Assessment of Breast Cancer Risk Factors Reveals Subtype Heterogeneity.

Johanna Holm, Louise Eriksson, Alexander Ploner, Mikael Eriksson, Mattias Rantalainen, Jingmei Li, Per Hall and Kamila Czene.

Assessment of Breast Cancer Risk Factors Reveals Subtype Heterogeneity

Author list:

Johanna Holm[1], Louise Eriksson[1,2], Alexander Ploner[1], Mikael Eriksson[1], Mattias Rantalainen[1], Jingmei Li[1], Per Hall[1,3], Kamila Czene[1]

Affiliations:

1 = Department of Medical Epidemiology & Biostatistics, Karolinska Institutet, Solna, Sweden

2 = Department of Oncology and Pathology, Karolinska Institutet and University Hospital, Stockholm, Sweden.

3 = Department of Oncology, Södersjukhuset, Stockholm, Sweden.

Running title: Subtype heterogeneity for breast cancer risk factors

Keywords: breast cancer, molecular subtypes, risk prediction, risk factors, PAM50

Funding:
The KARMA study was funded by the Märit and Hans Rausing's Initiative Against Breast Cancer. The LIBRO-1 study was funded by the Cancer Risk Prediction Center (CRisP), a Linneus Centre (Contract ID 70867902) financed by the Swedish Research Council. This work was further supported by the Swedish Research Council (Grant No. 2014-2271), Swedish Cancer Society (Grants No. CAN 2013/469 and CAN 2016/684), Stockholm County Council (Grant No. 20150642) and Cancer Society in Stockholm (Grant No. 141092).

Corresponding author: Johanna Holm. Email: Johanna.holm@ki.se Address: Karolinska Institutet, Department of Medical Epidemiology & Biostatistics, Nobels väg 12 A, SE-171 77 Solna. Phone: +46-707-911-847 Fax: +46-8-31 49 75

Conflict of interest statement: The authors declare no conflict of interest

This manuscript describes original work and is not under consideration by any other journal. All authors approved the manuscript and this submission.

Word count: Main text excluding references: 3,909, Abstract: 228

Files: 4 tables, 1 figure

Abstract

Subtype heterogeneity for breast cancer risk factors has been suspected, potentially reflecting etiological differences and implicating risk prediction. Reports are conflicting regarding presence of heterogeneity for many exposures.

To examine subtype heterogeneity across known breast cancer risk factors, we conducted a case-control analysis of 2,632 breast cancers and 15,945 controls in Sweden. Molecular subtype was predicted from pathology-record derived immunohistochemistry markers by a classifier trained on PAM50 subtyping. Multinomial logistic regression estimated separate odds ratios for each subtype by the exposures parity, age at first birth, breastfeeding, menarche, HRT use, somatotype at age 18, benign breast disease, mammographic density, polygenic risk score, family history of breast cancer and BRCA mutations.

We found clear subtype heterogeneity for genetic factors and breastfeeding. The polygenic risk score was associated with risk of all subtypes except for the basal-like (p heterogeneity < 0.0001). Parous women who never breastfed were at higher risk of basal-like subtype (OR 4.17; 95% CI 1.89 to 9.21) compared to both nulliparous (reference) and breastfeeding women. Breastfeeding was not associated with risk of HER2-overexpressing type, but protective for all other subtypes.

The observed heterogeneity in risk of distinct breast cancer subtypes for germline variants supports heterogeneity in etiology and has implications for their use in risk prediction. The increased risk of basal-like subtype among women who never breastfed merits more research into potential causal mechanisms and confounders.

Introduction

Breast cancer is a molecularly diverse disease. At least four subtypes have been robustly established following gene expression based characterization in the early 2000s[1,2]. These subtypes (Luminal A, luminal B, HER2-overexpressing and basal-like) behave differently in terms of age at onset and prognosis, but the question remains to what extent they also represent etiologically distinct cancers and differ in risk factors. Recently analysis of recurrent and metastasizing patient samples has shown that breast cancer can drift in molecular subtype throughout disease progression [3–6]. However, it has repeatedly been shown that in situ cancers display the full spectrum of molecular subtypes before any signs of invasiveness[7–11], pointing towards early determination of the cancer's nature. Presence of subtype heterogeneity for breast cancer risk factors could indicate separate etiology for one or several subtypes of breast cancer, and would also implicate clinical efforts in risk prediction and prevention.

