Validation of a Polygenic Risk Score for Breast Cancer in Unaffected Caucasian Women Referred for Genetic Testing

March 2018

Mary Helen Black, Shuwei Li, Holly Laduca, Jefferey Chen, Robert Hoiness, Stephanie Gutierrez, Hsiao-Mei Lu, Jill S. Dolinsky, Brigette Tippin-Davis

Ambry Genetics, Aliso Viejo, California

Abstract

BACKGROUND: While assessment of genetic contribution to breast cancer (BC) risk was once limited to high-penetrance genes such as *BRCA1/2*, genome-wide association (GWA) studies have identified many single nucleotide polymorphisms (SNPs), each with a modest contribution to BC risk. Several reports suggest that a score based on combined genotypes across a large number of SNPs may have substantial predictive value for risk stratification in the general population. However, few studies have examined the performance of such a score in high-risk women.

METHODS: We genotyped 100 BC-associated SNPs using next-generation sequencing in order to examine whether a risk score based on these SNPs was predictive of BC in 3,020 women (1,772 BC cases and 1,248 controls) referred for genetic testing at a single diagnostic laboratory. All women (mean±SD age at testing 52±13 years) were self-reported Caucasian, 18-84 years of age, provided family history information at the time of testing, and tested negative for pathogenic variants in BC-susceptibility genes. We constructed a population-standardized polygenic risk score (PRS), with each SNP weighted by per-allele relative risks in Caucasians from large genome-wide association studies and population-specific allele frequencies, and used logistic regression to test PRS association with BC.

RESULTS: The PRS was significantly higher in cases than controls (mean±SD 1.20±0.88 vs. 0.95±0.69, $p<0.0001$). Compared to women in the 1st quartile of PRS, those in the 2nd, 3rd and 4th quartile were 1.51 (95% CI: 1.23-1.87), 2.06 (95% CI: 1.67-2.55) and 2.69 (95% CI: 2.17-3.35) times as likely to have BC (all $p<0.0001$). PRS predictive performance was consistent with prior literature (AUROC=0.61).

CONCLUSION: These data suggest that a 100-SNP PRS assessed in high-risk patients performs similarly to risk scores reported in the broader population, which could have direct implications for their clinical management. Our ongoing analysis of the ability of the PRS to discriminate among specific pathologic subtypes, as well as validity and utility of a PRS combined with clinical models to estimate residual lifetime risk, has the potential to further inform screening guidelines and improve patient care.

Introduction

Since 2007, genome-wide association (GWA) studies have identified many common single nucleotide polymorphisms (SNPs), each with a modest contribution to breast cancer (BC) risk[1]. As these SNPs are associated with relative risks ranging from 1.03-1.41[2], no individual SNP is informative on its own. However, a score based on combined genotypes across a large number of SNPs may have substantial predictive value for risk stratification[3-7]. While the utility of such a score has been investigated in large studies conducted in the general population, few have assessed its performance in high-risk women referred for genetic testing[8, 9].

In fact, SNP-based scores may have clinically useful predictive power in women referred for genetic testing due to family history of disease. Sawyer *et al.* examined a 22-SNP polygenic risk score (PRS) comparing women ascertained from a familial cancer clinic with BC, who were either *BRCA1/2* carriers or *BRCA1/2* negative, to a set of controls[9]. They found that *BRCA1/2* negative cases had a significantly higher PRS than *BRCA1/2* carriers or controls, and that *BRCA1/2* negative cases in the highest quartile of the PRS distribution were more likely to have had early-onset BC (<30 years of age) and/or a second BC compared to those with a score in the lowest PRS quartile. Li *et al.* assessed a 24-SNP PRS among unaffected women from two familial BC cohorts who were prospectively followed for an average of 7.4 years, and observed that women in the highest quintile of the PRS distribution were more than three times as likely to develop BC as those in the lowest quintile[8]. While all of the women included were enriched for family history of BC, half were known to be negative for high-penetrance genes such as *BRCA1/2*, *PALB2* and *ATM*. Taken together, these data suggest that a SNP-based PRS may be most useful for risk stratification in women with family history of BC who are negative for high-penetrance BC-susceptibility genes.

