g(t-1) e^{\tau^p_r} = \sum_g \tau_g S^b_g(t-1) \lambda^b_g(t) \sum_{p_r} \phi_{p_r} S^o_g(t-1) e^{\tau^p_r} = \lambda^b_1(t) \left( \tau^p_0 S^b_0(t-1) \sum_{p_r} \phi_{p_r} S^o_0(t-1) e^{\tau^p_r} + \sum_{k=3} \tau^p_k S^b_k(t-1) \sum_{p_r} \phi_{p_r} S^o_k(t-1) e^{\tau^p_r} \right),
\]

\[
+ \lambda^b_2(t) \tau^p_1 S^b_1(t-1) \sum_{p_r} \phi_{p_r} S^o_1(t-1) e^{\tau^p_r} + \lambda^b_2(t) \tau^p_2 S^b_2(t-1) \sum_{p_r} \phi_{p_r} S^o_2(t-1) e^{\tau^p_r}
\]
Rearranging this gives the following equation for the baseline hazard function:

\[
\lambda_0^b(t) = \frac{i^b(t) \sum_{k=0} \tau_k S_k^b (t-1) \sum_{p_r} \phi_{p_r} S_{p_r}^o (t-1)^{x^{p_r}} - \lambda_1^b(t) \tau_1 S_1^b (t-1) \sum_{p_r} \phi_{p_r} S_{p_r}^o (t-1)^{x^{p_r}}}{\tau_0 S_0^b (t-1) \sum_{p_r} \phi_{p_r} S_{p_r}^o (t-1)^{x^{p_r}} + \sum_{k=3} \tau_k S_k^b (t-1) \sum_{p_r} \phi_{p_r} S_{p_r}^o (t-1)^{x^{p_r}}} \]

Incorporating SNPs into the risk prediction algorithm

Given that the risk $P(y_{i+1}^* | y)$ of an individual developing ovarian cancer between ages $t_0$ and $t_f$ is equal to

$$
\sum_{i=t_0}^{t_f} P(y_{i+1}^*)
$$

where $P(y_{i+1}^*)$ is the probability of the family phenotypes including the individual developing OvC at time point $t_i$, the probability $P(y_{i+1}^* | y, P_{Ki})$, of the same event conditional on the family genotype and the observed SNP genotypes, is equal to

$$
\frac{\sum_{i=t_0}^{t_f} P(y_{i+1}^*, P_{Ki})}{P(y, P_{Ki})} = \frac{\sum_{i=t_0}^{t_f} P(y_{i+1}^* | P_{Ki}) P(P_{Ki})}{P(y | P_{Ki}) P(P_{Ki})},
$$

This can then be rewritten in terms of the $n$th total polygenotype of the proband as

$$
\sum_{i=t_0}^{t_f} P(y_{i+1}^*, P_{Ki}) \sum_{n=0}^{2N+1} P(y_{i+1}^* | P_{Ki}) P(P_{Ki} | P^n) P(P^n) = \sum_{i=t_0}^{t_f} \sum_{n=0}^{2N+1} P(y_{i+1}^*, P^n) P(P_{Ki} | P^n),
$$

where $P(P_{Ki} | P^n)$ is the conditional normal density function given by

$$
P_{Ki} | P^n \sim N(p^n \sigma_k^2, \sigma_k^2, \sigma_k^2, \sigma_k^2),
$$

$P(y_{i+1}^* | y)$ can also be written in terms of $P_i$ as

$$
P(y_{i+1}^* | y) = \frac{\sum_{i=t_0}^{t_f} P(y_{i+1}^*)}{P(y)} = \sum_{n=0}^{2N+1} P(y_{i+1}^*, P^n).
$$

Thus the probability of an individual developing OvC conditional on their observed SNP genotype is the ratio of two likelihood function sums, each term of which can be
computed as the corresponding term in the family history-conditional risk probability multiplied by a conditional normal density.

