Discriminatory Accuracy From Single-Nucleotide Polymorphisms in Models to Predict Breast Cancer Risk

Mitchell H. Gail

One purpose for seeking common alleles that are associated with disease is to use them to improve models for projecting individualized disease risk. Two genome-wide association studies and a study of candidate genes recently identified seven common single-nucleotide polymorphisms (SNPs) that were associated with breast cancer risk in independent samples. These seven SNPs were located in FGFR2, TNRC9 (now known as TOX3), MAP3K1, LSP1, CASP8, chromosomal region 8q, and chromosomal region 2q35. I used estimates of relative risks and allele frequencies from these studies to estimate how much these SNPs could improve discriminatory accuracy measured as the area under the receiver operating characteristic curve (AUC). A model with these seven SNPs (AUC = 0.574) and a hypothetical model accuracy measured as the area under the receiver operating characteristic curve from these studies to estimate how much these SNPs could improve discriminatory cancer risk in independent samples. These seven SNPs were located in wide association studies and a study of candidate genes recently identified seven common SNP alleles that confer risk for breast cancer. I calculated how much disease risk factors X will develop breast cancer over a defined time interval. For a short interval, such as 5 years, r(X) is proportional to rr(X) because competing risks of death can be ignored. Thus, r(X) = k[rr(X)], where k is the risk for a woman with relative risk 1.0, which corresponds to the lowest level of risk for all risk factors. Hence, the observation of relative risks in the general population is F_i(t) = \sum X_i P(X) = F_i(t/k).

Affiliation of author: Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD.

Correspondence to: Mitchell H. Gail, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, Rm 8032, Bethesda, MD 20892-7244 (e-mail: gailm@mail.nih.gov).

See “Funding” and “Notes” following “References.” DOI: 10.1093/jnci/djn180

Published by Oxford University Press 2008.
CONTEXT AND CAVEATS

Prior knowledge
Two genome-wide association studies and a study of candidate genes recently identified seven common single-nucleotide polymorphisms (SNPs) that were associated with breast cancer risk in independent samples.

Study design
Estimates of relative risks and allele frequencies from these studies were used to estimate how much these SNPs could improve discriminatory accuracy measured as the area under the receiver operating characteristic curve (AUC). The discriminatory accuracy of these seven SNPs and a hypothetical model with 14 such SNPs were then compared with that of the National Cancer Institute’s Breast Cancer Risk Assessment Tool (BCRAT).

Contribution
The seven-SNP model (AUC = 0.574) and a hypothetical model with 14 such SNPs (AUC = 0.604) have less discriminatory accuracy than the National Cancer Institute’s BCRAT (AUC = 0.607). Adding the seven-SNP model to BCRAT increased the AUC to 0.632.

Implications
Experience to date and quantitative arguments indicate that a huge increase in the numbers of case patients with breast cancer and control subjects would be required in genome-wide association studies to find enough SNPs to achieve high discriminatory accuracy.

Limitations
Individual-level data on case patients and control subjects are needed to investigate interactions that may improve the models. The data used to estimate SNP effects did not permit estimation of interactions among SNPs or between SNPs and risk factors in BCRAT.

From the Editors

of risk in women who develop breast cancer (case patients) is

\[
FD_d(t) = \frac{1}{\sum_{X} r(X) p(X)} \left[ \sum_{X} r(X) p(X) \right]^{-1} \sum_{X} r(X) p(X)
\]

Likewise, the distribution of relative risks in case patients is

\[
FD_c(t) = \frac{1}{\sum_{X} r(X) p(X)} \left[ \sum_{X} r(X) p(X) \right]^{-1} \sum_{X} r(X) p(X)
\]

and it follows that \(FD(t) = FD_d(t/k)\).

The distribution \(F_c(t)\) is shown in Figure 1 for the seven-SNP model. The corresponding mean of \(\log[r(X)]\) (MLRR) is 0.841, with a standard deviation (SDLRR) of 0.262. This SDLRR describes the dispersion of relative risk and risk in the population and is related to discriminatory accuracy (7). A steep slope in the midrange of a locus in Figure 1 corresponds to a small SDLRR.

The curves in Figure 2 are plots of \([1−FD_d(t)]\) (ie, the probability that risk exceeds a given level, \(t\), in case patients) against \([1−F_c(t)]\) (ie, the probability that risk exceeds a given level, \(t\), in the population), as the risk level, \(t\), (not shown) varies from 0 to 1.0. Each point on a locus thus gives the probability that a case patient would have a risk greater than \(t\) on the ordinate and the probability that a member of the general population would have a risk greater than \(t\) on the abscissa. If most of the risk were concentrated in a small proportion of the population, the curve would rise quickly, indicating that most case patients had higher risks than members of the general population. In the curve corresponding to the seven-SNP model in Figure 2, only a fraction \([1−FD_d(t_{0.5})] = 0.606\) of case patients have risks higher than the median risk in the general population, defined by \([1−F_c(t_{0.5})] = 0.5\), indicating poor discrimination for the seven-SNP model. Another measure of discriminatory accuracy, the area under this curve \((6,8,9)\), is 0.574. For a rare disease, such as breast cancer in a 5-year interval, this area is very nearly equal to area under the receiver operating characteristic curve (AUC), which is the probability that a randomly selected case patient has a projected risk greater than that of a randomly selected control (non-case) subject (6). For these discrete risk models, I allow for ties in projected risk by computing the probability that the case risk exceeds the control risk (more precisely the risk in the general population) plus half the probability that the case risk equals the control risk.

