Original research

Evaluation of European-based polygenic risk score for breast cancer in Ashkenazi Jewish women in Israel


ABSTRACT

Background Polygenic risk score (PRS), calculated based on genome-wide association studies (GWASs), can improve breast cancer (BC) risk assessment. To date, most BC GWASs have been performed in individuals of European (EUR) ancestry, and the generalisation of EUR-based PRS to other populations is a major challenge. In this study,
we examined the performance of EUR-based BC PRS models in Ashkenazi Jewish (AJ) women.

Methods We generated PRSs based on data on EUR women from the Breast Cancer Association Consortium (BCAC). We tested the performance of the PRSs in a cohort of 2161 AJ women from Israel (1437 cases and 724 controls) from BCAC (BCAC cohort from Israel (BCAC-IL)). In addition, we tested the performance of these EUR-based BC PRSs, as well as the established 313-SNP EUR BC PRS, in an independent cohort of 181 AJ women from Hadassah Medical Center (HMC) in Israel.

Results In the BCAC-IL cohort, the highest OR per 1 SD was 1.56 (±0.09). The OR for AJ women at the top 10% of the PRS distribution compared with the middle quintile was 2.10 (±0.24). In the HMC cohort, the OR per 1 SD of the EUR-based PRS that performed best in the BCAC-IL cohort was 1.58±0.27. The OR per 1 SD of the commonly used 313-SNP BC PRS was 1.64 (±0.28).

Conclusions Extant EUR GWAS data can be used for generating PRSs that identify AJ women with markedly elevated risk of BC and therefore hold promise for improving BC risk assessment in AJ women.

WHAT IS ALREADY KNOWN ON THIS TOPIC
- Genome-wide association studies (GWASs) on breast cancer (BC) were, to date, mainly done on women of European (EUR) ancestry, and recent studies showed that polygenic risk score (PRS) based on these GWAS can effectively stratify EUR women according to their BC risk.
- However, PRS performance declines with the increase of the genetic distance between the population used in the GWAS and the population on which the PRS is applied.

WHAT THIS STUDY ADDS
- Here, we systematically evaluated the performance of EUR-based BC PRS on Ashkenazi Jewish (AJ) women from Israel. Our results demonstrate that extant EUR GWAS data can be used for generating PRSs that identify AJ women with markedly elevated risk of BC.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY
- Our study suggests the possibility of personalised BC screening programmes in Israel that could potentially improve early detection of BC while reducing overdiagnosis.

INTRODUCTION
Breast cancer (BC) is the most common cancer diagnosed among women in Western countries including Israel, where some 5500 BC cases are diagnosed annually. An early diagnosis of BC leads to a higher cure rate and improved survival. Thus, it is essential to develop accurate risk prediction methods for identifying women at high risk of BC. An ongoing debate over the optimal approach to BC screening has led to discordant professional society recommendations. Two fundamental questions—whether to screen annually or at a lower frequency and whether screening should start at the age of 40 or at a later point in life—have been debated for over 20 years.

In Israel, health providers generally recommend biennial mammography screening starting at age 50 for women, except for those with a family history of relevant cancer or carriers of pathogenic variants in BC-associated genes, who are recommended to start earlier and screen more frequently. This ‘one size fits all’ approach to nationwide BC screening might be suboptimal as it assumes an equal risk of developing BC to most women. A personalised screening strategy might improve early detection of BC while reducing overdiagnosis.

The decline in PRS performance in non-EUR populations might aggravate disparities in clinical genetics care between ethnic groups. Several studies showed that BC PRS generated from EUR GWAS summary statistics (EUR BC PRS) has lower performance on non-EUR women (eg, African-Americans). Yet, some studies demonstrated that EUR BC PRS performance on Latin American women—a large group with variable levels of Indigenous American, EUR and African ancestries—was similar to its performance on women of EUR ancestry.

The population in Israel is highly heterogeneous, with Ashkenazi Jews (AJ) being one of its largest ethnic group. Given the relatively low genetic distance between the EUR and AJ populations, we hypothesised that EUR BC PRS could be used to develop clinically relevant PRS models for AJ women in Israel. To that end, we used the massive genetic resource generated by the multinational Breast Cancer Association Consortium (BCAC), which also contains an Israeli cohort, to conduct a systematic evaluation of the predictive performance of EUR BC PRS models on Israeli AJ women. We demonstrate that an EUR BC PRS can be adjusted to the AJ population and identify women with markedly elevated BC risk (OR >2.0 for AJ women in the top 10% compared with the middle quintile). We substantiate these findings using an independent cohort of AJ Israeli women.

MATERIALS AND METHODS

BCAC dataset
We analysed 132 335 EUR women from the BCAC: 72 899 cases and 59 436 controls. In addition, the BCAC includes an Israeli cohort (BCINIS/BCAC cohort from Israel (BCAC-IL)) of 2161...
women: 1437 cases and 724 controls. According to the 'ethnOt' field in the BCAC phenotype file, all the women in the BCAC-IL cohort are tagged as 'Jewish Ashkenazi'. In addition, there are 73 samples in the EUR cohort that are tagged as AJ.

