Cancer Epidemiology, Biomarkers & Prevention

Predicting the Future of Genetic Risk Prediction

Nilanjan Chatterjee, Ju-Hyun Park, Neil Caporaso, and Mitchell H. Gail

Recent genome-wide association studies (GWAS) have identified many susceptibility loci for a variety of complex traits. There is intense interest in evaluating the potential utility of these variants for individualized risk prediction for chronic diseases such as breast cancer. A recent article in this journal (1) evaluated the potential discriminatory accuracy of breast cancer risk models that use a variety of genetic variants, some of which, but not all, were found through GWAS. We comment on the methodology and conclusions of that paper and compare it with other work that has assessed the potential role of genetic variants for risk prediction and considered the possible impact of discoveries of new variants in future studies.

The paper by van Zitteren et al. (1) used simulations to estimate the area under the operating characteristic curve (AUC) from breast cancer risk models based on single nucleotide polymorphisms (SNPs). From a review of meta-analyses and GWAS, they found 96 variants, 41 of which were nominally significant at the 0.05 level. They estimated the AUC as 0.67 based on these 41 variants and 0.68 based on all 96 variants. They also commented on the numbers of additional variants that would be needed to increase the AUC to a higher value, such as 0.80.

The work of van Zitteren et al. is interesting and surprising in several respects. Previous studies based on SNPs associated with breast cancer in GWAS yielded estimates of AUC of 0.574 for 7 SNPs (2), 0.579 for 10 SNPs, (3) and 0.585 for 11 SNPs (4). The SNPs in these studies attained genome-wide significance ($P < 10^{-7}$) and odds ratios per allele were based on independent data, thus reducing or eliminating the bias found in odds ratios in the discovery phase of GWAS, known as the "winners's curse" (5). It is very difficult to increase the AUC from 0.58 to 0.67 (2, 6).

Park et al. (7) examined the prospects for increasing the AUC based on further GWAS. They estimated the likely number and the distribution of effect sizes for yet to be discovered SNPs based on the effect sizes for known susceptibility loci and the power of the original GWAS that led to these discoveries. For breast cancer, they estimated that within the range of effect sizes of known loci, there could be a total of 67 common susceptibility SNPs, including those already identified, which is almost 4 times the number of known SNPs. However, the incremental contribution of the undetected SNPs to the genetic variance is likely to be modest as their effect sizes will tend to concentrate toward the lower end of the range. In particular, Park et al. estimated that the total of 67 susceptibility SNPs could explain a total of 17.1% of genetic variance of breast cancer and could lead to an AUC of only 0.635. Because detection of most of these additional SNPs at a genome-wide significance would require very large sample sizes, they concluded that an AUC of 0.635 would be the practical upper limit of the discriminatory power of risk model for breast cancer on the basis of common susceptibility SNPs alone.

Thus, it is impressive that van Zitteren et al. report an AUC of 0.67 based on 41 nominally significant variants. On the basis of a log-normal approximation of risk (8), we estimated that to achieve an AUC of 0.67, a genetic risk model will need to explain 27.9% of known heritability of breast cancer corresponding to a sibling recurrence risk of 2.0. In contrast, the 18 SNPs found in GWAS studies in Table 1 account for only 7.1% of the heritability. There has been recent speculation (9) on reasons why the SNPs from GWAS account for such a small portion of the theoretical heritability that one would anticipate from studies of familial aggregation of breast cancer (8, 10). Some possibilities are variants that contribute to risk but are not tagged by common SNPs in standard GWAS analyses. Such variants might include deletions, repeats, copy number variations, and uncommon and rare SNPs that are not in tight linkage disequilibrium with SNPs accessible with a GWAS platform, such as the Illumina Infinium 660K array. Thus, it is possible that van Zitteren et al. have tapped into some of the "missing heritability" to achieve an AUC of 0.67 from nominally significant variants on the basis of meta-analytic literature; some of these variants are deletions and repeats that might not be detected in GWAS. We will return to this possibility in discussing Table 1, which classifies the variants in van Zitteren et al. partly in terms of the ability of GWAS to detect them.

