Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes


Digital Object Identifier (DOI): 10.1016/j.ajhg.2018.11.002

Link: Link to publication record in Edinburgh Research Explorer

Document Version: Publisher's PDF, also known as Version of record

Published in: American Journal of Human Genetics

Publisher Rights Statement: This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

General rights Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 26. Feb. 2019
Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes

Nasim Mavaddat,1,* Kyriaki Michailidou,1,2 Joe Dennis,1 Michael Lush,1 Laura Fachal,1 Andrew Lee,1 Jonathan P. Tyrer,3 Ting-Huei Chen,4 Qin Wang,1 Manjeet K. Bolla,1 Xin Yang,1 Muriel A. Adank,5 Thomas Ahearn,6 Kristiina Aittomäki,7 Jamie Allen,1 Irene L. Andrulis,8,9 Hoda Anton-Culver,10 Natalia N. Antonenkovova,11 Volker Arndt,12 Kristan J. Aronson,13 Paul L. Auer,14,15 Päivi Auvinen,16,17,18 Myrto Barrdahl,19 Laura E. Beane Freeman,6 Matthias W. Beckmann,20 Sabine Behrens,19 Javier Benitez,21,22 Marina Bermisheva,23 Leslie Bernstein,24 Carl Blomqvist,25,26 Natalia V. Bogdanova,11,27,28 Stig E. Bojesen,29,30,31 Bernardo Bonanni,32 Anne-Lise Børresen-Dale,33,34 Hiltrud Brauch,35,36,37 Michael Bremer,27 Hermann Brenner,12,37,38 Adam Brentnall,39 Ian W. Brock,40 Angela Brooks-Wilson,41,42 Sara Y. Brucker,43 Thomas Brüning,44 Barbara Burwinkel,45,46 Daniele Campa,47 Brian D. Carter,48 Jose E. Castelao,49 Stephen J. Chanock,6 Rowan Chlebowski,50 Hans Christiansen,27 Christine L. Clarke,51 J. Margriet Colee,2 Emilie Cordin-Duverger,53 Sten Cornelissen,54 Fergus J. Couch,55 Angela Cox,40 Simon S. Cross,56 Kamila Czene,27

Stratification of women according to their risk of breast cancer based on polygenic risk scores (PRSs) could improve screening and prevention strategies. Our aim was to develop PRSs, optimized for prediction of estrogen receptor (ER)-specific disease, from the largest available genome-wide association dataset and to empirically validate the PRSs in prospective studies. The development dataset comprised 94,075 case subjects and 75,017 control subjects of European ancestry from 69 studies, divided into training and validation sets. Samples were genotyped using genome-wide arrays, and single-nucleotide polymorphisms (SNPs) were selected by stepwise regression or lasso penalized regression. The best performing PRSs were validated in an independent test set comprising 11,428 case subjects and 18,323 control subjects from 10 prospective studies and 190,040 women from UK Biobank (3,215 incident breast cancers). For the best PRSs (313 SNPs), the odds ratio for overall disease per 1 standard deviation in ten prospective studies was 1.61 (95%CI: 1.57–1.65) with area under receiver-operator curve (AUC) = 0.630 (95%CI: 0.628–0.651). The lifetime risk of overall breast cancer in the top centile of the PRSs was 32.6%. Compared with women in the middle quintile, those in the highest 1% of risk had 4.37- and 2.78-fold risks, and those in the lowest 1% of risk had 0.16- and 0.27-fold risks, of developing ER-positive and ER-negative disease, respectively. Goodness-of-fit tests indicated that this PRS was well calibrated and predicts disease risk accurately in the tails of the distribution. This PRS is a powerful and reliable predictor of breast cancer risk that may improve breast cancer prevention programs.

Introduction

Breast cancer is the most common cancer diagnosed among women in Western countries. While rare mutations in genes such as \textit{BRCA1} and \textit{BRCA2} confer high risks of developing breast cancer, these account for only a small proportion of breast cancer cases in the general population. Multiple common breast cancer susceptibility variants discovered through genome-wide association studies (GWASs)1–10 confer small risk individually, but their combined effect, when

1Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK; 2Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, 1683 Nicosia, Cyprus; 3Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge CB1 8RN, UK; 4Department of Mathematics and Statistics, Laval University, Québec City, QC G1V 0A6, Canada; 5Family Cancer Clinic, the Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, 1066 CX, the Netherlands; 6Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20850, USA; 7Department of Clinical Genetics, Helsinki University Hospital, University of Helsinki, Helsinki 00290, Finland; 8Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON MSG 1X5, Canada; 9Department of Molecular Genetics, University of Toronto, Toronto, ON MSS IAB, Canada; 10Department of Epidemiology, Genetic Epidemiology Research Institute, University of California Irvine, Irvine, CA 92617, USA; 11NN Alexandrov Research Institute of Oncology and Medical Radiology, Minsk 223040, Belarus; 12Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; 13Department of Public Health Sciences, and Cancer Research Institute, Queen’s University, Kingston, ON K7L 3N6, Canada; 14Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; 15Ziller School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53205, USA; 16Breast Cancer Research Center, Kuopio University Hospital, Kuopio 70210, Finland; 17Institute of Clinical Medicine, Oncology, University of Eastern Finland, Kuopio 70210, Finland; 18Translational Cancer Research Area, University of Eastern Finland, Kuopio 70210, Finland; 19Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; 20Department of Gynecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen 91054, Germany; 21Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain; 22Biomedical Network on Rare Diseases (CIBERER), Madrid 28029, Spain; 23Institute of Biochemistry and Genetics, Ufa Scientific Center of Russian Academy of Sciences, Ufa 450054, Russia; 24Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA 91010, USA; 25Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki 00290, Finland; 26Department of Oncology, Ourebo University Hospital, Ourebo 70185, Sweden; 27Department of Radiation Oncology, Hannover Medical School, Hannover, Germany; 28Breast Cancer Subtypes