All samples analysed were genotyped using the OncoArray chip. In our analysis, we used an imputed version of the data provided by BCAC. The imputation was done against the 1000 Genomes Project imputation panel. In BCAC-IL, 119 (5.5%) BRCA1/2 mutation carriers were identified by a self-reporting field provided by the BCAC.

**Hadassah Medical Center (HMC) cohort**

The HMC dataset contains 181 Israeli AJ women, of whom 118 are BC cases under the age of 45 years and 63 are controls older than 75 years. We validated either by sequencing or genotyping that none of the women carried one of the three AJ founder mutations in BRCA1/2. Samples were genotyped using the Axiom PMDA chip. Likely pathogenic variants in selected genes are covered by this chip. Three women carried such variants in BRCA1/2, and none bore pathogenic variants in other BC susceptibility genes.

We phased the data using SHAPEIT2 and imputed it using IMPUTE2. The imputation reference panel was generated using SHAPEIT2 from the EUR samples from the 1000 Genomes Project (n=503). Using PLINK, we filtered out SNPs with uncertainty greater than 0.1.

For the evaluation of the 313 PRS, we were able to map 304 SNPs, of which 248 were called (either by genotyping or imputation) in more than 90% of the samples.

**Quality check (QC) of discovery sets**

We performed QC on each discovery set using PLINK. We kept only SNPs with minor allele frequency (MAF) of ≥5%, HWE p value of ≥1e-6 and missing rate of ≤10%. In addition, we kept only samples where less than 10% of SNPs present in the set were missing. In addition, we filtered out ambiguous and duplicated alleles. A total of 4,617,515 SNPs remained in the BCAC-EUR cohort and 4,973,754 SNPs in the BCAC-EUR cohort after exclusion of the Polish samples.

Similarly, we used PLINK to perform QC on each target set. We kept the same HWE, missing rate and MAF thresholds as in the discovery set, filtered out duplicated alleles and kept samples where less than 10% of SNPs present in the set were missing. This process left 5,549,031 and 5,704,856 SNPs on the entire Israeli (BCAC-IL) and, as used as a control, the entire Polish (BCAC cohort from Poland (BCAC-PL)) cohorts, respectively. Note that in cross-validation (CV) analyses (see further), to avoid information leakage, we performed QC on each fold separately, so the number of SNPs in each fold slightly varied, depending on the subset of individuals in the fold.

**GWAS analysis**

We ran GWAS analyses for two sets: EUR (n=132,335) and EUR without the PL cohort (n=128,153). Both sets did not contain the BCAC-IL women. For each analysis, we ran PCA and GWAS using PLINK (with the --glm command) and generated GWAS summary statistics with the first five principal components as covariates.

**Nested CV**

We applied nested CV for optimising PRS models generated by four different methods (pruning and thresholding using European linkage disequilibrium (P+T EUR-LD), pruning and thresholding using linkage disequilibrium of the target population (P+T target set LD), LDpred2 and Lassosum; see further). Specifically, for each PRS method, we split the BCAC-IL cohort into six sets (each of size 360). Next, we held out one set (red box in figure 1) and used the other five sets (green boxes in figure 1) to perform a standard 5-fold CV, in which four out of five parts (training set; light green) are used to derive PRS model on the five CV folds with the chosen hyperparameters. Finally, we applied the resulting PRS model on the holdout set and measured the OR per 1 SD and for the top 10% OR. We repeated this entire process six times, each with a different holdout set. We applied this scheme to each of the four PRS methods included in our analysis (P+T EUR-LD, P+T target LD, LDpred2 and Lassosum). The method that obtained the highest average performance on the six holdout sets is selected as the best one. AJ, Ashkenazi Jewish; BCAC, Breast Cancer Association Consortium; BCAC-IL, BCAC cohort from Israel; CV, cross-validation; PRS, polygenic risk score; P+T EUR-LD, pruning and thresholding using European linkage disequilibrium; P+T target LD, pruning and thresholding using linkage disequilibrium of the target population.
models with different predefined sets of hyperparameters, and then the resulting models are applied on the fifth part (validation set, dark green). For each model, we measured the OR per 1SD (using logistic regression with the first six principal components as covariates) and OR of women at the top 10% of the PRS distribution compared with the middle quintile. After iterating over the five combinations of training and test sets, we chose the hyper-parameter set that performed the best on average (see detailed ranking criteria below). Then, using these optimal hyper-parameters, we retrained a PRS model on the entire five CV folds (green boxes). Finally, we applied the resulting PRS model on the holdout set and measured the OR per 1SD and top 10% OR. We repeated this entire process six times, each with a different holdout set. The method with the highest average on the six holdout sets is nominated as the best one.

In all analyses, PRS were standardised to the control samples of the respective target set.