In this report, we examined whether and to what extent a PRS, based on the combined effects of 100 SNPs previously reported in multiple large GWAs, is predictive of BC in high-risk women referred for genetic testing who tested negative for pathogenic or likely pathogenic variants in known BC-susceptibility genes.

Methods

Patient population

Patients were referred to Ambry Genetics for testing in 2015-2018, and were eligible for study inclusion if they were female, self-reported Caucasian race (non-Ashkenazi Jewish), 18-84 years of age at the time of testing, and provided family history information to ordering clinicians. Those who tested positive in the present study for a pathogenic or likely pathogenic in a BC-susceptibility gene (*ATM, BARD1, BLM, BRCA1, BRCA2, BRIP1, CDH1, CHEK2, FANCC, MRE11A, NBN, NF1, PALB2, PTEN, RAD50, RAD51C, RAD51D, STK11, TP53*) were excluded. Cases were identified as those with a personal history of BC, and were excluded if clinical history included other cancer primaries. Controls were unaffected with any cancer (not including basal or squamous cell carcinoma); those with a first- or second- degree relative with breast or ovarian cancer were further excluded from analysis.

Molecular Analysis

Sequencing quality for Illumina NextSeq 500 are monitored during the sequencing run, and include visualization of Intensity-vs-Cycle (IVC) plots, and cluster intensity over the duration of the run. Other quality metrics that are evaluated for the entire sequencing run upon completion of sequencing and demultiplexing of the samples include metrics for the % Perfect Index Reads, % of ≥ Q30 Bases, and overall Mean Quality Score. Samples passing the sequencing quality metrics were fed into proprietary NGS data processing pipeline in a parallelized fashion, starting with alignment of sequencing reads to human reference genome build (GRCh37/hg19), followed by variant and genotype calling on the panel genes and the 100 BC-associated SNP positions. Additionally, NGS coverage is evaluated for all 100 BC-associated SNPs for every sample, and any SNPs with no or low coverage (<20X) were excluded from genotype calling, and were not included in downstream statistical analysis.

Statistical Analysis

SNPs that met the following criteria were selected for inclusion in the analysis: 1) reported with genome-wide significance in >1 GWA analysis, based on a sample size of >500 cases and >500 controls in any population; 2) are not strongly correlated; 3) effects are race/ethnicity-specific (i.e. only SNPs that meet 1) and 2) in a specific population were included). For this study, we selected SNPs reported in studies of individuals of N. European ancestry[1, 2, 10-18]. We identified 100 SNPs meeting these criteria, of which 3 could not be directly genotyped and were substituted by tag SNPs (pairwise r^2 was 0.9-1.0 and median distance was 1,062 bp between each proxy and originally reported SNP).

NGS data were examined to assess missing rates for each sample, and each SNP. Samples were excluded if >10 SNPs were missing due to bioinformatics quality control thresholds (n=12; 0.4% of samples). SNP calls were checked for consistency with publically available databases (GRCh37/hg19; Ensembl release 91[19]) and literature-reported reference and risk alleles. We compared SNP allele frequencies among control subjects to those available in the 1000 Genomes EUR population to ensure consistency with the reference population. Hardy Weinberg Equilibrium (HWE) was assessed for all SNPs among controls using R package HardyWeinberg[20]. To assess the assumption of SNP effects consistent with a log additive model, we examined all possible pair-wise SNP*SNP interactions using logistic regression, with a 1-df test for the interaction and BC as the outcome. We additionally tested for higher-order SNP interactions using logic regression[21].