© 2018 The Authors. This is an open access article under the CC-BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
summarized as a polygenic risk score (PRS), can be substantial.3–5 Such genomic profiles can be used to stratify women according to their risk of developing breast cancer.6 This in turn holds the promise of improved breast cancer prevention and survival, by targeting screening or other preventative strategies at those women most likely to benefit. We previously derived a PRS based on 77 established breast cancer susceptibility single-nucleotide polymorphisms (SNPs) and reported levels of risk stratification achieved by this PRS.7 Based on our findings, several studies have investigated the potential for combining PRSs and other known risk factors for risk stratification and evaluated the impact of risk reduction strategies across risk strata defined by the PRS.8–10 Preliminary studies investigating the use of the PRS to inform targeted breast cancer screening programs are underway (see CORDIS...
and GenomeCanada in Web Resources). Empirical validation and characterization of the PRS in large-scale epidemiological studies has, however, not been carried out previously. In addition, more informative PRSs would improve the clinical utility of risk prediction. GWASs have now identified ~170 breast cancer susceptibility loci. Moreover, genome-wide heritability estimates indicate that these loci explain only ~40% of the heritability explained by all common variants on genome-wide SNP arrays. This suggests that the discrimination provided by the PRS could be improved by incorporating variants associated at more liberal significance thresholds. In addition, many variants confer risks that differ by breast cancer subtype (estrogen-receptor [ER]-positive or -negative), suggesting that subtype-specific PRSs might allow better prediction of subtype-specific disease, including the more aggressive ER-negative breast cancer, and enable selection of women for preventative medication.
Here, we used data from 79 studies conducted by the Breast Cancer Association Consortium (BCAC) to optimize PRSs for overall and subtype-specific disease, and we validate their performance in independent datasets.\textsuperscript{1,13–15}

**Material and Methods**

**Study Subjects and Genotyping**

The dataset used for development of the PRSs comprised 94,075 breast cancer-affected case subjects and 75,017 control subjects of European ancestry from 69 studies in the BCAC (Tables S1 and S2). Data collection for individual studies is described previously.\textsuperscript{1} Samples were genotyped using one of two arrays: iCOGS\textsuperscript{13,14}, and OncoArray.\textsuperscript{1,15} The dataset was divided into a training and validation set. The validation set was randomly selected (approximately 10% of case and control subjects) from studies that had been genotyped with the OncoArray, after excluding studies of bilateral breast cancer, studies or sub-studies oversampling for family history, and individuals with in situ cancers or case subjects with unknown ER status.

The best PRSs were evaluated in an independent test dataset comprising 11,428 invasive breast cancer-affected case subjects and 75,017 control subjects of European ancestry from 69 studies in the BCAC (Tables S1 and S2). Data collection for individual studies is described previously.\textsuperscript{1} Samples were genotyped using one of two arrays: iCOGS\textsuperscript{13,14}, and OncoArray.\textsuperscript{1,15} The dataset was divided into a training and validation set. The validation set was randomly selected (approximately 10% of case and control subjects) from studies that had been genotyped with the OncoArray, after excluding studies of bilateral breast cancer, studies or sub-studies oversampling for family history, and individuals with in situ cancers or case subjects with unknown ER status.

The best PRSs were evaluated in an independent test dataset comprising 11,428 invasive breast cancer-affected case subjects and 75,017 control subjects of European ancestry from 69 studies in the BCAC (Tables S1 and S2). Data collection for individual studies is described previously.\textsuperscript{1} Samples were genotyped using one of two arrays: iCOGS\textsuperscript{13,14}, and OncoArray.\textsuperscript{1,15} The dataset was divided into a training and validation set. The validation set was randomly selected (approximately 10% of case and control subjects) from studies that had been genotyped with the OncoArray, after excluding studies of bilateral breast cancer, studies or sub-studies oversampling for family history, and individuals with in situ cancers or case subjects with unknown ER status.

The best PRSs were evaluated in an independent test dataset comprising 11,428 invasive breast cancer-affected case subjects and 75,017 control subjects of European ancestry from 69 studies in the BCAC (Tables S1 and S2). Data collection for individual studies is described previously.\textsuperscript{1} Samples were genotyped using one of two arrays: iCOGS\textsuperscript{13,14}, and OncoArray.\textsuperscript{1,15} The dataset was divided into a training and validation set. The validation set was randomly selected (approximately 10% of case and control subjects) from studies that had been genotyped with the OncoArray, after excluding studies of bilateral breast cancer, studies or sub-studies oversampling for family history, and individuals with in situ cancers or case subjects with unknown ER status.

The best PRSs were evaluated in an independent test dataset comprising 11,428 invasive breast cancer-affected case subjects and 75,017 control subjects of European ancestry from 69 studies in the BCAC (Tables S1 and S2). Data collection for individual studies is described previously.\textsuperscript{1} Samples were genotyped using one of two arrays: iCOGS\textsuperscript{13,14}, and OncoArray.\textsuperscript{1,15} The dataset was divided into a training and validation set. The validation set was randomly selected (approximately 10% of case and control subjects) from studies that had been genotyped with the OncoArray, after excluding studies of bilateral breast cancer, studies or sub-studies oversampling for family history, and individuals with in situ cancers or case subjects with unknown ER status.

The best PRSs were evaluated in an independent test dataset comprising 11,428 invasive breast cancer-affected case subjects and 75,017 control subjects of European ancestry from 69 studies in the BCAC (Tables S1 and S2). Data collection for individual studies is described previously.\textsuperscript{1} Samples were genotyped using one of two arrays: iCOGS\textsuperscript{13,14}, and OncoArray.\textsuperscript{1,15} The dataset was divided into a training and validation set. The validation set was randomly selected (approximately 10% of case and control subjects) from studies that had been genotyped with the OncoArray, after excluding studies of bilateral breast cancer, studies or sub-studies oversampling for family history, and individuals with in situ cancers or case subjects with unknown ER status.
The overall breast cancer PRS was also evaluated among 190,040 women of European ancestry from the UK Biobank cohort who had not had any cancer diagnosis or mastectomy prior to recruitment. A total of 3,215 incident registry-confirmed invasive breast cancers developed over 1,381,019 person years of prospective follow-up. Follow-up started 6 months after age of baseline questionnaire. The primary endpoint was invasive breast cancer. Follow-up was censored at the earliest of: risk-reducing mastectomy, diagnosis of any type of cancer, death, or January 15, 2017.