To determine whether the modest discriminatory accuracy of the seven-SNP model could be improved, I supposed that there were seven more SNPs with identical properties to the first seven SNPs and that all were in linkage equilibrium. As shown in Figure 2, some improvement in discriminatory accuracy was observed, with an AUC of 0.604. The corresponding distribution of \([\log[r(X)]\) had an MLRR of 1.682 and an SDLRR of 0.371 (Figure 1). Note that 1.682 is twice the MLRR for the seven-SNP model and 0.371 is 2.05 times the SDLRR for the seven-SNP model, as

Figure 1. Cumulative distributions of the log relative risk, \(F_c(t)\), for the seven–single-nucleotide polymorphism (SNP) model (thin dashed line), a 14-SNP model with the original seven SNPs plus seven more SNPs with identical characteristics (thin solid line), the Breast Cancer Risk Assessment Tool (BCRAT; thick dashed line), and BCRAT plus the seven SNPs (thick solid line).
follows from the addition of independent log relative risks (Equation 1).

BCRAT (Gail model 2) is based on age at first live birth, age at menarche, number of first-degree relatives with breast cancer, and number of previous benign breast biopsy examinations. BCRAT has been criticized for lack of discriminatory accuracy (9). I obtained unbiased (weighted) estimates (10) of the joint distribution of these risk factors, X, for white women aged 50 years or older from the 2000 National Health Interview Survey (http://www.cdc.gov/NCHS/nhis/htm; data accessed on July 22, 2002). From the BCRAT relative risks (11), I used the methods described above to calculate an MLRR of 0.520 and an SDLRR of 0.359, corresponding to the thick dashed curve in Figure 1; the AUC was 0.607 (Figure 2). Thus, BCRAT had greater discriminatory accuracy measured by AUC than the seven-SNP model and a slightly greater AUC than the hypothetical 14-SNP model.

By assuming that odds ratios from the seven-SNP model multiplied those from the BCRAT and that the distribution of these SNPs was independent of that of the risk factors in BCRAT, I estimated how much the discriminatory accuracy of BCRAT could be improved by adding the seven SNPs. The resulting distribution of log [\(r(x)\)] (Figure 1) has an MLRR of 1.361 and an SDLRR of 0.445. The AUC increased to 0.632 (Figure 2). In a different population, Chen et al. (12) estimated that adding mammographic density to BCRAT increased the average age-specific AUC by 0.047, from 0.596 to 0.643. The corresponding increase in AUC from adding these seven SNPs to BCRAT was 0.025 (= 0.632 – 0.607). Thus, mammographic density adds more to the discriminatory accuracy of BCRAT than do the seven SNPs.

All the AUC values in these analyses describe the discriminatory power of risk factors, such as SNPs, in women of comparable age over a short interval, such as 5 years. Thus, these AUC values describe the discriminatory accuracy of risk factors apart from age. Some investigators compare case patients and control subjects over large age ranges. Because age is a strong predictor of breast cancer risk and is included in all risk models and because case patients tend to be older than control subjects, doing so increases the AUC value.

This presentation is focused on discriminatory accuracy. High discriminatory accuracy is required for some applications, such as screening for disease (6), but even risk models with modest discriminatory accuracy can be useful for some applications, such as deciding whether or not to take tamoxifen, which decreases the absolute risks of breast cancer and hip fracture but increases the absolute risks of endometrial cancer and stroke (6, 13). For such decision problems, for general counseling, and for designing prevention trials, it is important that the model accurately predict the risk in women with various risk factor combinations, a feature termed “calibration” (6, 9). To assess calibration, one will need to study a cohort to determine how many women develop breast cancer and then compare that number with how many cancers were predicted, overall and in groups of women with various combinations of genotypes and other risk factors. It will be of special interest to determine whether the risks for women with multiple adverse alleles are as high as predicted by the multiplicative model in Equation 1. Positive or negative interactions among such SNP effects or with other risk factors could lead to poor calibration in some subgroups. Although interactions can affect calibration, my unreported calculations indicate that they have little effect on discriminatory accuracy. The generalizability to various racial groups of a risk model that is based on SNPs might be affected by interactions between SNP effects and racial group because the magnitude and even the direction of an association of a marker allele with disease may vary by racial group (3).

The power to detect interactions between pairs of SNPs and between SNPs and other risk factors is limited. A recent study of prostate cancer risk (14) failed to detect such interactions and found that adding information from five SNPs increased the AUC for a model based on age, geographic region, and family history of prostate cancer by only 0.009, from 0.624 to 0.633. Another study of prostate cancer failed to demonstrate statistically significant interactions among disease-associated SNPs from seven different genomic regions (15). It would be
of interest to search for interactions of the effects of common SNP alleles on breast cancer risk with age, as have been found for rare high-risk mutations in BRCA1 and BRCA2 (16).