Studies addressing subtype heterogeneity of risk factors have so far employed various immunohistochemistry (IHC) marker combinations as proxy for the gene-expression based classifications, which limits precision of findings. Due to the majority of cases in study populations being luminal, one could expect known risk factors to be biased towards predicting luminal breast cancers. A few consistent observations have been made: compared to the most prevalent subtype luminal A, basal-like breast cancer is associated with younger age at onset, presence of BRCA1 mutations and recent African heritage[3]. However, results are less consistent for lifestyle and reproductive risk factors such as HRT use, parity, age at first birth, and age at menarche[4]. Most consistent is a stronger protective effect of breastfeeding on the basal-like subtype among parous women[5-8]. Intriguingly, although parity has been consistently associated with decreased risk of luminal subtypes, results vary from a decreased to an increased risk of the basal-like subtype[4]. Very little is known about risk factors for the luminal B and HER2-overexpressing types, related to difficulties in finding consistent IHC

proxies to represent them[4]. Moreover, because the relative proportions of subtypes vary across ethnicities, genetic variation has been put forward as explaining variation in subtype incidence[9,10] but could also be explained by environmental, lifestyle and reproductive differences[11-14]. Germline genetic risk factors beyond BRCA mutations should therefore also be included in studies of subtype heterogeneity.

We aimed to assess the associations between reproductive, genetic and hormonal exposures and breast cancer subtypes, addressing subtype heterogeneity of risk factors in a Swedish material. To improve precision in classification of the outcome, we made use of a subset of our data where PAM50-subtyping was available and predicted subtype on the full dataset from these, as well as used a previously established IHC proxy to define subtypes to assess the robustness of any findings.

Methods

Setting

This is a case-control study based on two Swedish breast cancer cohorts. Ethical approvals were granted from the regional ethical vetting board and all participants gave written informed consent.

Participants

Participants were recruited from the the KARolinska MAmmography Project for Risk Prediction of Breast Cancer (KARMA) and Libro-1 study cohorts[15-16]. KARMA is a prospective cohort study of 70,877 women with or without breast cancer, recruited in 2011 to 2013 from four mammography units situated in Skåne county and Stockholm conducting both population-based mammography screening and clinical mammography. Libro1 is a case-only, population-based cohort consisting of 5,715 women diagnosed with breast cancer in Stockholm years 2001 to 2008. Women in both studies answered questionnaires, donated

blood at enrollment, and consented to the retrieval of their mammograms and medical records. Questionnaires and study material were largely similar for both studies as Libro-1 was the pilot study of KARMA.

All primary invasive breast cancer cases from both studies diagnosed 2005 to 2015 were eligible for inclusion (n=4,598). The cutoff at 2005 was chosen as IHC markers HER2 and Ki67 were not stained for, and thus not available in medical records, prior to this year. Exclusion criteria were missed information on either of the immunohistochemistry (IHC) markers ER, PR, HER2 and Ki67 (n=1,265). Controls were randomly selected among breast-cancer free participants of the KARMA study, frequency-matched up to 1:5 to cases on age at enrollment (controls) to age at diagnosis (cases) in 5-year strata using a greedy nearest neighbor algorithm without replacement[17]. The final study sample consisted of 2,632 breast cancers and 15,945 controls. For the analysis of the polygenic risk score, all KARMA women who did not have breast cancer and who had available genotyping (n=5425) were used as controls.

Data on exposures

Data on parity, age at first birth, breastfeeding, hormone replacement therapy (HRT) use, somatotype at age 18, menarche, weight and height at enrollment,BRCA1/2 mutation carriership, country of birth, education and family history of breast cancer was collected from web- and paper study questionnaires answered at study enrollment, with all variables harmonized between the two studies prior to analysis. Recall of HRT use was aided by pictures of HRT brands dispensed in Sweden. Somatotypes were illustrated by 9 pictograms in the questionnaires and women were asked which pictogram most resembled them at age 18. As BRCA mutation status was self-reported and thus more likey to be known for cases, only data on BRCA status from cases were considered for this study.