Genotype calling, quality control, and imputation for iCOGS and OncoArray were performed as previously described. Briefly, imputation was performed for the iCOGS and OncoArray datasets separately using the Phase 3 (October 2014) release of the 1000 Genomes data as reference. We followed a two-stage approach using SHAPEIT for phasing and IMPUTE2 for the imputation. Where samples were genotyped with iCOGS and OncoArray, the OncoArray calling was used. SNPs with MAF > 0.01 and imputation $r^2 > 0.9$ for OncoArray and $r^2 > 0.3$ for iCOGS were included in this analysis (~7 million SNPs); a higher threshold was imposed for OncoArray to ensure accurate determination of the PRS in the validation and test datasets.

UK Biobank samples were genotyped using Affymetrix UK BiLEVE Axiom array and Affymetrix UK Biobank Axiom array and imputed to the combined 1000 Genomes Project v3 and UK10K reference panels using SHAPEIT3 and IMPUTE3. The lowest imputation info score for the SNPs used in these analyses was 0.86. Samples were included on the basis of female sex (genetic and self-reported) and ethnicity filter (Europeans/White British ancestry subset). Duplicates, individuals with high degree of relatedness (>10 relatives), and one of each related pair of first degree relatives were removed. Samples were also excluded using standard quality control criteria.

Participants provided written informed consent, all studies were approved by the relevant ethics committees, and procedures followed were in accordance with the ethical standards of these committees.

**Statistical Analysis**

The general aim was to derive a PRS of the form:

$$
PRS = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \beta_0
$$

where $\beta_k$ is the per-allele log odds ratio (OR) for breast cancer associated with SNP $k$, $x_k$ is the allele dosage for SNP $k$, and $n$ is the total number of SNPs included in the PRS. Previous analyses found no evidence for statistically significant interactions between SNPs and little evidence for departures from a log-additive model for individual SNPs. Assuming this is true in general, the PRS summarizes efficiently the combined effects of SNPs on disease risk.

The main challenge is how to determine which SNPs to include and the weighting parameters $\beta_k$ to assign. Inclusion of only those SNPs reaching a stringent significance threshold ("genome-wide significant," $p < 5 \times 10^{-8}$) threshold ignores information from larger numbers of SNPs that are likely, but not certain, to be associated with the risk of breast cancer. We used two general
approaches for model selection: “hard-thresholding,” based on a stepwise regression model that retained SNPs significantly associated with overall or subtype-specific disease at a given threshold, and penalized regression using lasso.\textsuperscript{21,22} A schema for the analyses is shown in Figure S1.

To prioritize SNPs for analysis, single SNP association tests were first conducted in the training set. Per-allele ORs and standard errors were estimated separately in the iCOGS and OncoArray datasets, adjusting for study and nine ancestry informative principal components (PCs) in the iCOGS dataset and by country and ten PCs in the OncoArray dataset, using a purpose-written program.\textsuperscript{1} Combined p values were then derived using a fixed-effects meta-analysis with the software METAL.\textsuperscript{21} SNPs were sorted by p value and filtered on LD, such that uncorrelated SNPs (correlation \(r^2 < 0.9\)) with lowest p value for association with overall breast cancer in the training set were retained (more rigorous pruning, for example at \(r^2 < 0.2\), would have removed from consideration informative SNPs from regions with multiple correlated signals\textsuperscript{23,24}).

In the hard thresholding approach, a series of stepwise forward regression analyses were first carried out in 1 Mb regions centered on SNPs significant at a pre-specified threshold for association with either overall and/or subtype-specific disease in the training set. Only SNPs passing the specified p value thresholds were included in each 1 Mb region. Two analyses were performed in parallel: for overall breast cancer and ER-negative disease. At each stage the SNP with the smallest (conditional) p value for any analysis was added to the model, the threshold for the stepwise regression being the same as that for pre-selection. The process was repeated until no further SNPs could be added at the pre-defined threshold. A second stage of stepwise regressions were then carried out across all regions in each chromosome, to take into account correlated SNPs in different regions. Finally, the effect sizes for the selected SNPs were jointly estimated in a single logistic regression model.

For the best-performing PRSs, SNPs associated with ER-positive at \(p < 10^{-6}\) but not with overall breast cancer (at \(p < 10^{-5}\)) were added at the end of the final SNP list. A third round of stepwise forward regression was then carried out with p value for selection of \(p < 10^{-7}\) for ER-positive disease. For completeness we added to this final PRS two rarer variants (\(BRC\)A2 p.Lys3326X and CHEK2 p.Ile157Tyr) which are established to confer a moderate risk of breast cancer and were genotyped on the OncoArray but did not pass the allele frequency threshold in the PRS development phase.

For the penalized regression using lasso, we used the program \textit{glmnet}.\textsuperscript{22} SNPs with \(p < 0.001\) in overall BC or ER-negative disease in the training set were pre-selected for inclusion in the lasso, and \(BRC\)A2 p.Lys3326X and CHEK2 p.Ile157Thr were added. Covariates for 19 PCs (9 for iCOGS and 10 for Oncoarray) and country were included in each model. For overall breast cancer, the penalty parameter (lambda) giving the best overall breast cancer PRS in the validation set was selected.