Another possibility is that some of the variants used by van Zitteren et al. are not truly associated with breast cancer and represent chance findings with possibly exaggerated odds ratios. Many of the meta-analyses cited by van Zitteren et al. focused on a "candidate gene." Findings from many early candidate gene studies were not reproducible, and a number of authors sought to explain why (11, 12). Important reasons for these failures were the

Authors' Affiliation: Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institute of Health, Rockville, Maryland

Corresponding Author: Nilanjan Chatterjee, 6120 Executive Blvd, EPS 8052, Rockville MD 20852. Phone: 301-402-7933; Fax: 1-301-402-0081; Email: chattern@mail.nih.gov

doi: 10.1158/1055-9965.EPI-10-1022

Table 1. Three categories of variants associated with breast cancer risk in van Zitteren et al. (1)

Gene	Variant in van Zitteren	Variant in GWAS literature	Allele frequency in van Zitteren	Per allele odds ratio in van Zitteren	Per allele odds ratio in GWAS literature	*P* value in GWAS[a]	Power to detect SNP in GWAS[a,b]
Category 1: 18 SNPs found in GWAS							
1p11	**rs11249433 (T/C)**	**same**	0.3935	1.16	1.16	1×10^{-2}	
2q35	**rs13387042 (G/A)**	**same**	0.500	1.11	1.12	2×10^{-10}	
3p24 SLC4A7	**rs4973768 (C/T)**	**same**	0.470	1.12	1.11	6×10^{-7}	
5p12	**rs2067980 (A/G)**	**rs10941679**	0.160	1.08	1.19	3×10^{-3}	
5p12	**rs7716600 (C/A)**		0.2205	1.10			
6q25	**rs2046210 (G/A)**	**same**	0.3645	1.36	1.15	2×10^{-5}	
8q24	**rs13281615 (T/C)**	**same**	0.40	1.06	1.08	2×10^{-5}	
Chr 9	**rs1011970 (C/A)**	**Turnbull[a]**	0.1675	1.07	1.09	3×10^{-8}	
Chr 10	**rs2380205 (G/A)**	**Turnbull[a]**	0.4295	0.95	0.94	5×10^{-7}	
Chr 10	**rs10995190 (G/A)**	**Turnbull[a]**	0.149	0.84	0.86	5×10^{-15}	
Chr 10	**rs704010 (G/A)**	**Turnbull[a]**	0.3945	1.11	1.07	4×10^{-9}	
Chr 11	**rs614367 (G/A)**	**Turnbull[a]**	0.152	1.16	1.15	3×10^{-15}	
14q24	**rs999737 (T/C)**		0.2385	0.94			
17q23	**rs6504950 (G/A)**		0.2700	1.04			
FGFR2	**rs2981582 (C/T)**		0.379	1.18	1.26	4×10^{-31}	
LSP1	**rs3817198 (T/C)**		0.300	1.06	1.07	6×10^{-4}	
MAP3K1	**rs889312 (A/C)**		0.2795	1.13	1.12	5×10^{-9}	
TNRC9	**rs3803662 (C/T)**		0.2505	1.23	1.19	3×10^{-15}	
Category 2: 13 SNPs not found in GWAS							
AKAP9	**M463I (G/T)**		0.366	1.08			0.087
CCND1	**G870A**		0.4595	1.12			0.641
CYP1B1	**Val432Leu (G/C)**		0.5445	1.60 (het) 1.1 (hom)			1.00
eNOS	**G894T**		0.2265	1.22 (recess)			0.010
eNOS	**T(-786)C**		0.8215	0.74			1.00
ESR1	**rs3020314 (T/C)**		0.3105	1.05			0.002
ESR2	**rs4986938 (A/G)**		0.361	0.94			0.017
FAS	**G(-1377)A**		0.148	1.18			0.710
HER2	**I655V (G/A)**		0.1975	1.05			0.000
IGFBP3	**A(-202)C**		0.4895	1.03			0.000
NBS1	**G8360C**		0.320	0.97			0.000