Criteria for choosing an optimal PRS model
We tested the performance of each PRS method with a predefined set of hyper-parameters (see below). For each method, we ranked runs with different hyper-parameters using two metrics: (1) OR per 1SD and (2) top-10% OR, and combined these rankings by taking their sum. We broke ties using the model with the higher OR per 1SD, as this metric is less noisy.

Pruning and thresholding using European linkage disequilibrium
Using PLINK, we clumped the GWAS results according to LD in the EUR population derived from the EUR samples in the 1000 Genomes Project (n=503) with $r^2 = 0.2$. Then, we filtered the remaining SNPs based on a significance threshold (T). We tested the following threshold values T:

- $5 \times 10^{-8}$, $10^{-7}$, $10^{-6}$, $10^{-5}$, $10^{-4}$, $10^{-3}$,
- $5 \times 10^{-3}$, $10^{-2}$, $5 \times 10^{-2}$, 0.1, 0.2, 0.3, 0.4, 0.5

For each T, we calculated the PRS from the SNPs that passed the filtering.

Pruning and thresholding using linkage disequilibrium of the target population
Here, when applying LD clumping in PLINK, we used LD inferred from the training set. The training set comes from the same population as the target set. Namely, in each fold of the CV, LD was calculated using the genotype data of individuals in the training set. On the HMC cohort, we used the LD from the BCAC-IL cohort. The subsequent steps of the analysis are identical to the P+T EUR-LD method.

LDpred2
LDpred2 (grid mode) generates a PRS model using SNP correlations calculated from genotype data (ie, the training set). We supplied LDpred2 with a training set that comes from the same population as the target set, as for the P+T method previously. We ran LDpred2 using the set of hyper-parameter values for the proportion of causal variants, heritability, and sparseness that were recommended by. The rest of the hyper-parameters were left with their default values.

Lassosum
Lassosum generates a PRS model using a reference panel calculated from genotype data (ie, the training set). We supplied Lassosum with a training set that comes from the same population as the target set, as above. We ran Lassosum using LD blocks option “EUR.hg19” and the values of the regularisation hyper-parameter $\lambda$ that were recommended by. The rest of the hyper-parameters were left with their default values.

313-SNPs EUR BC PRS model
We downloaded the weights for the EUR PRS model from. Originally, the model consisted of 313 SNPs. In the imputed data, we managed to retain all the 313 SNPs for the BCAC-IL cohort and 304 SNPs for the HMC cohort. Risk scores for each sample were calculated using PLINK.

RESULTS
We set to build and evaluate EUR-based BC PRS for AJ women from Israel. For this task, we used an Israeli cohort of 2161 AJ women (1437 BC cases and 724 controls) that is a part of the BCAC (Methods). We refer to the Israeli sub-cohort of the BCAC as BCAC-IL. In order to avoid inflation of the predictive performance, the target set should be independent of the discovery set. Therefore, we could not reliably assess how the commonly used EUR BC 313-SNP PRS performs on the BCAC-IL cohort since this PRS was derived from BCAC GWAS, which included the EUR-IL cohort. Therefore, we first removed the Israeli women from the EUR BCAC cohort and recomputed GWAS summary statistics using only data from the 132 335 non-Israeli EUR women (72 899 cases and 59 436 controls; Methods). A PCA on the BCAC genotype data confirmed the close genetic relatedness of AJ to the EUR population (figure 2).

Next, we set to adapt an EUR-based BC PRS for AJ women from Israel. We constructed PRS models from the GWAS we generated using four different methods: P+T, EUR-LD; P+T using LD of the target (AJ) population (P+T target LD), LDpred2 and Lassosum. We used two metrics to evaluate the models produced by these algorithms: (1) the OR per 1 unit SD and (2) the OR of women in the top 10% of the PRS distribution relative to those in the middle quintile (top 10% OR). We constructed and evaluated the PRS models using a nested CV scheme (see the Materials and methods section). The outline of our evaluation procedure is depicted in figure 1.

Of the four methods we tested, Lassosum performed best, obtaining an OR per 1 SD of 1.56 (±0.09) and a top 10% OR of 2.1 (±0.24) (table 1 and online supplemental figure S1; see online supplemental table S1 for performance on the validation sets in the CV). We also examined the OR of other deciles of the PRS (compared with the middle quintile) and found that it increased nearly monotonically (figure 3). Further, women in the top 10% were estimated to have fourfold higher OR for BC compared with AJ women in the bottom 10% (figure 3, online supplemental figure S2). Notably, these top and bottom 10% OR estimates that we obtained for AJ women were comparable to those reported using EUR BC PRS on women of EUR ancestry.12

Next, to estimate the decline in the performance of EUR-based BC PRS when applied to AJ women relative to women of EUR ancestry, we compared the performance obtained on women from BCAC-IL and women from BCAC-PL. We compared BCAC-IL to the Polish cohort as the AJ population is mainly from Eastern Europe. Specifically, we now excluded the Polish and Israeli samples from the BCAC discovery set and reran a GWAS analysis (see the Materials and methods section). Then, we applied the same nested CV scheme to the BCAC-PL (4537 women: 2318 cases and 2219 controls) and BCAC-IL cohorts using the same four PRS methods as previously discussed. As expected, the results obtained on BCAC-PL were mostly higher
than those on BCAC-IL, reflecting the greater genetic distance of the AJ population from the EUR population (table 2; see online supplemental table S2 for performance on the validation sets).