To construct subtype-specific PRSs, we evaluated four different methods: (1) using effect sizes for overall breast cancer (for each of the subtypes), (2) using effect sizes for subtype-specific (ER-positive or ER-negative) disease, (3) using a hybrid method, in which effect sizes were estimated in the relevant subtype for SNPs passing a certain optimal significance threshold in a case-only logistic regression (ER-positive versus ER-negative disease), and otherwise, using effect sizes estimated for overall breast cancer, or (4) by estimating case-only ORs using lasso and combining these with the overall breast cancer ORs to derive subtype-specific estimates, using the formulae:

\[
\beta_{\text{ERpositive}} = \beta_{\text{overall}} + \eta \cdot \beta_{\text{case-only}}
\]

\[
\beta_{\text{ERnegative}} = \beta_{\text{overall}} - (1 - \eta) \cdot \beta_{\text{case-only}}
\]

where \(\eta = 0.27\) was the proportion of ER-negative tumors in the validation set.

For the lasso analysis, effect sizes for subtype-specific disease were estimated using method 4 above, combining the estimates from a case-only lasso analysis with the coefficients for overall breast cancer from the lasso analysis. The lambda for the case-only model giving the best subtype-specific PRS in the validation set was selected.

To evaluate the performance of each potential PRS, we standardized the PRSs to have unit standard deviation (SD) in the validation set of control subjects. The association of the standardized PRSs was evaluated in the validation and test (prospective studies) datasets, by logistic regression. We used a Cox proportional hazards regression model to assess the association with risk of breast cancer in UK Biobank. Models were also compared in terms of the area under the receiver operator characteristic curves (AUC), adjusted for study, calculated using the Stata command \textit{comprc}. Meta-analysis of study-specific effects was carried out using the Stata command \textit{metan}.

The goodness of fit of the continuous model (i.e., assuming a linear association between log(OR) and risk) was tested using the Hosmer-Lemeshow (HL) test to compare the observed and predicted risks by quantile and using the tail-based test proposed by Song et al.\textsuperscript{26} In addition, we considered specifically the risks in the highest and lowest 1% of the distribution.

Effect modification of the PRS by age and family history of breast cancer in first-degree relatives was evaluated by fitting additional interaction terms in the model. The validation and prospective test datasets were combined for this analysis.

The absolute risks of developing breast cancer (overall and subtype-specific disease) were calculated taking into account the competing risk of dying from causes other than breast cancer, as described previously,\textsuperscript{7} with the PRS modeled as a continuous covariate and including a linear “age × PRS” interaction term. The absolute risk of developing subtype-specific disease was obtained constraining to the incidence of overall incidence of ER-negative and ER-positive disease in the UK. Women are at risk of developing both ER-negative and ER-positive disease, so the absolute risks were calculated given that the individual has been free of breast cancer of any subtype.

Analyses were carried out in R v.3.0.2 and Stata v.14.2. All tests of statistical significance were two-sided. Further details are provided in the Supplemental Material and Methods.

Results

Development of the PRS

We tried several approaches to develop PRSs; here we report results for models giving the highest prediction accuracy. Using stepwise forward selection, the best PRS for prediction of overall breast cancer was obtained at a p value threshold for pre-selection and stepwise regression of \(p < 10^{-5}\) (Table 1). The OR per unit standard deviation (SD) for this 305-SNP PRS with overall breast cancer in
the validation set was 1.65 (95% CI: 1.58–1.72), compared with 1.59 (95% CI: 1.52–1.66) using a “genome-wide” (p < 5 × 10^{-8}) threshold (123 SNPs).

Using lasso regression, the best PRS (OR = 1.71, 95% CI: 1.64–1.79) was more predictive than the best PRS developed using the stepwise regression model. In the best model (λ = 0.003), 3,820 SNPs were selected (Table 1).

**Optimizing the PRS for Prediction of Subtype-Specific Disease**

For evaluation of subtype-specific models following stepwise regression, SNP effect sizes were estimated, in the first instance, in each disease subtype. The best subtype-specific PRSs using this method were also obtained at a p value threshold of p < 10^{-5} (Table S5). The 305-SNP PRS was supplemented with 6 additional SNPs associated with ER-positive at p value < 10^{-6} and, in addition, by two known rare breast cancer susceptibility variants in the BRCA2 and CHEK2 genes, bringing the total number of SNPs included to 313 (PRS313).

The optimum subtype-specific PRS was obtained when a subset of these 313 SNPs (196 SNPs with a case-only p value for association with ER-negative versus ER-positive disease of p < 0.025) were given subtype-specific weights, while the remaining SNPs were given overall breast cancer weights. For ER-negative disease, the OR improved from OR = 1.45 (95% CI: 1.35–1.56) to OR = 1.47 (95% CI: 1.37–1.58) using the hybrid method compared with using only subtype-specific estimates, while for ER-positive disease the results were similar (OR = 1.74) (Tables S6 and S7).

Subtype-specific prediction using the lasso analysis was optimized using case-only lasso analysis. The OR per 1 SD in the validation set was 1.81 (95% CI: 1.73–1.89) for ER-positive and 1.48 (95% CI: 1.37–1.59) for ER-negative disease (Tables 2 and S8).

**Validation of the PRS in the Prospective Test Dataset**

The final PRSs were evaluated using data from 11,428 invasive breast cancer-affected case subjects and 18,323 control subjects from ten prospective studies. The ORs for both the overall and subtype-specific PRSs were slightly lower in the prospective test set compared to the validation set (Table 2). The difference between validation and test set may reflect some overfitting due to choosing the optimum p value threshold and for the lasso, the optimum lambda, in the validation set, but could also be due to somewhat different characteristics of the prospective studies. The ORs for overall and ER-positive, but not ER-negative, breast cancer were slightly higher for the 3,820-SNP PRS (PRS3820) compared with PRS313.

The odds ratio (OR) for overall disease per 1 standard deviation (SD) of the PRS313 in the prospective studies was 1.61 (95% CI: 1.57–1.65) while for the 77-SNP PRS (PRS77) derived previously OR = 1.46 (95% CI: 1.42–1.49). For ER-negative disease the difference was OR = 1.45 (95% CI: 1.37–1.53) versus 1.35 (95% CI: 1.27–1.43) (Table 2).