To build a model of absolute risk, one can couple the relative risk estimates from case–control data in genome-wide association studies with cancer incidence rates from registry data, as described previously (5,11). To do so requires data on the joint distribution of all risk factors in representative case or control subjects. In my analysis, it was assumed that the SNP genotypes were mutually independent and also independent of the factors in BCRAT. The effect of positive correlations between these SNPs and family history of breast cancer, if any, would be to diminish the discriminatory accuracy that these SNPs add to BCRAT because family history is included in BCRAT.

Very large relative risks are needed for a single factor to achieve good discriminatory accuracy (17). Even adding a strong risk factor with a large attributable risk, such as mammographic density, only increased the AUC of a model like BCRAT from 0.596 to 0.643 (12). Thus, it is not surprising that adding seven SNPs with small relative risks would increase the AUC of BCRAT only modestly.

It is tempting to speculate on how much additional discriminatory accuracy can be achieved by identifying further common SNPs and what effort would be required to find them. Pharoah et al. (7) assumed that the natural logarithm of risk was normally distributed, which provides a good approximation if many independent SNPs satisfy Equation 1 and if risk is proportional to relative risk, as was assumed in my analysis. Based on segregation analyses (18) and considerations of the recurrence risk among siblings, Pharoah et al. (7) estimated an SDLRR of 1.2 in the general population and showed that the logarithm of risk in case patients would be normally distributed with the same variance but with the mean increased by $1.2^2 = 1.44$. From these values, I calculated an AUC of 0.800. This result supports arguments (7) that knowing which SNPs give rise to this polygenic component of risk (which is independent of risk from BRCA1 and BRCA2 mutations) might have some value for screening the population. The seven-SNP model has an SDLRR of 0.262. To achieve an SDLRR of 1.2, one would need 147 [= 7(1.2/0.262)^2] SNPs like the seven SNPs already identified. The geometric mean of the per allele odds ratios from these seven SNPs was 1.15. The study by Easton et al. (2) used approximately 400 case patients with strong family histories of breast cancer in the SNP discovery phase, which might be equivalent in statistical power to approximately 1600 population-based case patients (19). Stacey et al. (3) used 1600 population-based case patients in the discovery phase. Calculations as in Gail et al. (20) show that approximately 65% of disease-associated SNPs with an odds ratio of 1.15 would have among the 25 000 smallest P values in a scan of 500 000 SNPs if 1600 case patients and control subjects are used in the discovery phase. Thus, increasing the number of case patients and control subjects in the discovery phase to 5000 or more (20) might increase the number of such SNPs that would eventually be confirmed in subsequent phases to 11 (= 7/0.65). Improvements in SNP chip technology might yield a few more such SNPs, but even a 50% increase would yield only 17 (= 11 × 1.5) SNPs. There are probably many other disease-associated SNPs with smaller odds ratios, but their detection will require larger numbers of case patients and control subjects both in the discovery and validation phases. For example, if remaining disease-associated SNPs have a geometric mean OR of 1.10, one would need (20) approximately 2.15 [= (log(1.15)/log(1.10))^2] times as many case patients and control subjects in the discovery phase as was required for an OR of 1.15. The contribution of an SNP to the variance of the log relative risk is $2(1-\rho)[log(OR)]^2$. It follows that if 10 additional SNPs can be identified with properties like those of the seven SNPs found so far but the rest of the SNPs have an OR of 1.10, one will need to find about 280 [= (147 − 17) × 2.15] additional low-risk SNPs to achieve the desired SDLRR of 1.2. Although these numbers are only illustrative, they show that a huge increase in the numbers of case patients and control subjects would be required in genome-wide association studies to find enough SNPs to achieve an SDLRR of 1.2.

This study had several limitations. To investigate interactions that may improve the models, individual level data on case patients and control subjects are needed. The published data (2–4) used to estimate SNP effects did not permit estimation of interactions among SNPs or between SNPs and risk factors in BCRAT. Several assumptions were needed to speculate on prospects for finding additional common disease-associated alleles that will achieve high discriminatory accuracy. Further research may indicate the extent to which these assumptions and the resulting broad conclusions hold.

References


**Funding**

Intramural Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute and National Institutes of Health.

**Notes**

I would like to thank Sir Bruce A. J. Ponder for stimulating discussions leading to this work, Dr Montserrat Garcia-Closas for discussions on an empirical study to evaluate risk prediction models including single-nucleotide polymorphisms with other risk factors, the reviewers and Dr Ruth M. Pfeiffer for helpful comments, and Mr David Pee for providing estimates of the distribution of risk factors for Breast Cancer Risk Assessment Tool from the 2000 National Health Interview Survey.

Manuscript received February 11, 2008; revised May 2, 2008; accepted May 6, 2008.