Pathogenic variants in BRCA1/2 confer a very high risk of BC. In BCAC-IL, 119 women were flagged as carriers of the BRCA1/2 mutation (106 cases and 13 controls). To test the impact of the inclusion of these BRCA1/2 carriers on PRS performance, we measured the performance of the P+T EUR-LD PRS on the BCAC-IL cohort after excluding these 119 samples. As shown in online supplemental figure S3, there was no significant difference between the two runs in the estimates for the OR per 1 SD and the top 10% OR.

To further examine the performance of EUR-based BC PRS on AJ women in Israel, we genotyped an independent sample of 181 Israeli AJ women recruited at the HMC in Jerusalem. This cohort comprises 118 patients with BC and 63 healthy women as controls. All the patients in the HMC cohort were diagnosed with BC at an early age (<45 years old) and tested negative for the three AJ founder variants in BRCA1/2. The controls were women aged 75 years and over who were never diagnosed with

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Performance of different PRS methods on the BCAC-IL cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>OR per 1SD</td>
</tr>
<tr>
<td>P+T EUR-LD</td>
<td>1.43±0.08</td>
</tr>
<tr>
<td>P+T target set LD</td>
<td>1.39±0.07</td>
</tr>
<tr>
<td>LDpred2</td>
<td>1.31±0.07</td>
</tr>
<tr>
<td>Lassosum</td>
<td>1.56±0.09</td>
</tr>
</tbody>
</table>

Performance of different PRS methods on the BCAC-IL cohort. ORs per 1 SD and top 10% OR were obtained using the nested CV outlined in figure 1. The last column is the average number of SNPs. Shown are means and SEMs over the six holdout sets. BCAC-IL, BCAC cohort from Israel; CV, cross validation; PRS, polygenic risk score; P+T EUR-LD, pruning and thresholding using European linkage disequilibrium; P+T target set LD, pruning and thresholding using linkage disequilibrium of the target population.
We first evaluated how the EUR BC 313-SNP PRS (313 PRS) performs on this cohort. Notably, the OR per 1 SD of the 313-PRS model was 1.64±0.28 on the HMC cohort, similar to the effect reported for this PRS model on EUR women (1.65 OR per 1 SD, 95% CI 1.59 to 1.79)\(^{12}\). For comparison, we also measured the performance of the 313 PRS on the BCAC-IL cohort and obtained OR per 1 SD of 1.77±0.09. This result is likely inflated due to the inclusion of the BCAC-IL in the discovery set used to infer the 313-PRS model. On the other hand, the OR estimate for the BCAC-IL cohort was less noisy than the one obtained in the HMC cohort due to its larger size (the BCAC-IL cohort is >10 times larger than the HMC).

Last, we evaluated Lassosum—the best performing method on BCAC-IL—on HMC. Using the EUR GWAS we generated, we trained the PRS model on the BCAC-IL cohort in fivefold CV (online supplemental figure S4). Applying this PRS to the HMC cohort yielded an OR of 1.58±0.27 per 1 SD (number of SNPs: 4540).

Overall, the results obtained on the HMC cohort reaffirm that EUR-based BC PRS has clinically relevant predictive capacity for Israeli AJ women.

**DISCUSSION**

PRS models have the potential to play an essential role in detecting women’s risk of developing BC. Nevertheless, at present, clinically relevant BC PRS models have been constructed primarily for women of EUR ancestry, for whom large discovery sets are currently available.\(^8\) Whether these models perform well on women of other ancestries and how they can be adapted for women of other ancestries are key open questions. Our study focuses on a major ethnic group in Israel, the Ashkenazi Jewish (AJ) population, which is genetically close to the EUR population. We tested whether a large number of available EUR genotypes of patients with BC and healthy women could be used to generate a clinically relevant BC PRS model for AJ women in Israel.

We evaluated four PRS methods on the Israeli cohort from BCAC (BCAC-IL) and found that Lassosum had the best prediction performance. Notably, there was a fourfold increased BC risk between women in the top and bottom 10% of the PRS distribution (figure 3 and online supplemental figure S2), suggesting that BC PRS models derived from EUR GWAS may help fit personalised recommendations for BC preventive screening for Israeli AJ women. The results obtained on the independent HMC cohort further support this conclusion. While the BCAC-IL cohort is too small to calculate reliable risk estimates for women in the top 5% and 1%, the monotonic increase of the OR with the deciles (figure 3) and results by similar BC PRS on EUR women\(^{12}\) suggest that this model has the capacity to identify at its very top percentiles AJ women with even higher risk of developing BC. Follow-up studies with larger samples of AJ women are needed to substantiate this expectation.