The associations between the PRS and overall, ER-positive, and ER-negative breast cancer by percentiles of the PRS313 are shown in Figure 1 and Table S9. Compared with women in the middle quintile (40th to 60th percentile), those in the highest 1% of risk for the subtype-specific PRS313 had 4.37 (95% CI: 3.59–5.33) and 2.78 (95% CI: 1.83–4.24)-fold risks, and those in the lowest 1% had 0.16 (95% CI: 0.09–0.30)- and 0.27 (95% CI: 0.09–0.86)-fold risks of developing ER-positive and ER-negative disease, respectively. The ORs by percentile of the PRS3820 were similar (Table S10).

**Goodness of Fit of the PRS**

The remaining analyses concentrated on PRS313. The associations between the PRS and breast cancer risk by

---

**Table 1. Comparison of Methods for Deriving the PRS: Results for Overall Breast Cancer in the Validation Set**

<table>
<thead>
<tr>
<th>p Value Cutoffa</th>
<th>SNPs Entering Model (n)</th>
<th>SNPs Selected (n)</th>
<th>ORb</th>
<th>95% CI</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published PRS7</td>
<td>77</td>
<td>77</td>
<td>1.49</td>
<td>1.44–1.56</td>
<td>0.612</td>
</tr>
<tr>
<td>Hard-Thresholding Stepwise Forward Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;5 × 10^{-8}</td>
<td>1,817</td>
<td>123</td>
<td>1.59</td>
<td>1.52–1.66</td>
<td>0.626</td>
</tr>
<tr>
<td>&lt;10^{-6}</td>
<td>2,603</td>
<td>197</td>
<td>1.62</td>
<td>1.55–1.68</td>
<td>0.634</td>
</tr>
<tr>
<td>&lt;10^{-5}</td>
<td>3,818</td>
<td>305</td>
<td>1.65</td>
<td>1.58–1.72</td>
<td>0.637</td>
</tr>
<tr>
<td>&lt;10^{-4}</td>
<td>6,743</td>
<td>669</td>
<td>1.62</td>
<td>1.56–1.69</td>
<td>0.631</td>
</tr>
<tr>
<td>&lt;10^{-3}</td>
<td>14,760</td>
<td>1,707</td>
<td>1.55</td>
<td>1.49–1.62</td>
<td>0.623</td>
</tr>
<tr>
<td>Penalized Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>15,032</td>
<td>3,820</td>
<td>1.71</td>
<td>1.64–1.79</td>
<td>0.647</td>
</tr>
</tbody>
</table>

---

*The p value cut off refers to the SNPs considered based on their marginal associations in the training set; the same p value threshold was used in each case in the stepwise regression. Parameter selection and effect size estimation for derivation of the PRS was carried out in the training set as described in the Material and Methods.

*OR per 1 SD for the PRS. OR for association with breast cancer in the validation set was derived using logistic regression adjusting for country and ten PCs. AUCs were adjusted for country. The lasso was carried out after pre-selecting SNPs at p < 10^{-5} based on their marginal association in the training set. For the lasso λ = 0.003 gave the optimal PRS in the validation set.*
percentiles of the risk score were compared with those predicted under a simple polygenic model with the PRS considered as a continuous covariate. The effect sizes did not differ from those predicted, and in particular the estimates for the highest and lowest centile were consistent with the predicted estimates (Table S9).

Further tests for goodness of fit and tail-based tests (see Material and Methods) were not statistically significant at $p < 0.05$.

There was no evidence of heterogeneity in the effect sizes among studies (Figure 2). All studies showed a significant association with similar effect sizes for overall and ER-positive breast cancer, and all but one study (FHRISK, based on only six case subjects) showed a significant effect for ER-negative breast cancer.

In the UK Biobank, the estimated hazard ratio (HR) for overall breast cancer per unit PRS (including 306 of the 313 SNPs) was $HR = 1.59$ (95%CI: 1.54–1.64) (Figure 2).

By way of comparison, we also evaluated a PRS based on 177 previously published susceptibility loci. The effect size for this PRS (OR = 1.61, 95%CI: 1.57–1.65) in the ten prospective studies was similar to the PRS313. However, this estimated effect size is biased because the validation and test datasets used here contributed to the GWAS discovery datasets; in the UK Biobank this PRS (based on 174 of 177 available SNPs) performed worse (HR = 1.53, 95%CI: 1.48–1.58).

### Combined Effects of PRS and Breast Cancer Family History

The association between PRS and disease risk was observed for women with and without a family history (Table 3). However, there was some evidence that for ER-positive disease, the PRS OR was smaller in women with a family history (interaction OR = 0.91, $p = 0.004$). The log OR for family history was attenuated by 21% (1.59 to 1.44) and 12% (1.66 to 1.56) for ER-positive and ER-negative disease, respectively, after adjusting for the PRS (Tables 3 and S12).

### Absolute Risk of Developing Breast Cancer According to the PRS

Estimated lifetime and 10-year absolute risks for UK women in percentiles of the PRS are shown in Figure 3. For ER-positive disease, the estimated lifetime absolute risk by age 80 years ranged from 2% for women in the lowest centile to 31% in the highest centile, while for ER-negative disease, the absolute risks ranged from 0.55% to 4%. The average 10-year absolute risk of breast cancer for a 47-year-old woman (i.e., the age at which women become eligible to enter the UK breast cancer screening program) in the general population is 2.6%. However, the 19% of women with the highest PRSs will attain this level of risk by age 40 years.
Discussion

We report development and independent validation of polygenic risk scores for breast cancer, optimized for prediction of subtype-specific disease and based on the largest available GWAS dataset. The best PRS based on a hard thresholding approach included 313 SNPs and was significantly more predictive of risk than the previously reported 77-SNP PRS (OR per 1 SD in the prospective test set: 1.61 versus 1.46; Table 2). The effect sizes were remarkably

![Figure 1. Association between the 313 SNP Polygenic Risk Score and Breast Cancer Risk](image)

Association between the 313 SNP polygenic risk score (PRS) and breast cancer risk in women of European origin for (A) overall breast cancers, (B) estrogen receptor (ER)-positive disease, and (C) ER-negative disease, in the validation (dashed line) and test (solid line) sets. Odds ratios are for different quantiles of the PRS relative to the mean PRS. Odds ratios and 95% confidence intervals are shown.