Using an approach consistent with prior literature[4, 5, 22, 23], we computed a SNP-based population-standardized PRS for each patient. Using previously published estimates of the per-allele odds ratio (OR) and risk allele frequency (p) for each SNP, and assuming independent and additive risks on the log OR scale, we computed the unscaled population average risk as:

$$\mu = (1 - p)^2 + 2p(1 - p)OR + p^2OR^2$$

Adjusted risk values were then calculated as:

$$\frac{1}{\mu}, \frac{OR}{\mu}, \frac{OR^2}{\mu}$$

for the 3 genotypes defined by the number of risk alleles: 0, 1 or 2, respectively. Missing genotypes were assigned a population average risk of 1.0. Adjusted risk values for each SNP were multiplied to compute the overall PRS-associated risk for each individual based on his/her observed genotypes. Logistic regression models were used to estimate the ORs for BC by quartile of the PRS, with the 1st quartile category (<25th percentile) as the reference. We also report the OR per standard deviation of continuous PRS. Area under the receiver-operating curve (AUROC) was computed using R package pROC[24]. R (v.3.3.3) was used for all statistical analyses; all statistical tests were two sided, and p-values <0.05 were considered nominally statistically significant.

Results

A total of 3,020 patient samples (1,772 BC cases and 1,248 controls) underwent NGS for the present study. After assessment of quality control and inclusion/exclusion criteria, data from 1,689 BC cases and 1,160 controls were available for analysis. The mean±SD age at testing for cases and controls was 55.7±11.3 and 47.5±12.9 years, respectively. Among cases, the mean±SD age at first diagnosis was 51.0±10.9 years. While 92.0% had at least one close relative (1st, 2nd or 3rd degree) with cancer 74.8% had a close relative and 39.7% had at least one first degree relative with breast and/or ovarian cancer, specifically. Approximately 21.8% of cases had estrogen receptor negative, and 14.0% had triple negative BC.

The mean±SD SNP call rate, or the proportion of individuals for whom a genotype was successfully determined for a given SNP, was 99.7%±1.1% (range 92.2% to 100.0%). SNP risk allele frequencies (RAF) among controls ranged from 0.8% to 93.5%, and were consistent with the 1000 Genomes non-Finnish EUR population (range: 1.0% to 93.3%; mean±SD absolute difference among SNPs: 0.5%±2.5%, p=0.05). One SNP was monomorphic in both cases and controls (RAF=0%), as observed in the 1000 Genomes non-Finnish EUR population; the Finnish population carries the risk allele with a frequency of 2.5%, and a frequency of 0.7% has been reported among

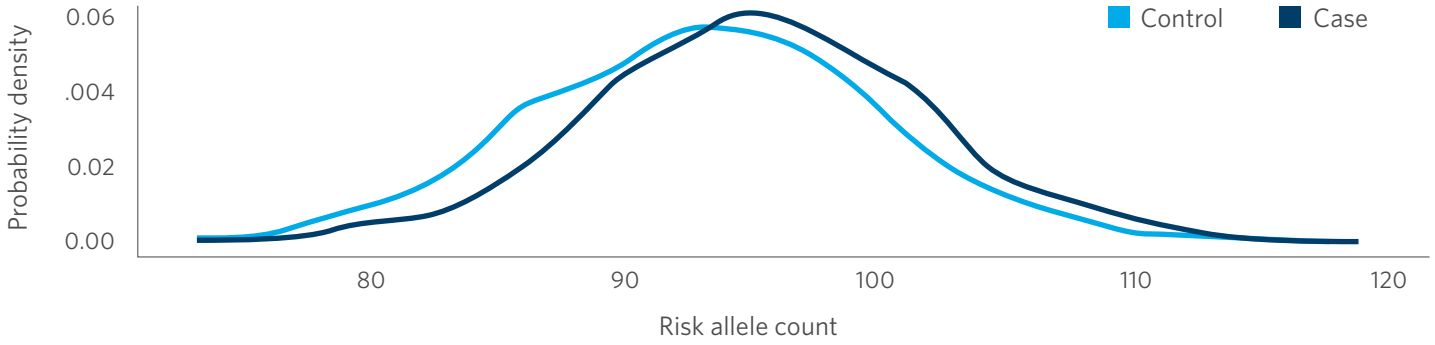

Figure 1. Distribution of the sum of risk alleles across 100 SNPs, for cases compared to controls. Probability density on the y-axis represents the proportion of cases and controls, respectively, with a given risk allele count (x-axis). The mean±SD risk allele count in cases vs. controls: 95.3±6.5 vs. 93.1±6.7, p<0.0001.

controls in the literature[2]. Consistent with the findings of previous studies[3, 5, 25], we did not detect any significant pairwise or high-order interactions among the SNPs after Bonferroni or FDR correction for multiple testing.