Notably, the HMC cohort has extreme age differences between the case and control arms: healthy women are older than 75 and patients with BC are younger than 45. Thus, the high prediction performance of the BC PRS models on this cohort suggests that EUR-based PRS models may also be relevant for detecting early-onset cases of BC among Israeli AJ women. In addition, these results indicate that for AJ women, low-impact common genetic variants—and not only pathogenic variants with high and moderate impact—play an important role in predisposing women to early-onset BC.

One limitation of our study is that BRCA1/2 carriers were identified in the BCAC-IL only by self-reporting. Thus, there might be additional women carrying BRCA1/2 variants who were marked as non-carriers as identified by.\(^30\) Still, our analysis indicates that inclusion of a limited group of patients who carry pathogenic variants in BRCA1/2 genes does not have a significant impact on the PRS performance (online supplemental figure S3).

As the patients with BC at HMC were under 45, we could not directly generalise the prediction performance obtained on HMC for older AJ Israeli patients. However, online supplemental figure S5 indicates that there is no substantial difference in the PRS between age groups of BCAC-IL patients, consistent with previous findings on EUR population.\(^12\)

Our finding indicates that the currently available EUR BC GWAS data can be used to generate BC PRS models for Israeli AJ women. Nevertheless, this observation should not nullify the effort to genotype a higher number of individuals in Israel. First, an increased sample of AJ women would provide more accurate risk estimates for women at the top tail of the PRS distribution. Second, the Israeli population is highly heterogeneous, comprising many different ethnic groups, including North African and Middle Eastern Jews, as well as Palestinians, Druzes and Bedouins. Moreover, many of the younger generation in Israel are of mixed ethnicities. Therefore, to cover additional groups in nationwide BC prevention programmes, large-scale genotyping initiatives should include women from other ethnic groups in Israel, including admixed groups. Such data would allow a systematic evaluation of EUR-derived PRS BC models on non-AJ Israeli populations. We hope that this study will expedite the realisation of the potential for personalised BC risk stratification and encourage the development of screening protocols for high-risk women.

**Author affiliations**

1. The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel
2. Department of Human Molecular Genetics and Biochemistry, Tel Aviv University, Tel Aviv, Israel
3. Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel
4. Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel
5. Institute of Oncology, Davidoff Cancer Center, Robin Medical Center, Beilinson Hospital, Petah Tikva, Israel
6. Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel
7. Department of oncology, Hadassah Medical Center, Jerusalem, Israel
8. Hebrew University of Jerusalem, Jerusalem, Israel

---

**Table 2** Performance of EUR PRS when excluding the Polish and Israeli cohorts from the discovery set and using these respective populations as the target cohorts

<table>
<thead>
<tr>
<th>Method</th>
<th>Target set cohort</th>
<th>OR per 1 SD</th>
<th>Top 10% OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>P+T (EUR)</td>
<td>BCAC-IL</td>
<td>1.37±0.06</td>
<td>2.41±0.76</td>
</tr>
<tr>
<td>P+T target LD</td>
<td>BCAC-IL</td>
<td>1.36±0.03</td>
<td>1.64±0.17</td>
</tr>
<tr>
<td>LDpred</td>
<td>BCAC-IL</td>
<td>1.50±0.06</td>
<td>1.82±0.33</td>
</tr>
<tr>
<td>Lassosum</td>
<td>BCAC-IL</td>
<td>1.50±0.06</td>
<td>1.82±0.33</td>
</tr>
</tbody>
</table>