![Figure 2. Prospective Validation for the 313 SNP Polygenic Risk Score](image)

Prospective validation for the 313 SNP polygenic risk score (PRS) by study for (A) overall breast cancer, (B) ER-positive disease, and (C) ER-negative disease. Association between the 313 SNP PRS and breast cancer risk in women of European origin. Odds ratios and 95% confidence intervals are shown. I-squared and p value for heterogeneity were calculated using fixed effect meta-analysis.
consistent among the 10 cohorts in the prospective test set, and also consistent with that in the UK Biobank cohort (HR = 1.59, 95%CI: 1.54–1.64).

Recently, Khera et al.\textsuperscript{27} derived a PRS using our publicly available summary statistics based on analysis of the BCAC data.\textsuperscript{1} We were able to construct a PRS based on 5,194 of their 5,218 listed SNPs and compared this to our 313-SNP PRS. In our analysis of this PRS in the prospective UK Biobank data, we obtained a HR of 1.49 (95%CI: 1.44–1.54), substantially lower than that for our PRS\textsubscript{313}. The corresponding AUCs were 0.613 (95%CI: 0.603–0.623) for their 5,194-SNP PRS versus AUC 0.630 (95%CI: 0.620–0.640) for our PRS\textsubscript{313}. Similarly, PRS\textsubscript{313} performed better than the Khera et al. PRS in a Biobank dataset consisting of 7,113 case subjects diagnosed before entry and 183,536 control subjects (AUC = 0.642 versus AUC = 0.627). Khera et al. report a much higher AUC (0.68), perhaps reflecting the inclusion of predictors other than SNPs in their model (for example age or principal components).

We specifically aimed to improve prediction for ER-negative breast cancer as to date prediction of this more aggressive disease has been poor. SNP selection was based on association with either ER-negative or overall breast cancer, and the optimum subtype-specific PRSs were derived by weighting a subset of SNPs according to subtype-specific effect sizes, with overall breast cancer weights used for the remaining SNPs. These results are consistent with the observation from genome-wide analyses that the heritability of ER-positive and ER-negative disease are partially correlated.\textsuperscript{2} The performance of the PRS\textsubscript{313} in predicting ER-negative disease was considerably improved over the PRS\textsubscript{57} reported previously (OR = 1.45 versus 1.35). Nevertheless, the prediction is still better for ER-positive than ER-negative disease, reflecting the fact that ER-negative disease is more infrequent and hence the GWAS data are less powerful. The estimated heritability of ER-negative disease is similar to that of overall breast cancer,\textsuperscript{1,2} suggesting that more powerful ER-negative PRSs should be achievable with larger sample sizes.

The best PRS developed using lasso was more predictive for ER-positive disease but slightly less predictive for ER-negative disease in the prospective studies. Given the small differences between the models, we focused on PRS\textsubscript{313} since this should be more straightforward to implement in diagnostic laboratories using next generation sequencing. However, this will change with developing technology, and the cost effectiveness of using a large marker panel should be further investigated.

From a clinical viewpoint, an important consideration is the performance of the PRS in the tails of the distribution. According to the standard polygenic model, under which the effects of variants combine multiplicatively, the relationship between the PRS and the log-OR should be linear. The PRS was well calibrated at different quantiles. Even in this large study, we observed no deviation from this model, and in particular the observed risks in the highest and lowest centile were consistent with the predicted risk. The sample sizes in the extreme tails, however, were still relatively small, particularly for ER-negative disease.

While the AUC may appear modest, the predicted risk differences in the tails of the distribution are large. For the new PRS\textsubscript{313}, the women in the top 1% of the distribution have a predicted risk that is approximately 4-fold larger than the risk in the middle quintile. The lifetime risk of overall breast cancer in the top centile of the PRSs, based on UK incidence and mortality data, was 32.6%. Women in the top centile would therefore meet the UK NICE definition of high risk (see Web Resources). In the general population, an estimated 3.6%, 12%, 21%, and 35% of all breast cancers would be expected to occur in women in the highest 1%, 5%, 10%, and 20% of the new

\textbf{Table 3.} Associations between the 313-SNP PRS (PRS\textsubscript{313}) and Breast Cancer Risk by First-Degree Family History of Breast Cancer in the Combined Validation and Prospective Test Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>ER-Positive Disease</th>
<th></th>
<th>ER-Negative Disease</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR\textsuperscript{a}</td>
<td>95% CI</td>
<td>OR\textsuperscript{a}</td>
<td>95% CI</td>
</tr>
<tr>
<td>PRS unadjusted</td>
<td>1.67</td>
<td>1.62–1.72</td>
<td>1.44</td>
<td>1.37–1.54</td>
</tr>
<tr>
<td>PRS in women without family history</td>
<td>1.71</td>
<td>1.65–1.78</td>
<td>1.45</td>
<td>1.36–1.57</td>
</tr>
<tr>
<td>PRS in women with family history</td>
<td>1.55</td>
<td>1.48–1.65</td>
<td>1.40</td>
<td>1.27–1.55</td>
</tr>
<tr>
<td>Interaction between PRS and family history</td>
<td>0.91</td>
<td>0.85–0.97 (p = 0.004)</td>
<td>0.96</td>
<td>0.85–1.09 (p = 0.53)</td>
</tr>
</tbody>
</table>

Association between Family History and Breast Cancer Risk (Adjusted and Unadjusted for PRS)

