The Role of Genome Sequencing in Personalized Breast Cancer Prevention

Supplementary Materials and Methods

This Supplement contains additional information on the following topics: development of the full and partial breast cancer risk scores and their variance in the US population, evaluating the performance of strategies that classify young women into high and low risk groups on the basis of their risk scores, and the genetic susceptibility variants used to develop the partial risk scores.

Breast Cancer Risk Scores

We modeled the lifetime probability of developing breast cancer for an individual with genetic risk score *s* as $R = 1 - \exp\left(-ce^{s}\right)$, where *c* is a positive constant. When the risk scores *s* are not large, the risks of this model can be approximated as $R \simeq ce^{s}$, and for consistency with the results of Pharoah (1,2) we used this approximation in our calculations. We also assumed that the risk scores have Gaussian distributions in the population. The latter assumption is based on the Central Limit Theorem for the sum of the genetic factors contributing to a woman's scores. These assumptions imply that a woman's lifetime genetic risk has approximately a lognormal distribution in the population (1,3).

Full risk scores and their variance

Pharoah (1,2) approximated the variance of the full risk scores of European-American females as $\sigma^2 \sim \log\left(\lambda_{MZ}\right)$, where λ_{MZ} is the ratio of mean lifetime risk among monozygotic twins of breast cancer cases relative to that of the population. The relation follows from the following argument. When the population distribution of log risk is approximately Gaussian with parameters $\left(\mu, \sigma^2\right)$, then the distribution of log risk among breast cancer cases also is approximately Gaussian, but with parameters $\left(\mu + \sigma^2, \sigma^2\right)$ (1). Therefore, since the mean of a

log-normally distributed variable with parameters (μ,σ^2) is $e^{\mu+\sigma^2/2}$, the mean risks in the population and among cases are, respectively, $e^{\mu+\sigma^2/2}$ and $e^{\mu+3\sigma^2/2}$. Letting Y denote an indicator for lifetime breast cancer occurrence, we can thus express the MZ risk-ratio as

$$\lambda_{MZ} \cong \frac{E_S\left[Ie^s \mid Y=1\right]}{E_S\left[Ie^s\right]} = \frac{e^{\mu+3\sigma^2/2}}{e^{\mu+\sigma^2/2}} = e^{\sigma^2},$$

which gives $\sigma^2 = \log \lambda_{MZ}$.

Partial risk scores and their variances

Epidemiologic data suggest that the locus-specific contributions to breast cancer risk combine additively when risk is represented on the logistic scale *log[R/(1-R)]* (4,5). Since the logistic and log scales are roughly equivalent when risk *R* is small, this log-additive model implies that $\log R \sim \log c + s$, where $s = \beta^T g = \beta_1 g_1 + \cdots + \beta_{86} g_{86}$ is a linear combination of genotypes at the 86 unlinked loci listed in Supplementary Table S1, and the coefficients $\beta_1,...,\beta_{86}$ specify the effect sizes of the 86 risk alleles, obtained as the log per-allele hazard-ratios or odds-ratios.

The population variance of the partial risk scores depends on the multi-locus genotype probabilities

$$\Pr\left[G=(g_1,...,g_{86})\right] = \prod_{k=1}^{86} \binom{2}{g_k} p_k^{g_k} (1-p_k)^{2-g_k} \qquad (1)$$

where p_k is the risk allele frequency of the k^{th} locus in Table 1. However calculating these $3^{86} = 10^{41}$ probabilities is computationally infeasible. Instead we approximated the distribution φ by sampling from the set of all 3^K genotype vectors in proportion to their probabilities. We implemented this sampling procedure in the following steps:

1. Sample a genotype vector g_i using the genotype probabilities of equation (1), assuming linkage equilibrium among loci.
2. Label g_i with score $s_i = \beta^T g_i$.
3. Repeat steps 1 and 2 a total of n times to obtain n pairs (g_i, s_i), $i = 1,\ldots,n$.
4. Approximate the distribution $\varphi(s)$ of partial scores by the empirical distribution $\hat{\varphi}(s) = n^{-1} \#\{i : s_i = s\}$.

We used this procedure with a sample of size n = 10^6 genotype vectors to obtain an empirical distribution $\hat{\varphi}(s)$ and used this distribution to estimate the variance of the partial risk scores.

Performance of risk-score-based classification

We evaluated how well the risk scores perform when used to classify a proportion α of the European-American female population as high-risk, for selected values of α. Specifically, we defined a woman as high-risk if her risk score exceeded the $100(1-\alpha)^{th}$ percentile of the relevant population distribution. For the partial risk scores, we took the population distribution to be the empirical distribution $\hat{\varphi}(s)$ described above. For the full risk scores, we approximated the risk for a woman with centered scores as $R \simeq ce^s$ and assumed a Gaussian distribution of risk scores with variance 1.44. We also assumed the mean lifetime risk among European-American females is $\bar{R}$ = 12.68% (6). The performance measures depend on the population mean risk $\bar{R}$ and the relative risk Ψ_α among high-risk women compared to low-risk women. This relative risk is

$$\Psi_\alpha = \frac{\bar{R}_H}{\bar{R}_L} = \frac{1-\alpha}{\alpha} \frac{\int_{z_\alpha \sigma}^{\infty} e^{s - s^2/2\sigma^2}\, ds}{\int_{-\infty}^{z_\alpha \sigma} e^{s - s^2/2\sigma^2}\, ds},$$

where $\bar{R}_L$ and $\bar{R}_H$ denote the mean risks in low-and high-risk subgroups, respectively, z_α denotes the $100(1-\alpha)^{th}$ percentile of the standard Gaussian distribution, and σ^2 denotes the variance of the risk scores in the population. The relation $\bar{R} = (1-\alpha)\bar{R}_L + \alpha\bar{R}_H = \bar{R}_L\left[1+\alpha(\Psi_\alpha - 1)\right]$ gives the mean risks in high- and low-risk groups as $\bar{R}_L = \bar{R}/\left[1+\alpha(\Psi_\alpha - 1)\right]$ and $\bar{R}_H = \Psi_\alpha \bar{R}_L$. The positive predictive value of the classification is $PPV = \bar{R}_H$ and its negative predictive value is $NPV = 1-\bar{R}_L$. Its sensitivity and specificity are $Sn = \alpha\bar{R}_H / \bar{R}$ and $Sp = (1-\alpha)(1-\bar{R}_L)/(1-\bar{R})$. The risk among women classified as low risk, relative to that of the population, is $\left[1+\alpha(\Psi_\alpha - 1)\right]^{-1}$.

References

1. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA. Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet. 2002;31:33-6.
2. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med. 2008;358:2796-803.
3. Begg CB, Pike MC. Comment on "the predictive capacity of personal genome sequencing". Sci Transl Med. 2012;4:135le133; author reply 135lr133.
4. Li H, Beeghly-Fadiel A, Wen W, Lu W, Gao YT, Xiang YB, et al. Gene-environment interactions for breast cancer risk among Chinese women: a report from the Shanghai Breast Cancer Genetics Study. Am J Epidemiol. 2013;177:161-70.
5. Farewell VT. The combined effect of breast cancer risk factors. Cancer. 1997;40:931-6.
6. Howlader N, Noone A, Krapcho M, Garshell J, Neyman N, Altekruse S, et al. SEER Cancer Statistics Review, 1975-2010, National Cancer Institute. Bethesda, MD, based on November

2012 SEER data submission, posted to the SEER web site, April 2013. Available from: http://seer.cancer.gov/csr/1975_2010/.