Shown are average ORs per 1 SD and top 10% ORs. Errors were measured using SEM for the six holdout sets. BCAC-IL, BCAC cohort from Israel; BCAC-PL, BCAC cohort from Poland; P+T, pruning and thresholding; P+T target LD, pruning and thresholding using linkage disequilibrium of the target population.
Acknowledgements We thank all the individuals who took part in these studies and all the researchers, clinicians, technicians and administrative staff who enabled this work to be carried out. ABCFS thanks Maggie Angelakos, Judi Maskiell and all the researchers, clinicians, technicians and administrative staff who enabled this work to be carried out. ABCFS thanks Maggie Angelakos, Judi Maskiell and all the researchers, clinicians, technicians and administrative staff who enabled this work to be carried out.
Follow-up investigations, the teams that recruited the participants and continue working on follow-up, and the many thousands of Melbourne residents who continue to participate in the study. The MDISS study group acknowledges the former principal investigator, Professor Henrik Olsson. MSSkCC thanks Marina Coimbra and Lauren Jacobs, MTLSBECs thanks Martine Tranchant (CHU de Québec – Université Laval Research Centre, Marie-France Valois, Annie Turgeon and Lea Héry-Carcione (McGill University Health Center, Royal Victoria Hospital; McGill University) for DNA extraction, sample management and skilled technical assistance. JS is chair holder of the Canada Research Chair in Oncogenomics. The following are NBCS collaborators: Kristine K Salihbeg (PhD), Anne-Lise Berresen-Dale (Prof Emeritus), Lars Oostedt (MD), Rolf Käresen (Prof Emeritus), Dr Ellen Schlüchtung (MD), Marti Muri Holten (MD), Toril Sauer (MD), Vilide Haakenlsen (MD), Olav Engebretsen (MD), Bjørn Naume (MD), Alexander Fossa (MD), Cecile E. Kiserud (MD), Kristin V. Reinertsen (MD), Aslauk Helland (MD), Margit Ris (MD), Jürgen Geisler (MD), OSARCE and Grethe I Grenaker Alnas (MSc). NBHS and SBCGG thanks the study participants and their staff for contributions to the study. We thank the participating staff and the NHS for their valuable contributions as well as the following state cancer registries for their help. AL, AZ, AR, CA, CO, CT, DE, FL, GA, IA, ID, IN, IA, KS, KY, LA, ME, MA, MI, MN, MO, MS, MT, NC, ND, NE, OK, OR, PA, RI, SC, TN, TX, VA, WA and WY. The authors assume full responsibility for analyses and interpretation of these data. OBCS thanks Arja Jukkola-Vuorio, Mervi Grip, Saja Kaupilla, Meeri Otsuka, Leena Keski-Katajan and Kari Mononen for their contributions to this study. The OFBRC thanks Teresa Salander, Nayanee Weerasooriya and Steve Gallinger. ORIGO thanks Er Krol-Warmerdale and J Blom for patient accrual, administering questionnaires and managing clinical information. The LUMC survival data were retrieved from the Leiden hospital-based cancer registry system (ONCOCODE) with the help of Dr J Molenaar. PBCS thanks Louise Brinton, Mark Sherman, Neoliana Szelesz-Dabrowoska, Beata Peplonska, Witold Zatonski, Pei Chao and Michael Stagner. We thank staff in the Experimental Cancer Medicine Centre, which supported the Faculty of Medicine Tissue Bank and the Faculty of Medicine DNA Banking resource. The authors acknowledge the role of the Breast Cancer Now Tissue Bank in collecting and making available the samples and/or data, and the patients who have generously donated their tissues and shared their data to be used in the generation of this publication. PREFACE thanks Sonja Oesser and Silke Lindhardt. The RBCS thanks Jannet Blom, Saskia Gelders, Wendy J. Crapper—van de Smitten, and the Erasmus MC Family Cancer Clinic. SBCS thanks Sue Higham, Helen Cramp, Dan Connelly, Ian Brock, Sabapathy Balasubramanian and Malcolm W R Reed. We thank the SEARCH and EPIC teams. SGBCC thanks the participants and all colleagues for their excellent help with recruitment, data and sample collection. SKKOKZFS thanks all study participants, clinicians, family doctors, researchers and technicians for their contributions and commitment to this study. We thank the SUCCESS Study teams in Munich, Duesseldorf, Erlangen and Ulm. UBCS thanks all study participants as well as the ascertainment, laboratory, analytics and informatics teams at Huntsman Cancer Institute and Intermountain Healthcare for their important contributions to this study. UCBS thanks Irene Masunaka. UKGBS thanks Breast Cancer Now and the Institute of Cancer Research for support and funding of the Generations Study, and the study participants, study staff and the doctors, nurses and other health care providers and information sources who have contributed to the study. We acknowledge NHS funding to the Royal Marsden/ICR NIHR Biomedical Research Centre.


Contributors RE and KS conceived the project, HL designed and developed and performed the analyses and supervised the preparation and submission of RE and RS. SC, SR, NE and SBS were consulted on the analyses. NE, SBS, RE, DZ, and YP collected and provided the HMC data. The rest of the authors are part of Breast Cancer Association Consortium (BCAC), which contributed the BCAC dataset. All authors wrote the manuscript and approved the final manuscript. Guarantor: RE.