The sum of the risk alleles across the 100 SNPs was approximately normally distributed among cases and controls, and ranged from 75 to 119 and 73 to 111, respectively (mean±SD risk allele count: 95.3±6.5 vs. 93.1±6.7, p<0.0001; Figure 1). The mean±SD population standardized PRS was significantly higher for cases compared to controls (1.20±0.88 vs. 0.95±0.69, p<0.0001). The OR for BC per standard deviation of the PRS was 1.45 (95% CI: 1.32-1.59). Compared to women in the 1st quartile of PRS, those in the 2nd, 3rd and 4th quartile were 1.51 (95% CI: 1.23-1.87), 2.06 (95% CI: 1.67-2.55) and 2.69 (95% CI: 2.17-3.35) times as likely to have BC (all p<0.0001; Figure 2). Maximum AUROC for PRS discrimination of cases and controls was reached at a threshold of 0.83, corresponding to a PPV=0.67 and NPV=0.50 (AUROC=0.61, 95% CI: 0.59-0.63; Figure 3).

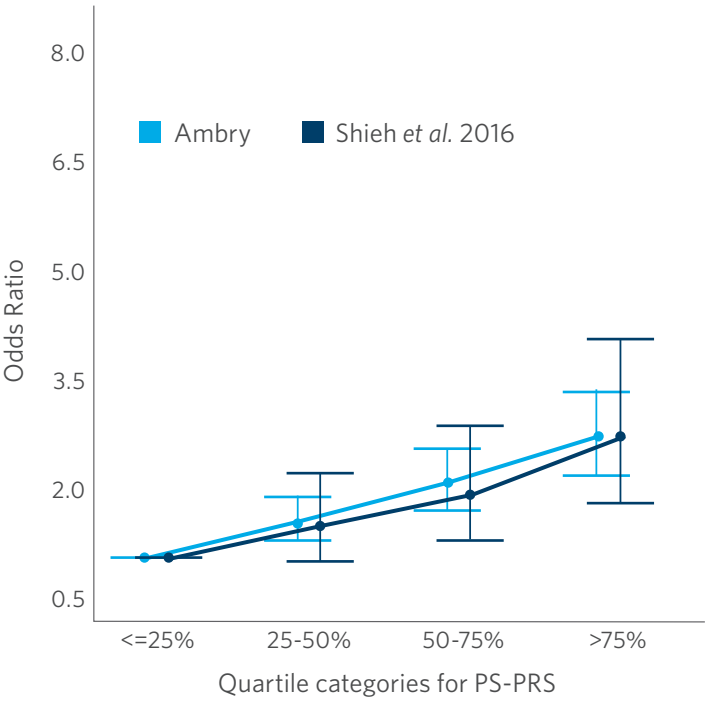

Figure 2. ORs (95% CI) for BC per quartile of PRS estimated in the Ambry case/control set compared to those reported by Shieh *et al.* 2016 [7]. First quartile of PRS is the referent category (OR=1.0).