History of benign breast disease (BBD) was obtained from pathology records. Only BBD diagnoses at least one year prior to breast cancer for cases, or before 2013 for the controls, were included in analysis. Mammograms were collected from radiology departments and available for 90% of the study population. Mammographic density (MD) was measured using an automated method previously described[15,22]. Image pairs from a reading within 4 years prior to diagnosis or study enrollment were selected for cases and controls respectively and absolute MD from the left and right mediolateral oblique view were averaged. Data on SNP markers was obtained from blood donated by participants at enrollment, which had been genotyped on a custom Illumina iSelect Array (iCOGS Array)[23]. Missing genotypes were imputed using 1000 Genomes (phase I integrated variant set release [v3] in National Center for Biotechnology Information build 37 [hg19] coordinates). Polygenic risk scores (PRS) were constructed using 77 breast cancer risk SNPs discovered in large consortia studies. The score for each patient was calculated by summing the number of alleles for each SNP (0, 1 or 2), weighted by the per-allele odds ratios for the minor alleles reported by Mavvadat et al[24]. Two scores were generated, one general score using breast cancer risk odds ratios as weights and the other score using weights from associations to ER negative breast cancer only. Full details on the construction of the PRS were described previously[25].

Data on tumor characteristics (cases only)

Data on molecular markers were retrieved in 2015-2016 from medical and pathology records at treating hospitals. Percent estrogen receptor (ER) and progesterone receptor (PR) staining was dichotomized into positive or negative status with a cutoff at >= 10% as positive during this period. HER2 status was dichotomized into positive or negative according to the Swedish Society of Pathology's guidelines[18]: HER2 was considered negative if protein expression showed 0 or 1+, or was higher with no confirmed gene amplification by fluorescence in-situ hybridization (FISH), and positive if FISH showed gene amplification. Proliferation marker

Ki67 was measured in hotspot regions according to contemporary guidelines[18] and reported as percent staining. Information on tumor invasiveness and prior breast cancer diagnoses was obtained through merges to the Swedish National Cancer Register[19] and the Regional Breast Cancer Quality Register[20] using the unique, Swedish personal identity numbers[21].

Outcome classification

A random forest algorithm was used to construct a subtype classifier using the caret R package26 (v. 6.0.58). As part of the Clinical Sequencing of Cancer in Sweden (Clinseq) project, 237 of the cases had had tumors RNA-sequenced and assigned into PAM-50 molecular subtypes as described previously[26]. The algorithm was trained to predict subtype on the subset of the data with PAM50 subtype information available ('training data', n=237). Binary ER, PR, HER2, continuous Ki67 and age at diagnosis were entered as input and the algorithm run with 5-times repeated 10-fold cross validation to avoid overfitting. Accuracy and kappa values were used to select the best performing algorithm. The resulting best algorithm (hereafter denoted 'classifier') then assigned PAM50 subtype to all remaining cases based on their age, ER,PR, HER2 and Ki67 status.

As a sensitivity analysis, the St Gallen method of using immunohistochemistry markers as a proxy for gene-expression based subtyping[28]was used to assign subtype, with a modified cutoff for Ki67 of 25 % instead of 14% due to the lack of whole-slide % for Ki67. This proxy defines luminal A as ER+/PR+/HER2- and KI67 low, luminal B as either ER+/PR-/HER2-, or ER+/PR+/HER2-, KI67 high, or ER+/HER2+/any PR, any KI67. HER2-overexpressing is defined as ER-/PR -/HER2 +, and basal-like defined as ER-/PR-/HER2- (triple negative).

To assess the performance of our classifier and the St Gallen proxy, we obtained accuracy and kappa statistics for both methods as compared against gene-expression based PAM50 subtyping, by resampling from the 237 observations with PAM50 data (function "resamples" from the caret package in R). Resampled values were necessary in order to avoid over-fitted,

over-optimistic statistics for the classifier which had been trained on the same data. Confusion Matrices of actual verses predicted subtype were tabulated for both methods, in the training dataset.

Statistical analysis

Multivariable regression

In multivariable regression analysis, breast tumors of different subtypes were considered as separate outcomes and their respective risks were modelled relative to healthy controls via multinomial regression. For simplicity, we refer to the resulting relative risk ratios as odds ratios. Heterogeneity in odds ratios was formally assessed with a global Wald test, testing the null hypothesis that the risk associated with the exposure was the same across all subtypes. Additionally, multinomial logistic regression was performed in a case-only design using luminal A cases as the reference group. For the sake of comparison to subtype-specific odds ratios, odds ratios comparing all breast cancer cases to controls were obtained using unconditional logistic regression. All analyses were minimally adjusted for age (matching variable), education level (<10 years, 10 to 12 years, university, other) and country of birth (binary, Sweden/other), further covariates were included as potential confounders in the models based on subject matter knowledge. The sets of covariates included in each of the fully adjusted models are stated in the respective tables. Only case-only analysis was performed pertaining to exposure BRCA mutation status.