Funding This study was supported in part by grants from the Israel Science Foundation (number 3165/19, within the Israel Precision Medicine partnership program; number 22062/22 to RS, and number 407/17 to SC), from the Tel Aviv University Center for AI and Data Science (TAD) to RE and RS, by a joint program grant from the Cancer Biology Research Center, Djerassi Oncology Center, Edmond J. Safra Center for Bioinformatics and TAD to RE, and by the Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics to RE and RS. HL was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. The Breast Cancer Association Consortium (BCAC) is funded by the Confluence project, which is funded with intramural funds from the National Cancer Institute Intramural Research Program, National Institutes of Health. Additional funding for BCAC is provided by Cancer Research UK grant: PP/RPGM-Nov20100002, the European Union’s Horizon 2020 Research and Innovation Programme (grant numbers 634935 and 633784 for BRIDGES and B-CAST, respectively), and the PERSPECTIVE I&I project, funded by the government of Canada through Genome Canada and the Canadian Institutes of Health Research, the Ministère de l’Économie et de l’Innovation du Québec through Genome Québec, the Quebec Breast Cancer Foundation. The EU Horizon 2020 Research and Innovation Programme funding source had no role in study design, data collection, data analysis, data interpretation or writing of the report. Genotyping of the OncotypeDX was funded by the NIH (grant U19 CA164587), and Cancer Research UK (grant C1287/A16563) and the PERSPECTIVE project supported by the government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344), and the Ministère de l’Économie, Science et Innovation du Québec through Genome Québec and the PSI-SIRI-701 grant and the Quebec Breast Cancer Foundation. Funding for iCOGS came from the European Community’s Seventh Framework Programme (FP7/2007-2013) grant agreement 223175. (COGS), Cancer Research UK (C1287/A10118, C1287/ A10170, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15070, C5047/A10692 and C13197/A16565), the National Institutes of Health (CA128978) and Post-Cancer GWAS Initiative (U19 CA145387, U19 CA134005 and U19 CA148112 – the GAME-ON initiative), the Department of Defence (WB1XXH-10-004), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, and Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund. The BRIDGES panel sequencing was supported by the European Union Horizon 2020 research and innovation program BRIDGES (grant number, 634935) and the Wellcome Trust (v203477/2/16/2). The Australian Breast Cancer Foundation (ABCFS) was supported by grant U116 CA146920 from the National Cancer Institute (USA). The content of this manuscript does not reflect the views of, nor does it necessarily represent endorsement by the US Government or the BCFR. The ABCFS was also supported by the National Health and Medical Research Council of Australia, the New South Wales Cancer Council, the Victorian Health Promotion Foundation (Australia) and the Victorian Breast Cancer Research Consortium. JLH is a National Health and Medical Research Council (NHMRC) Senior Principal Research Fellow. MCS is an NHMRC senior research fellow. The ABCS study was supported by the Dutch Cancer Society (grants NKI 2007-3839 and 2009 4363) and an institutional grant of the Dutch Cancer Society and of the Dutch Ministry of Health, Welfare and the Sport. The Australian Breast Cancer Tissue Bank was supported by the National Health and Medical Research Council of Australia; the Cancer Institute New South Wales and the National Breast Cancer Foundation. The AHS study is supported by the intramural research program of the National Institutes of Health, the National Cancer Institute (grant number 201-CP010119), and the National Institute of Environmental Health Sciences (grant number 201-ES049030). The work of the BCFR was partly funded by ELAN-Fond of the University Hospital of Erlangen. The BBCS is funded by Cancer Research UK and Breast Cancer Now and acknowledges NHS funding to the NIHR Biomedical Research Centre, and the National Cancer Research Network (NCRN). The BCEE was funded by the National Health and Medical Research Council, Australia and the Cancer Council Western Australia and acknowledges funding from the National Breast Cancer Foundation (JS). For the BCFR-NY, BCFR-PA, BCFR-UT this work was supported by grant U116 CA146920 from the National Cancer Institute. The content of this manuscript does not necessarily reflect the views, policies or the National Cancer Institute or any of the collaborating centers in the BCFR, nor does mention of trade names, commercial products, or organisations imply endorsement by the US Government or the BCFR. The BCNIS study is supported in part by the Breast Cancer Research Foundation (BCRF). For BIGGS, ES is supported by NHIR Comprehensive Biomedical Research Centre, Guy’s & St. Thomas’ NHS Foundation Trust in partnership with King’s College London, United Kingdom. It is supported by the Oxford Biomedical Research Centre Research and Development Fund and Cancer Research UK (C8620/AB372/A15106) and the Institute of Cancer Research (UK). BOCs acknowledges NHS funding to the Royal Marsden / Institute of Cancer Research NIHR Specialist Cancer Biomedical Research Centre. The Breast Oncology Galician Network is funded by Acción Estratégica de Salud del Instituto de Salud Carlos III FIS PI12/01215/Colfonadiado y FECEDER FEDER-Acción Estratégica de Salud del Instituto de Salud Carlos III [PI13/10136]; Programa Grupos Emergentes, Cancer Genetics Unit, Instituto de Investigacion Biomedica Galicia Sur. Xerencia de Xestion Integrada de Vigo-SEGRAS, Instituto de Salud Carlos III, Spain; Grant 10CSA012E, Consellera de Industria.