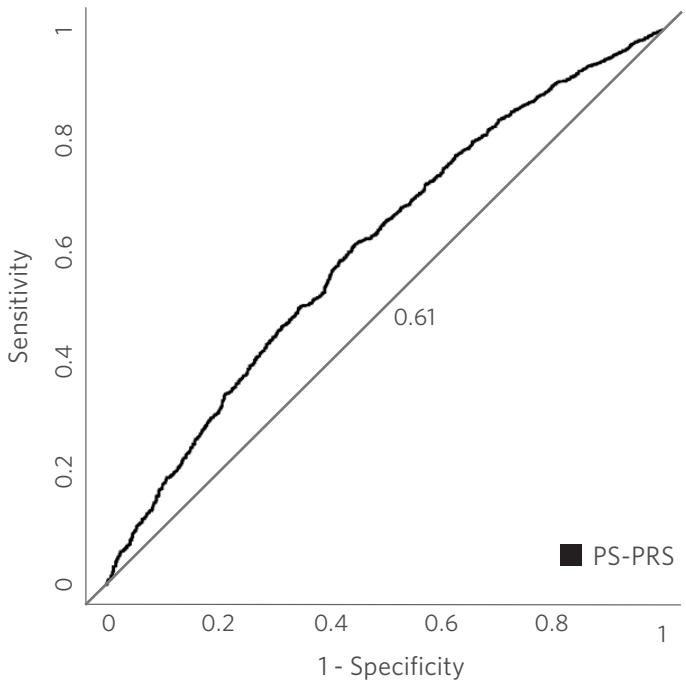

Figure 3. Area under the receiver operating curve (AUROC) for the accuracy of the PRS in distinguishing between BC cases and controls (AUROC=0.61, 95% CI: 0.59-0.63).

Discussion

We examined the performance of a PRS, based on 100 SNPs previously reported at genome-wide significance for association with BC, in a high-risk population of women referred for genetic testing who were negative for mutations in BC-susceptibility genes. Our results demonstrate that the PRS-associated BC risk for unaffected women enriched for family history of breast and/or ovarian cancer and/or other clinical characteristics is similar to that reported in broader populations. We also found that the predictive performance of the 100 SNP PRS, while modest, was consistent with prior literature.

Dite *et al.* investigated the utility of a 77-SNP population standardized PRS in 1,155 Caucasian females (750 cases and 405 controls) from the Australian Breast Cancer Family Registry who were also known to be negative for mutations in *BRCA1/2*. In this study, investigators reported an OR per standard deviation of the PRS of 1.46 (95% CI: 1.29-1.64), nearly identical to the OR per standard deviation we report for our 100-SNP PRS[4]. Shieh *et al.* examined an 83-SNP PRS in a case-control study of approximately 1,000 females (>80% Caucasian) from the California Pacific Medical Center (CPMC) Research Institute Cohort, and observed unadjusted ORs for BC of 1.34 (95% CI: 0.90 - 2.00), 1.76 (95% CI: 1.18 - 2.62) and 2.54 (95% CI: 1.69 -3.82) for the 2nd, 3rd and 4th quartile of PRS compared to the 1st quartile[7]. Interestingly, when adjusted for family history, these ORs increased to 1.45 (95% CI: 0.96-2.19), 1.89 (95% CI: 1.26 -2.85), and 2.67 (95% CI: 1.76-4.05), respectively. Our ORs for the 2nd and 3rd PRS quartile were only slightly greater, and well within the 95% CI reported by Shieh *et al.* (cf. Figure 2). Moreover, the ability of various PRSs to discriminate between BC cases and controls as assessed by AUROC in prior reports ranged 0.55-0.68[3-5, 7-9, 23, 26]; our PRS therefore has similar predictive performance to those previously published in broader populations.

In conclusion, these data suggest that a 100-SNP PRS assessed in high-risk patients performs similarly to risk scores reported in other groups of individuals, which may have implications for clinical management. Our ongoing analysis of the ability of the PRS to discriminate among specific pathologic subtypes, as well as validity and utility of a PRS combined with clinical models to estimate residual lifetime risk in both Caucasian and non-Caucasian populations, has the potential to further inform screening guidelines and improve patient care.