(Continued on the following page)

Table 1. Three categories of variants associated with breast cancer risk in van Zitteren et al. (1) (Cont'd)

Gene	Variant in van Zitteren	Variant in GWAS literature	Allele frequency in van Zitteren	Per allele odds ratio in van Zitteren	Per allele odds ratio in GWAS literature	*P* value in GWAS[a]	Power to detect SNP in GWAS[a,b]
WDR79	**R68G C/G)**		0.1205	1.08			0.004
XRCC1	**R399Q (A/G)**		0.660	1.12 (recess)			0.078
Category 3: 10 non-SNP variants							
AR	**CAG repeat**		0.539	0.61 (dom.)			
CASP8	**D302H**		0.444	0.89	0.88	0.14 (ns)	
CHEK2	**1100delC**		0.500 binary	2.40			
Cyp19	**$TTTA_{10}$**		0.024 binary	1.52			
ERCC2	**A751C**		0.634 binary	1.13			
GSTM1	**Deletions**		0.565 binary	1.10			
GSTT1	**Deletions**		0.493 binary	1.11			
MTHFR	**C677T**		0.336 binary	1.04			
TGFB1	**L10P**		0.5265	1.05			
VDR	**Fok1**		0.146 binary	1.14			

[a]The GWAS was reported in ref. 14.

[b]Power calculations for category 2 SNPs ($n = 13$) were made at significance level 10^{-7} for the 2-stage design of Turnbull et al. (14). Power calculations for SNPs with recessive and codominant effects were based on a 2 degree-of-freedom test. For all other SNPs, the power calculations were done for the 1 degree–of-freedom trend test under a model that assumes that the relative odds are 1, OR, or OR^2, for per allele odds ratio OR.

limited power and lack of stringency of significance levels used, which yielded low positive predictive value for the studies, and hence many false positive findings. Most of these studies did not include an independent confirmatory study of the findings, which would have limited false positive reports. The situation was so problematic that scientists and journal editors developed guidelines and recommendations for reporting on genetic association studies (13). The use of meta-analyses on the basis of multiple studies with large cumulative numbers of cases and controls, as in Table 2 of van Zitteren et al., ameliorates some of these problems, but the possibility of false positives cannot be ruled out due to the low significance threshold ($P < 0.05$) used for inclusion of the variants in the analysis. Further, publication bias can also create false positives and estimates of exaggerated odds ratios in large meta-analyses.

To explore the contributions of different types of variants in the calculation of AUC by van Zitteren et al., we divided them into 3 categories (Table 1). (a) We considered a set of 18 SNPs that have been associated with risk of breast cancer with genome-wide significance (typical $P < 10^{-7}$) in recent GWAS. (b) We considered a set of 13 common candidate SNPs that were not detected in GWAS, but reached borderline significance ($P < 0.05$) in the meta-analysis of van Zitteren et al. If these associations were real, recent GWAS should have had sufficient power to detect some of these variants through tagging SNPs in the same region. (c) We considered a set of 10 non-SNP variants that may not be detectable in GWAS even if the associations were real.

Using a log-normal model for risk, as in Pharoah et al. (8), we estimated that SNPs in category 1 lead to an AUC of only 0.587. This estimate is very near the value 0.585 calculated by Gail (4) for 11 known SNPs and lies between the value of AUC obtained by Wacholder et al. (3) for 10 known SNPs, and the value obtained by Park et al. for 67 potentially detectable SNPs. Adding the SNPs in category 2 further increases the AUC to 0.623, and further addition of the non-SNP variants to the model leads to an AUC of 0.640. A refined calculation for the 41 variants that treats the 8 combinations of *AR*, *CHEK2*, and *Cyp19* exactly, and uses the log-normal approximation for the other variants with smaller odds ratios yields AUC of 0.646. Increases in AUC from 0.587 to 0.623 and again to 0.646 represent impressive gains. Thus, these comparisons show that the variants in categories 2 and 3 that have not been reported in GWAS have made a substantial contribution to the estimate of AUC given by van Zitteren et al. Whether such an AUC value would be reproducible in an independent data set will depend on how many of the variants listed in categories 2 and 3 reflect true association, and whether the meta-analyses of van Zitteren et al. yielded unbiased estimate of risk for these variants.