Wellcome Trust (Wellcome Trust Strategic Award, ‘Stratifying Resilience and Depression Longitudinally’, reference 104362/2/14/F2), Funding for identification of cases and contribution to BCAC was provided in part by the Wellcome Trust Seed Award, ‘Temporal trends in incidence and mortality of molecular subtypes of breast cancer to inform public health, policy and prevention’ (reference 207800/1/17). The GEPIA study was conducted by the Dietmar-Hopp-Foundation, Helmholtz Society and the German Cancer Research Center (DKFZ). CBCS was funded by the Danish Cancer Society and the German Cancer Research Center (DKFZ). CBCS was supported by the Deutsche Krebshilfe e.V. (70492) and the DKFZ. GLACIER was supported by Breast Cancer Now, CRUK and Biomedical Research Centre at Guy’s and St Thomas’ NHS Foundation Trust and King’s College London. The HABCS study was supported by the Hungarian Research Foundation (DFG Do761/15-1), the Claudia von Schilling Foundation for Breast Cancer Research, by the Lower Saxonian Cancer Society and by the Rudolf Bartling Foundation. The HABCS was financially supported by the Helsinki University Hospital Research Fund, the Heidi Josel Foundation and the Cancer Foundation Finland. The HEBON study is supported by the Dutch Cancer Society (grants NK1998-1854, NKI2004-3088 and NKI2007-3576), the Netherlands Organisation of Scientific Research (grant NWO 11109024), the Pink Ribbon (grants 110005 and 2014-187.W076), the BBMRI (grant NWO 184.021.007.CP46) and the Transcan (grant JTC 2012 Cancer 12-054). The HMBCS was supported by the Hungarian Research Foundation (DFG Do761/15-1), a grant from the Friends of Hannover Medical School and by the Rudolf Bartling Foundation. The HUBCS was supported by German Research Foundation (DFG Do761/15-1), a grant from the German Federal Ministry of Education and Research (RUS08/01). BM was supported by the Russian Foundation for Basic Research (grants 17-44-020498 and 17-29-00614). DP was supported by the Russian Foundation for Basic Research (grant 18-29-09125A) and the program of the President of the Russian Federation (2020-220-08-2197), and the study was performed as part of the assignment of the Ministry of Science and Higher Education of the Russian Federation (NAAAA-I-16620350022-1). ICICLE was supported by Breast Cancer Now, CRUK and Biomedical Research Centre at Guy’s and St Thomas’ NHS Foundation Trust and King’s College London. Financial support for KARBC was provided through the Regional agreement on medical training and clinical research (ALTT) between Stockholm County Council and Karolinska Institutet, the Swedish Cancer Society, The Gustav V Jubilee foundation and Bert von Kantonis foundation. The KARMA study was supported by Märts and Hans Rausing Initiative Against Breast Cancer. The KCPC was financially supported by the special Government Funding (VTR) of Kuopio University Hospital grants, Cancer Fund of North Savo, the Finnish Cancer Organizations South of Finland, the strategic fund of the University Hospital of Eastern Finland. KOnFab is supported by a grant from the National Breast Cancer Foundation, and previously by the NHMRC, the Queensland Cancer Fund, the Cancer Council of New South Wales, Victoria, Tasmania and South Australia, and the Cancer Foundation of Western Australia. Financial support for the AOCS was provided by the United States Army Medical Research and Materiel Command (DAMD17-01- 0297-0045), Cancer Council Victoria, Australia; Cancer Council South Australia, The Cancer Foundation of Western Australia, The Cancer Council Tasmania and the NHMRC (400413, 400281 and 196000). GCT and PW were supported by the NHMRC. RR was a Cancer Institute NSW Clinical Research Fellow. LMBC was supported by the ‘Stichting tegen Kanker’. DL was supported by the FWO. The MBACS study is funded by the Research Centre for Genetic Engineering and Biotechnology ‘Georgi D. Efremov’, MAIA. The MARIE study was supported by the National Foundation ‘University of Helsinki’ (grant 553/2009, 108419, 110826 and 110828), the Hamburg Cancer Society, the German Cancer Research Center (DKFZ) and the BMBF Germany (01KX0402). MBCSG was supported by grants from the AIRC. The MCBCS was supported by the NIH (grants R01CA253187, R01CA192393, R01CA116167 and R01CA176785), an NIH Specialized Program of Research Excellence (SPORE) in Breast Cancer (PS01CA162100) and the Breast Cancer Research Foundation. The Melbourne Collaborative Cohort Study (MCCS) cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further augmented by the Australian National Health and Medical Research Council (grants 209057, 396414 and 107483) and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry and the Australian Cancer Registry, including the Victorian Neonatal Death Registry. The CAN was supported by the Australian Cancer Database. The MEC was supported by NIH (grants CA63464, CA54249, CA087585, CA132839 and CA164973). The MISS study was supported by funding from ERC-2011-294576 Advanced grant, Swedish Cancer Society CAN 2018/675, Swedish Research Council, local hospital funds, Berta Kampford Foundation FBKS 2021-19, Gunnar Nilsson. MSKCC is supported by grants from the Breast Cancer Research Foundation Robert and Kate Niehaus Clinical Cancer Research Initiative. The work of supported by the German Cancer Research Foundation, the Canadian Institutes of Health Research for the ‘CHIR Team in Familial Risks of Breast Cancer’ programme (grant number CRN-B7521) and the Ministry of Economic Development, Innovation and Export Trade (grant number PSR-SIRI-701). The NBCS has received funding from the K.G. Jebsen Centre for Breast Cancer Research; The Research Council of Norway (grant 193387/V50 to A-B-L-D and VNK, and grant 193388/10 to the Scottish Breast Cancer Research Unit); the Danish Cancer Society; the Icelandic Cancer Society and the Norwegian Cancer Society (to A-B-L-D and VNK). The NBHS was supported by NIH (grant R01CA100374). Biological sample preparation was conducted by the Survey and Biospecimen Shared Resource, Levi H, et al. J Med Genet 2023;60:1186–1197. doi:10.1136/jmg-2023-109185