References

1. Easton, D.F., *et al.*, Genome-wide association study identifies novel breast cancer susceptibility loci. Nature, 2007. **447**(7148): p. 1087-93.
2. Michailidou, K., *et al.*, Association analysis identifies 65 new breast cancer risk loci. Nature, 2017. **551**(7678): p. 92-94.
3. Mavaddat, N., *et al.*, Prediction of breast cancer risk based on profiling with common genetic variants. J Natl Cancer Inst, 2015. **107**(5).
4. Dite, G.S., *et al.*, Breast Cancer Risk Prediction Using Clinical Models and 77 Independent Risk-Associated SNPs for Women Aged Under 50 Years: Australian Breast Cancer Family Registry. Cancer Epidemiol Biomarkers Prev, 2016. **25**(2): p. 359-65.
5. Mealiffe, M.E., *et al.*, Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. J Natl Cancer Inst, 2010. **102**(21): p. 1618-27.
6. Reeves, G.K., *et al.*, Incidence of breast cancer and its subtypes in relation to individual and multiple low-penetrance genetic susceptibility loci. Jama, 2010. **304**(4): p. 426-34.
7. Shieh, Y., *et al.*, Breast cancer risk prediction using a clinical risk model and polygenic risk score. Breast Cancer Res Treat, 2016. **159**(3): p. 513-25.
8. Li, H., *et al.*, Breast cancer risk prediction using a polygenic risk score in the familial setting: a prospective study from the Breast Cancer Family Registry and kConFab. Genet Med, 2017. **19**(1): p. 30-35.
9. Sawyer, S., *et al.*, A role for common genomic variants in the assessment of familial breast cancer. J Clin Oncol, 2012. **30**(35): p. 4330-6.
10. Michailidou, K., *et al.*, Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet, 2013. **45**(4): p. 353-61, 361e1-2.
11. Michailidou, K., *et al.*, Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet, 2015. **47**(4): p. 373-80.
12. Gold, B., *et al.*, Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci U S A, 2008. **105**(11): p. 4340-5.
13. Couch, F.J., *et al.*, Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. Nat Commun, 2016. **7**: p. 11375.
14. Milne, R.L., *et al.*, Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet, 2017. **49**(12): p. 1767-1778.
15. Garcia-Closas, M., *et al.*, Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nat Genet, 2013. **45**(4): p. 392-8, 398e1-2.
16. Fletcher, O., *et al.*, Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. J Natl Cancer Inst, 2011. **103**(5): p. 425-35.
17. Orr, N., *et al.*, Genome-wide association study identifies a common variant in *RAD51B* associated with male breast cancer risk. Nat Genet, 2012. **44**(11): p. 1182-4.
18. Lindstrom, S., *et al.*, Genome-wide association study identifies multiple loci associated with both mammographic density and breast cancer risk. Nat Commun, 2014. **5**: p. 5303.
19. Zerbino, D.R., *et al.*, Ensembl 2018. Nucleic Acids Res, 2018. **46**(D1): p. D754-d761.
20. Graffelman, J. and J.M. Camarena, Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. Hum Hered, 2008. **65**(2): p. 77-84.
21. Schwender, H. and K. Ickstadt, Identification of SNP interactions using logic regression. Biostatistics, 2008. **9**(1): p. 187-98.
22. Cuzick, J.,*et al.*, Impact of a Panel of 88 Single Nucleotide Polymorphisms on the Risk of Breast Cancer in High-Risk Women: Results From Two Randomized Tamoxifen Prevention Trials. J Clin Oncol, 2017. **35**(7): p. 743-750.
23. Allman, R., *et al.*, SNPs and breast cancer risk prediction for African American and Hispanic women. Breast Cancer Res Treat, 2015. **154**(3): p. 583-9.
24. Robin, X., *et al.*, pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 2011. **12**(1): p. 77.
25. Milne, R.L., *et al.*, A large-scale assessment of two-way SNP interactions in breast cancer susceptibility using 46,450 cases and 42,461 controls from the breast cancer association consortium. Hum Mol Genet, 2014. **23**(7): p. 1934-46.
26. Vachon, C.M., *et al.*, The contributions of breast density and common genetic variation to breast cancer risk. J Natl Cancer Inst, 2015. **107**(5).