To understand whether the SNPs in category 2 in Table 1 are truly associated with breast cancer risk, we calculated the power of the large GWAS by Turnbull et al. (14) to detect them at a genome-wide significance level of $P = 10^{-7}$. We used the odds ratios in van Zitteren et al. (and Table 1) and assumed that these SNPs were perfectly correlated with tagging SNPs used by Turnbull et al. Although the power was less than 0.10 for 9 of the 13 SNPs, the powers were 0.64, 1.00, 1.00, and 0.71 for *CCND1*, *CYP1B1*, *eNOS*, and *FAS*, respectively. The fact that none of these SNPs was detected in the GWA study of Turnbull et al. raises the suspicion that some of the associations in this category will not be confirmable.

Our methods for calculating AUC yield qualitatively similar results to those in van Zitteren et al., but our AUC of 0.646 for all 41 variants is substantially lower than the value 0.67 that they report. Both analyses assume that relative odds multiply across loci and that genotypes at the various loci are independent. These assumptions justify the log-normal approximation by an application of a central limit theorem. However, there are some differences in the methods used. First, we reviewed the literature and chose different odds ratios than used by van Zitteren et al. in a very few instances (Table 1). For example, for *6q25*, we used the odds ratio 1.15 instead of 1.36, because the former corresponds to a Caucasian population and the latter to an Asian population. Second, we calculated the AUC using the odds ratios in Table 1, whereas van Zitteren et al. reported an average AUC obtained by resampling odds ratios at each locus. Because the AUC is a nonlinear function of these odds ratios, the mean exceeds the value we calculated. From unreported simulations, we estimated that this phenomenon accounts for an increase of approximately 0.012 in the AUC compared with our estimate based on the original odds ratios. We regard this increment of 0.012 as an upward bias, because it is induced by sampling error in the relative risk estimates rather than by random variation in the underlying true relative risks. With the method of van Zitteren et al., smaller studies with less precise estimates of relative risk would yield higher AUC estimates than larger studies with the same point estimates of relative risk. The remaining difference, $0.670 - 0.646 - 0.012 = 0.012$ might be accounted for by some differences in the modeling that we do not understand because of our difficulty in reconstructing some of the modeling steps followed by van Zitteren et al.

Even higher AUC levels than the value 0.67 found by van Zitteren et al. would be desirable for many public health applications, such as defining subsets of the population for which a chemopreventive agent like tamoxifen might be recommended, or allocating public health resources (4). To investigate the feasibility of attaining higher AUC values, van Zitteren et al. calculated that if 2, 3, 4, or 5 times as many variants are found in the future with an odds ratio distribution similar to that of the 41 variants in Table 1, then the AUCs could be 0.73, 0.77, 0.80, and 0.82 respectively. Moreover, they described

combinations of numbers of variants and odds ratios needed to bring the AUC of the model to 80%. One can assess how realistic these scenarios are from current GWAS data. As Park et al. noted, more susceptibility SNPs are likely to exist, but the distribution of effect sizes for the undetected loci will be shifted toward smaller values than previously detected loci. Thus, simulating additional variants on the assumption that their effect size distribution will be similar to that of known loci, as in van Zitteren et al., leads to a substantial overestimate of AUC for any given number of assumed loci. Another way to see whether a particular combination of number of variants and odds ratios is realistic would be to compare the number of expected discoveries under such a model to number of observed discoveries in recent GWAS. For example, we estimated that if there were 100 SNPs each with odds ratio 1.3 and minor allele frequency 0.3, a scenario that will lead to an AUC of 0.80 according to Table 3 in van Zitteren et al., one would expect all of these SNPs to be detected at $P < 10^{-7}$ in the GWA study of Turnbull et al. (14), who reported only 12 SNPs, including 4 novel SNPs, at this significance level.