**Distribution of ovarian cancer risk and implications for ovarian cancer prevention.**

Based on our model, given a log-normal polygenic risk of $e^{y_p}$ in the general population, where $y_p$ is predicted to follow a normal distribution with standard deviation $\sigma$ and mean $-\sigma^2/2$, rescaled so that the average risk $E(e^{y_p})$ is equal to 1, it is easily established that the distribution of initial risk among individuals diagnosed with cancer is also log-normal with the log-risk $y_c \sim N\left(\sigma^2/2, \sigma^2\right)$ (see [1], methods). Computing the area under the two normal curves to the right of any given risk point gives us an estimate of the proportion of the population with risk greater than a given level and of the proportion of all cancer cases which will occur within this subgroup. Comparing these values gives a potentially informative measure of the relationship between risk distribution in the population and among cancer cases.

A measure of the predictive power of a 17-SNP genotype risk score could be informative as an indicator of how useful these SNPs are in combination for predicting OvC risk-distribution in the general population – a risk score that can be used to identify a high proportion of all cancers in a relatively low proportion of the population is very useful while one which would need almost half the population to be closely monitored to detect little more than 50% of cancers is almost as costly and ineffective as following 50% of the population at random. A comparison of the predictive power of the SNP risk with that of a total polygenic risk based on explicit family history could also give an indication of how much familial OvC still remains unaccounted for.
The combined log-effects of the seventeen SNPs were assumed to have a normal distribution with variance $V = \sum V_i$ where $V_i = \log\left(\frac{1 - p_i + p_i \exp(2r_i)}{(1 - p_i + p_i \exp(r_i))^2}\right)$ is the variance of the log-risk distribution from the $i^{th}$ SNP, with frequency $p_i$ and log-risk $r_i$ [2] [3]. The proportions of the population and of cancer cases at different levels of SNP risk and polygenic risk were plotted against each other for comparison purposes.
Results: Sample statistics, pathology and genetic information.

### Supplementary Table 1. Sample size, age and case distribution for the probands and their relatives

<table>
<thead>
<tr>
<th></th>
<th>Probands</th>
<th>Mothers</th>
<th>Sisters</th>
<th>Daughters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;=50</td>
<td>&gt;50</td>
<td>Proband&lt;50</td>
<td>Proband&gt;50</td>
</tr>
<tr>
<td>Individuals</td>
<td>415</td>
<td>1133</td>
<td>356</td>
<td>984</td>
</tr>
<tr>
<td>Families (no. with 1 or more)</td>
<td>1548</td>
<td>1340</td>
<td>1404</td>
<td>1144</td>
</tr>
<tr>
<td>Mean age (SD)</td>
<td>43.9 (6.4)</td>
<td>60.4 (5.7)</td>
<td>69.9 (10.8)</td>
<td>74.5 (11.7)</td>
</tr>
<tr>
<td>No. ovarian cancers</td>
<td>415</td>
<td>1133</td>
<td>20</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>1548</td>
<td>51</td>
<td>26</td>
<td>3</td>
</tr>
</tbody>
</table>
Supplementary table 2: SNPs and associated Odds Ratio estimates used in the construction of the Polygenic Risk Score.