Exposure parameterizations

First-degree family history of breast cancer was modelled as a binary variable, defined as having a mother or sister with breast cancer, yes/no. Continuous variables for polygenic risk scores were scaled prior to modeling and modelled per 1-standard deviation (SD) increase. Both PRS's were additionally modelled as categorical variables, cut into quartiles of the scores. Parity was modelled as a categorical (0, 1 to 2, >2 children) and continuous variable.

Age at first birth was dichotomized into < 30 and >= 30 years of age, restricting analysis to parous women. Breastfeeding was categorized into 0, >0 to 1.5 years and >1.5 years, after summarizing the total length of breastfeeding across all children. Breastfeeding was also assessed as a composite variable including parity, by using nulliparous women as reference group. HRT use was modelled as 'Ever'/'Never' use. Women who reported use of locally administered HRTexclusively were coded as never users. Somatotype was modelled as a semi-continuous variable by assigning integers to each somatotype 1-9 from lowest to highest adiposity. Menarche was modelled per year's delay as a continuous variable. Mammographic density was scaled prior to modelling and assessed per-1 SD increase. Benign breast disease was separated into non-proliferative and proliferative, non-atypical lesions and modeled as binary variables of 'ever'/'never' diagnosed. Atypical proliferative lesions were not included in analysis due to insufficient numbers.

All statistical tests were two-sided with a pre-determined cutoff for statistical significance at alpha = 0.05. Software R[29] v.3.2.2 was used for all statistical analysis.

Results

Outcome classification

The average resampled accuracy and kappa values were 0.73 and 0.55 for the random forest classifier, and 0.64 and 0.46 for the St Gallen IHC proxy. For all main tables, the random forest classifer was used for outcome classification. The mode of IHC marker combinations for cases classified as luminal A or B was ER+/PR+/HER2-, observed for 80% and 54% respectively, for HER2-overexpressing ER-/PR-/HER2+, observed for 45%, and for basal-like triple negative for all markers, observed for 85%. Average percentage Ki67 staining was 14% among luminal A tumors, 46% among luminal B tumors, 36% among HER2-overexpressing tumors and 69% among basal-like tumors (Table 1). Confusion matrixes of true vs. predicted

subtypes revealed that both the classifier and the St Gallen IHC proxy were good at capturing luminal A and basal-like status, but performed worse for luminal B and HER2-overexpressing tumors (Supplementary Table 1).

Descriptive cross-tabulations

Table 1 shows crude descriptive contingency tables of cases verses controls, and cases by subtypes for adjusting variables and selected exposures of interest. Cases tended to be more highly educated, more often born abroad, had a higher frequency of family history of breast cancer and had breastfed less than controls. Within subtypes, variations in age, BRCA mutations and breastfeeding were observed. Although age ranges were similar across subtypes, luminal A cases were on average older than the other categories (59 years) and basal-like cases were youngest at diagnosis (52 years) (Table 1).

Adjusted case-control odds ratios

Multivariable regression analysis of genetic background risk factors for each subtype is shown in Table 2. The general 77-SNP PRS was associated with all subtypes except for the basal-like subtype, and most strongly associated with the luminal A subtype. The Wald test showed significant heterogeneity across subtype odds ratios ($p_{heterogeneity}$ <0.0001). In contrast to the general PRS, the PRS weighted on associations to ER negative breast cancer yielded no evidence of heterogeneity of effect across subtypes ($p_{heterogeneity}$ = 0.43, Table 2). There was no evidence of heterogeneity for first-degree family history across breast cancer subtypes (Table 2).