van Zitteren et al. used the AUC statistic as the main metric for evaluating utility of risk models. Several authors have recommended other criteria based on risk-benefit analyses in specific applications as more appropriate (4, 15–18). Nonetheless, the ordering of AUC values gave a good indication of how various risk models would perform in several risk-benefit applications in public health (4). van Zitteren et al. commented that SNPs and other variants might be combined with other factors to improve discriminatory power (AUC). Previous work has shown that adding mammographic density to standard epidemiologic risk factors, such as family history, increases AUC even more than adding SNPs (2). Presumably, adding SNPs and other variants as well as mammographic density could increase AUC even more. Adding *BRCA1/2* mutations with carrier frequency 0.0032 and relative risk 10, based on data in ref. 19, we found that the AUC increased from 0.646 without *BRCA1/2* to 0.655 with *BRCA1/2* in our models for the general population. To achieve a large increase in AUC will require including strong, fairly common risk factors, such as mammographic density. It is possible that well-validated strong biomarkers, such as sensitivity to mutagens, might contribute, as has been seen for lung (20) and bladder cancer (21). For women with biopsies, detailed pathologic and molecular data may yield strong risk factors (22). Perhaps one day a nearly painless way to obtain such pathologic or molecular data will be found that would be widely applicable and would increase the discriminatory accuracy of breast cancer risk predictions for women in the general population.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Grant Support

The research was supported by the intramural program of the National Cancer Institute, National Institute of Health.

Received September 27, 2010; accepted November 12, 2010; published online January 6, 2011.

References

1. van Zitteren M vdNJ, Kundu S, Freedman AN, van Duijn CM, Janssens ACJW. Genome-based prediction of breast cancer risk in the general population: a modeling study based on meta-analyses of genetic associations. Cancer Epidemiol Biomarkers Prev 2010;20:9–22.
2. Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. J Natl Cancer Inst 2008;100:1037–41.
3. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, , et al. Performance of Common Genetic Variants in Breast-Cancer Risk Models. N Engl J Med 2010;362:986–93.
4. Gail MH. Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. J Natl Cancer Inst 2009;101:959–63.
5. Zollner S, Pritchard JK. Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet 2007;80:605–15.
6. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 2004;159:882–90.
7. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42:570–5.
8. Pharoah PDP, 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.
9. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 2010;11:446–50.
10. Pharoah PDP, Antoniou AC, Easton DF, Ponder BAJ. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med 2008;358:2796–803.
11. Caporaso NE. Why have we failed to find the low penetrance genetic constituents of common cancers?Cancer Epidemiol Biomarkers Prev 2002;11:1544–49.
12. Wacholder S, Chanock S, Garcia-Closas M, El ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 2004;96:434–42.
13. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, et al. Replicating genotype-phenotype associations. Nature 2007;477:655–60.
14. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 2010;42:504–7.
15. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. Biostatistics 2005;6:227–39.

16. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc Ser A 2009;172: 729–48.
17. Pauker SG, Kassirer JP. Therapeutic decision-making—cost-benefit analysis. N Engl J Med 1975;293:229–34.
18. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26: 565–74.
19. Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, et al. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. Br J Cancer 2008;98: 1457–66.
20. Spitz MR, Etzel CJ, Dong Q, Amos CI, Wei Q, Wu X, et al. An expanded risk prediction model for lung cancer. Cancer Prev Res 2008;1:250–4.
21. Wu X, Lin J, Grossman HB, Huang M, Gu J, Etzel CJ, et al. Projecting individualized probabilities of developing bladder cancer in white individuals. J Clin Oncol 2007;25:4974–81.
22. Hartmann LC, Sellers TA, Frost MH, Lingle WL, Degnim AC, Ghosh K, et al. Benign breast disease and the risk of breast cancer. N Engl Med 2005;353:229–37.