<table>
<thead>
<tr>
<th>Locus</th>
<th>SNP</th>
<th>Minor allele frequency</th>
<th>Per-allele odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>3p157</td>
<td>rs15789171</td>
<td>0.049</td>
<td>1.45</td>
</tr>
<tr>
<td>9p16</td>
<td>rs3814113</td>
<td>0.32</td>
<td>0.83</td>
</tr>
<tr>
<td>8p129</td>
<td>rs1400482</td>
<td>0.13</td>
<td>0.85</td>
</tr>
<tr>
<td>19p17</td>
<td>rs4808075</td>
<td>0.30</td>
<td>1.12</td>
</tr>
<tr>
<td>17p40</td>
<td>rs62065444</td>
<td>0.18</td>
<td>1.15</td>
</tr>
<tr>
<td>17p43</td>
<td>rs10069690</td>
<td>0.26</td>
<td>1.09</td>
</tr>
<tr>
<td>5p1</td>
<td>rs2252894</td>
<td>0.66</td>
<td>0.89</td>
</tr>
<tr>
<td>2p176</td>
<td>rs12450786</td>
<td>0.26</td>
<td>1.13</td>
</tr>
<tr>
<td>8p82</td>
<td>rs74544416</td>
<td>0.067</td>
<td>1.20</td>
</tr>
<tr>
<td>17p33</td>
<td>rs3744763</td>
<td>0.59</td>
<td>0.94</td>
</tr>
<tr>
<td>10p22</td>
<td>rs12779865</td>
<td>0.34</td>
<td>1.09</td>
</tr>
<tr>
<td>1p36</td>
<td>rs56318008</td>
<td>0.15</td>
<td>1.11</td>
</tr>
<tr>
<td>1p34.3</td>
<td>rs58722170</td>
<td>0.23</td>
<td>1.08</td>
</tr>
<tr>
<td>4q26</td>
<td>rs17329882</td>
<td>0.24</td>
<td>1.09</td>
</tr>
<tr>
<td>6p22.1</td>
<td>rs1161331104</td>
<td>0.31</td>
<td>0.93</td>
</tr>
<tr>
<td>9q34.2</td>
<td>rs635634</td>
<td>0.19</td>
<td>1.11</td>
</tr>
<tr>
<td>17q11.2</td>
<td>chr17:29181220:15</td>
<td>0.28</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Supplementary Table 3: Predicted number of families with ovarian cancer under each model fitted

<table>
<thead>
<tr>
<th>Model</th>
<th>Only mother diagnosed with Ovarian cancer</th>
<th>Only 1 sister diagnosed with ovarian cancer</th>
<th>Mother and 1 sister diagnosed with ovarian cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed number of families</td>
<td>38</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>Base</td>
<td>19.9</td>
<td>7.98</td>
<td>0.39</td>
</tr>
<tr>
<td>Major Dominant</td>
<td>29.83</td>
<td>10.73</td>
<td>2.20</td>
</tr>
<tr>
<td>Major Recessive</td>
<td>26.76</td>
<td>18.90</td>
<td>1.24</td>
</tr>
<tr>
<td>Major General</td>
<td>29.87</td>
<td>10.74</td>
<td>2.20</td>
</tr>
<tr>
<td>Polygenic</td>
<td>36.27</td>
<td>14.54</td>
<td>1.35</td>
</tr>
<tr>
<td>Mixed Dominant</td>
<td>35.11</td>
<td>13.50</td>
<td>1.95</td>
</tr>
<tr>
<td>Mixed Recessive</td>
<td>33.94</td>
<td>17.86</td>
<td>1.55</td>
</tr>
<tr>
<td>Mixed General</td>
<td>35.61</td>
<td>13.64</td>
<td>2.04</td>
</tr>
</tbody>
</table>

*These assume no other cancers, breast or ovarian cancer in other family members (ie the three scenarios are mutually exclusive)
Supplementary figures

Supplementary Figure 1: Predicted risks of ovarian cancer over time to a BRCA1-carrier born in the 1940 birth cohort by family history.

Supplementary Figure 2: Predicted risks of ovarian cancer over time to a BRCA2-carrier born in the 1940 birth cohort by family history.

Supplementary Figure 3: Estimated ovarian cancer cumulative risk to a 50-year old female born in the 1940 birth cohort in the general population (family history information not considered), by PRS percentile.

Supplementary Figure 4: Estimated ovarian cancer cumulative risk to a 50-year old female born in the 1940 birth cohort with a mother diagnosed with ovarian cancer at 65, by PRS percentile.

Supplementary Figure 5: Estimated ovarian cancer cumulative risk to a 50-year old female born in the 1940 birth cohort with mother and sister diagnosed with ovarian cancer at 65 and 50, by PRS percentile.
References