<table>
<thead>
<tr>
<th>Family history</th>
<th>OR\textsuperscript{a}</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family history unadjusted for PRS</td>
<td>1.59</td>
<td>1.46–1.72</td>
</tr>
<tr>
<td>Family history adjusted for PRS</td>
<td>1.44</td>
<td>1.33–1.57</td>
</tr>
</tbody>
</table>

Association with breast cancer risk was tested for using logistic regression adjusting for study and ten PCs. For these analyses the validation and test datasets were combined. Analyses were restricted to women with known age and family history information. For ER-negative disease, 4,440 women with and 13,132 women without a family history of breast cancer were included in these analyses. For ER-positive disease, 6,787 women with and 17,351 women without a family history of breast cancer were included in these analyses. For ER-negative disease, reflecting the fact that ER-negative disease is more infrequent and hence the GWAS data are correlated.\textsuperscript{2} The performance of the PRS\textsubscript{313} in predicting ER-negative disease is more infrequent and hence the GWAS data are less powerful. The estimated heritability of ER-negative disease is similar to that of overall breast cancer,\textsuperscript{1,2} suggesting that more powerful ER-negative PRSs should be achievable with larger sample sizes.

The best PRS developed using lasso was more predictive for ER-positive disease but slightly less predictive for ER-negative disease in the prospective studies. Given the small differences between the models, we focused on PRS\textsubscript{313} since this should be more straightforward to implement in diagnostic laboratories using next generation sequencing. However, this will change with developing technology, and the cost effectiveness of using a large marker panel should be further investigated.

From a clinical viewpoint, an important consideration is the performance of the PRS in the tails of the distribution. According to the standard polygenic model, under which the effects of variants combine multiplicatively, the relationship between the PRS and the log-OR should be linear. The PRS was well calibrated at different quantiles. Even in this large study, we observed no deviation from this model, and in particular the observed risks in the highest and lowest centile were consistent with the predicted risk. The sample sizes in the extreme tails, however, were still relatively small, particularly for ER-negative disease.

While the AUC may appear modest, the predicted risk differences in the tails of the distribution are large. For the new PRS\textsubscript{313}, the women in the top 1% of the distribution have a predicted risk that is approximately 4-fold larger than the risk in the middle quintile. The lifetime risk of overall breast cancer in the top centile of the PRSs, based on UK incidence and mortality data, was 32.6%. Women in the top centile would therefore meet the UK NICE definition of high risk (see Web Resources). In the general population, an estimated 3.6%, 12%, 21%, and 35% of all breast cancers would be expected to occur in women in the highest 1%, 5%, 10%, and 20% of the new
PRS313, respectively, compared to only 9% of breast cancers in women in the lowest 20% of the distribution.

We observed a decline in the relative risk with age for ER-positive disease but not ER-negative disease. Even for ER-positive disease, however, the predicted relative risk, under a linear model, only declined from 1.89 at age 40 to 1.67 at age 70. While there was some indication of a lower relative risk below age 40 (estimated as 1.63 in the test set; Figure S2), these results indicate that PRS313 is broadly applicable at all ages. We observed an attenuation of the association between breast cancer family history and breast cancer risk after adjustment for the PRS (~21% for ER-positive, ~12% for ER-negative disease). This finding is broadly in line with the predicted contribution of the PRS to the familial relative risk of breast cancer. The PRS was predictive in women with and without a family history of breast cancer, but the OR was slightly lower in women with a family history, at least for ER-positive disease. This might reflect a weaker relative effect of the PRS in carriers of BRCA1 or BRCA2 mutations.28 We note, however, that the absolute differences in risk by PRS will be larger in women with a family history. These results indicate that the joint effects of family history and PRS need to be considered in risk prediction.

Figure 3. Cumulative and 10-Year Absolute Risk of Developing Breast Cancer
Cumulative and 10-year absolute risk of developing breast cancer for (A) overall breast cancer, (B) ER-positive disease, and (C) ER-negative disease by percentiles of the 313 SNP polygenic risk scores (PRSs). Note different scales and PRS categories in the different panels. The red line shows the 2.6% risk threshold corresponding to the mean risk for women aged 47 years. Absolute risks were calculated based on UK incidence and mortality data and using the PRS relative risks estimated as described in the Material and Methods.
Although we used the largest training dataset available to date for development of the PRS, further improvement should still be possible. We previously estimated using GWAS data that the theoretically best PRS, if the effect sizes of all common SNPs were known with certainty, would explain ~41% of the familial risk of breast cancer, corresponding to a standardized OR ~2.1: the PRS explains ~45% of this “chip” heritability. This implies that larger features into the analysis did not improve the predictive value, presumably because the enrichment effect was too small to overcome the increased complexity of the model. Better definition of genomic features to predict causal variants, and more sophisticated methods for integrating external biological information into prediction models, may improve the PRS.

The PRS has the potential to improve stratification for screening, while ER-specific PRSs may be informative for prevention with endocrine therapies. Previous studies have suggested that the earlier PRS was more predictive for screen-detected breast cancers than interval cancers, and that breast cancers arising among women with a low PRS are more aggressive compared with those arising in women with a high PRS, perhaps reflecting the stronger associations with ER-positive disease. It will therefore be important to evaluate carefully the associations between the new PRS and other tumor characteristics. Clinical translational studies are required to assess the risks and benefits of including the PRS in the context of current screening protocols.

While the PRS provides powerful risk discrimination, better risk discrimination will be obtained by combining the PRS with family history and other risk factors. This can be accomplished by incorporating the PRS into risk prediction models, in particular BOADICEA, which can allow for the explicit effects of family history, age, genetic, and other risk factors. (See Supplemental Material and Methods). However, further studies to validate risk models for individualized risk prediction based on the combined effects of genetic and lifestyle risk factors will be needed. In addition, it is important to note that the PRSs generated in this study were developed and validated in white European populations and need to be validated and potentially adapted for other populations.

Accession Numbers
Requests for access to this dataset should be made to the BCAC coordinator, contact provided in Web Resources.

