Supplementary Methods

Health databases

For identifying participants who had breast cancer, we used two administrative health databases (AHD): 1) the MED-ÉCHO AHD: this database contains all the Quebec Health Insurance Board (RAMQ) diagnoses, hospitalizations and physician claims of insured patients (about 98% of Quebec residents [1]), excluding private healthcare; in the case of cancers, all patients are treated in the public sector. Data were available from January 1st, 1998 to March 31st, 2016. Dates of death were also retrieved from the RAMQ; 2) the Quebec Breast Cancer Registry: it contains information about the Quebec Breast Cancer Screening Program, such as mammograms' results and breast cancers histological confirmation. Data were available from May 15th, 1998 to December 31st, 2017.

References

1 RAMQ. Table PA.01 - Nombre de personnes inscrites et admissibles au régime d’assurance maladie du Québec selon le sexe, le groupe d’âge et la région sociosanitaire. 2017.https://www4.prod.ramq.gouv.qc.ca/IST/CD/CDF_DifsnInfoStats/CDF1_CnsulInfoStatsCNC_iut/DifsnInfoStats.aspx?ETAPE_COUR=3&IdPatronRapp=8&Annee=2017&Per=0&LANGUE=en-CA (accessed 25 Nov 2019).

Genetic data

Genotypes were included in the CaG database and were obtained from hybridation upon three different chips: Illumina Omni 2.5M (7.7% of the participants), Affymetrix Axiom UK biobank (8.2%) and Illumina Infinium Global Screening Array (84.1%). A quality control (QC) was made before the imputation (detailed pipeline can be found at www.cartagene.qc.ca/info-genetic-data): 1) QC sample: for replicated samples, samples with the lowest call rates were removed. Sample with a call rate below 95% were removed. Samples pairs with an identity by state (IBS) higher than 0.20 and similar to at least 50% of the whole set were removed. Then, for pair of samples with an IBS higher than 0.85, when the correct sample could not be identified with certainty, both samples of the pair were removed. Samples with discrepancy between sex chromosome genotypes and reported gender were removed. 2) QC SNP: SNPs with a call rate lower than 95% or deviating from Hardy–Weinberg equilibrium (with a 10-6 threshold) were removed.

For the imputation, data were prepared using the Will Rayner toolbox (www.well.ox.ac.uk/~wrayner/tools/) with the Haplotype Reference Consortium (HRC) as reference panel [1]. To impute missing SNPs of our cohort, we used the Michigan Imputation Server with the Minimac4 algorithm [2], with separate chromosomes and chips. Imputation reference panel was the HRC r1.1 2016 European population, and the phasing was made with Eagle v2.4 [3]. A total of 39,131,578 SNPs were retrieved.

After imputation and after merging chromosomes, we used men and women to perform a sample QC based on the Anderson *et al.* protocole [4]: samples with a call rate lower than 95% and an heterozygosity higher than 3 standard deviation were removed. After LD pruning (window size: 50kb; step size: 5 variants; pairwise r^2 threshold: 0.2), for pair of participants with an IBS higher than 0.1875, the sample with the lowest call rate was removed. To remove samples with divergent ancestries, we used the two first principal components with the HapMap phase III reference panel. As we would like to have all SNPs available for calculating PRS, we did not perform an additional SNPs QC. QC process was performed using PLINK v1.90b6.2 and v2.00a2LM 64-bit ([5,6]; URL: pngu.mgh.harvard.edu/purcell/plink/).

References

1. the Haplotype Reference Consortium, McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016 août;48:1279.
2. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284–7.
3. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48(11):1443–8.
4. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010 Sep;5(9):1564–73.

5. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.
6. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015 Dec;4(1):7.

Absolute risk of breast cancer

The absolute risk of breast cancer over an established period $[t_0,t_1]$ (five years in this study) is the probability that a woman who is free of a breast cancer at age t_0 and has a risk score S will be diagnosed with breast cancer over the period $[t_0,t_1]$.

Under the assumption of a multiplicative proportional hazard model (or Cox model), this latter conditional probability (denoted $AR(t_0,t_1;S)$) can be written such as:

$$AR\left(t_0,t_1;S\right)=\int_{t_0}^{t_1}\lambda_0\left(t\right)e^{S}\exp\left[-\int_{t_{\square}}^{t_0}\lambda_0\left(u\right)e^{S}+\gamma\left(u\right)du\right]dt$$

where $\lambda_0(t)$ and $\gamma(t)$ are the baseline age-specific hazard rate for breast cancer and the age-specific mortality hazard rate from other causes (competing risks), respectively. In practice, the absolute risk is computed using piece-wise constant hazard rates.

These baseline hazard rates are calculated using marginal (or composite) hazard rates obtained from registries, together with either the attributable hazard function or the risk factor distribution.

In this work, the timescale of the analyses was age of an individual so that t_0 was the age of a woman at entry into the cohort and t_1 was the age five years later.

For the IBIS model, the baseline age-specific hazard rate for breast cancer is replaced by a hazard rate estimate obtained from the segregation model conditionally on the woman’s family history.

Variables extraction and coding

Age at inclusion was calculated using the birthdate. We retrieved from the CARTaGENE questionnaire the first menstrual period, first live birth, number of first-degree relatives with breast cancer, ethnicity, menopause occurrence and age at menopause, height, weight, hormonal replacement therapy (HRT) use, length of HRT and last HRT use. If first menstrual period occurred after first live birth, both were considered as missing. We retrieved from the Quebec Breast Cancer Registry the previous breast biopsy and the number of biopsy with hyperplasia, atypical hyperplasia and lobular carcinoma *in situ*. We retrieved from the RAMQ the occurrence and age of ovary cancers.

How the variables were coded for the IBIS model can be found online (https://ems-trials.org/riskevaluator/), in the Documentation section, file "Risk program input file format (v6-8)"