Supplemental Data
Supplemental Data include 2 figures, 12 tables, Supplemental Acknowledgments, and Supplemental Material and Methods and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.11.002.

Consortia
ABCTB Investigators are Christine Clarke, Rosemary Balleine, Robert Baxter, Stephen Bray, Jane Carpenter, Jane Dahlstrom, John Forbes, C. S. Lee, Deborah Marsh, Adrienne Morey, Nirmala Pathmanathan, Rodney Scott, Peter Simpson, Allan Spigelman, Nicholas Wilcken, Desmond Yip, and Nikolajs Zeps.

kConFab/AOCS Investigators are Adrienne Sexton, Alex Dobrovic, Alice Christian, Alison Trainer, Allan Spigelman, Andrew Fellows, Andrew Shelling, Anna De Fazio, Anneke Blackburn, Ashley Crook, Bettina Meiser, Briony Patterson, Christine Clarke, Christobel Saunders, Clare Hunt, Clare Scott, David Amor, David Gallego Ortega, Deb Marsh, Edward Edkins, Elizabeth Salisbury, Eric Haan, Finlay Macrea, Gelareh Farshid, Geoff Lindeman, Georgia Trench, Graham Mann, Graham Giles, Grantley Gill, Heather Thorne, Ian Campbell, Ian Hickie, Liz Caldon, Ingrid Winship, James Cui, James Flanagan, James Kollias, Jane Vissvader, Jennifer Stone, Jessica Taylor, Jo Burke, Jodi Saunus, John Forbes, John Hopper, Jonathan Beesley, Judy Kirk, Juliet French, Kathy Tucker, Kathy Wu, Kelly Phillips, Laura Forrest, Lara Lipton, Leslie Andrews, Lizz Lobb, Logan Walker, Maira Kentwell, Mandy Spurdle, Margaret Cummings, Margaret Gleeson, Marion Harris, Mark Jenkins, Mary Anne Young, Martin Delatycki, Mathew Wallis, Matthew Burgess, Melissa Brown, Melissa Southey, Michael Bogwitz, Michael Field, Michael Friedlander, Michael Gattas, Mona Saleh, Morteza Aghmesheh, Nick Hayward, Nick Pachter, Paul Cohen, Pascal Duijff, Paul James, Pete Simpson, Peter Fong, Phyllis Butow, Rachael Williams, Rick Kefferd, Rodney Scott, Roger Milne, Rosemary Balleine, Sarah-Jane Dawson, Sheau Lok, Shona O’Connell, Sian Greening, Sophie Nightingale, Stacey Edwarsd, Stephen Fox, Sue-Anne McLachlan, Sunil Lakhani, Tracy Dudding, and Yoland Antill.

NBCS collaborators are Kristine K. Sahlberg, Lars Ottestad, Rolf Kåresen, Ellen Schlichting, Marit Muri Holmen, Toril Sauer, Vilde Haakensen, Olav Engebraeten, Bjorn Naume, Alexander Fosså, Cecile E. Kiserud, Kristin V. Reinertsen, Åslaug Helland, Margit Riis, Jürgen Geisler, and OSBREAC.

Acknowledgments
BCAC was funded by Cancer Research UK (C1287/A16563) and by the European Community’s Seventh Framework Programme under grant agreement no. 223175 (HEALTH-F2-2009-223175) (COGS) and by the European Union’s Horizon 2020 Research and Innovation Programme under grant agreements 633784 (B-CAST) and 634935 (BRIDGES). Genotyping of the OncoArray was principally funded by Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344), the Ministère de l’Économie, de la Science et de l’Innovation du Québec through Genome Québec, the Quebec Breast Cancer Foundation; NIH grants U19 CA148065 and X01HG007492; and Cancer Research UK (C1287/A10118 and C1287/A16563). Genotyping of the iCOGS array was funded by the European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710), the Canadian Institutes of Health Research for the “CIHR Team in Familial Risks of Breast Cancer” program, and the Ministry of Economic Development, Innovation and Export Trade of Quebec (grant # PSR-SIRI-701). Combining the GWAS data was supported in part by the National Institutes
of Health (NIH) Cancer Post-Cancer GWAS initiative grant; No. 1 U19 CA 148065 (DRIVE, part of the GAME-ON initiative). We thank all the individuals who took part in this study and all researchers, clinicians, technicians, and administrative staff who enabled this work to be carried out. For other acknowledgments and sources of funding, see Supplemental Acknowledgments.

Declaration of Interests

D.G.E. reports grants from AstraZeneca and AmGen, outside the submitted work; U.M. has stock ownership and has received research funding from Abcodia Pvt Ltd.; A. Smeets reports other from MSD, outside of the submitted work; P.A.F. reports grants and personal fees from Novartis and personal fees from Pfizer, Roche, Teva, and Celgene, outside the submitted work; R.C. declares personal fees from Novartis, AstraZeneca, and Genentech, outside the submitted work. B.R. reports funding for the conduct of the clinical Success trial paid to her institution from AstraZeneca, Chugai, Lilly, Novartis, Veridex (now Janssen Diagnostics), and Sanoﬁ Aventis. M. Robson reports grants, personal fees, and non-ﬁnancial support from AstraZeneca, personal fees from McKesson, grants and personal fees from Pfizer, non-ﬁnancial support from Myriad, non-ﬁnancial support from Invitae, and grants from AbbVie, Tesaro, and Medivation, outside the submitted work; and M.P.L. reports personal fees from Novartis, Pfizer, Roche, Teva, AstraZeneca, Lilly, and Eisai, outside the submitted work.

Received: August 9, 2018
Accepted: November 3, 2018
Published: December 13, 2018

Web Resources

BCAC data access, http://bcac.ccge.medschl.cam.ac.uk
Nomis (26 March 2018), https://www.nomisweb.co.uk/
West Midlands Cancer Intelligence Unit, http://www.wmcieu.nhs.uk/

References