Multivariable regression analysis of reproductive factors by subtypes is shown in Table 3. There was no statistical evidence of heterogeneity between subtypes for parity, however judging from point estimates, parity had a protective effect on all subtypes except for basal-like subtype. Similarly, in analysis of parous women, no heterogeneity was observed for age at first birth (Table 3). The effects of breastfeeding were heterogeneous across subtypes

($p_{heterogeneity}$ = 0.01). With nulliparous women as reference group, parous women who never breastfed had an increased risk of basal-like breast cancer (OR 4.17; 95 % CI 1.89 to 9.21). Breastfeeding returned the risk of basal-like breast cancer to that of nulliparous women (OR breastfeeding <1.5 years 1.02; 95 % CI 0.59 to 1.76), OR breastfeeding >= 1.5 years 0.81; 95 % CI 0.43 to 1.60) (Table 3). In contrast, the risk of developing luminal A breast cancer was no different from nulliparous women for parous women never breastfeeding (OR 1.01; 95 % CI 0.74 to 1.39), but a protective effect was seen for breastfeeding (breastfeeding <1.5 years, OR 0.69; 95 % CI 0.59 to 0.82), breastfeeding >= 1.5 years, OR 0.63; 95 % CI 0.52 to 0.76). Luminal B showed point estimates similar to those of luminal A, whereas HER 2-overexpressing type was unaffected by breastfeeding (Table 3).

Multivariable analysis of lifestyle and non-reproductive hormonal factors by subtypes is shown in Table 4. Ever use of hormone replacement therapy (HRT) showed heterogeneity ($p_{heterogeneity}$ = 0.05), with increased risk for the luminal A (Ever HRT use, OR 1.43, 95 % CI 1.28 to 1.61) but no effect among the other subtypes (Table 4). Point estimates for age at menarche, somatotype at age 18 and benign breast disease also suggested differences by subtype but no statistically significant heterogeneity was observed. Still, for age at menarche, a protective effect of increasing age was observed for the luminal A and B subtypes alone. Increasingly endomorph somatotype at age 18 was associated with a protective effect for the luminal and HER2-subtypes whereas basal-like appeared null-associated. Proliferative non-atypical lesions were associated with luminal A cancers but showed similar estimates for other subtypes except luminal B, whereas non-proliferative lesions were null-associated with all subtypes except for a possible increase for the HER2-subtype. There was no evidence of subtype heterogeneity for mammographic density, with an increased risk for all subtypes of disease with increasing density. (Table 4).

Summary of findings for risk factors that displayed subtype heterogeneity are shown graphically as forest plots in figure 1, separately by subtype. Luminal A and B showed very similar estimates, whereas the basal-like subtype displayed distinct features (Figure 1).

All analyses were repeated within a case-only design with luminal A as reference, yielding the same conclusions regarding heterogeneity for PRS, breastfeeding and HRT use as the case-control analysis (Supplementary Tables 2-4). Case-only analysis of self-reported BRCA mutations among cases showed that basal-like tumors had higher prevalence of BRCA mutations than other subtypes, with an odds ratio of 11.31 (95% CI 5.37 to 23.07) relative to luminal A tumors (Supplementary Table 2).

Sensitivity analysis

All findings were replicated albeit attenuated when using the St Gallen IHC proxy to define subtypes, except for the heterogeneity found for HRT use, which was not observed (Supplementary Tables 5-7).

Discussion

Analysis of 2,632 breast cancer cases revealed evidence of subtype heterogeneity for three categories of risk factors: Genetic susceptibility, HRT use and breastfeeding. The 77-SNP PRS was exclusively associated with risk of non-basal like subtypes, with the largest effect size for luminal A breast cancers. HRT use was associated with risk of the luminal A subtype only. Compared to nulliparous women, never breastfeeding was associated with an increased risk of basal-like breast cancer but not with risk of the other subtypes. Among parous women, breastfeeding was protective for the luminal A, B and basal-like subtypes but null-associated with the HER2-overexpressing subtype. Luminal A and B breast cancers were very similar in associations with most risk factors.

Both germline BRCA mutations (Supplementary Table 2) and PRS (Table 2) were differentially associated with subtypes, suggesting that in addition to the previously observed heterogeneity for BRCA1 mutations[1,2], inherited low-risk variants could also differentially increase risk of specific breast cancer subtypes. BRCA1 mutations have been shown experimentally to result in accumulation of undifferentiated luminal progenitor cells[30-33], the suspected cell-of-origin of basal –like breast cancer[34,35], but it is not known how or if low-risk SNP's could represent an etiological difference between subtypes. Individual SNPs have been found to be associated with either subtype or ER status[36-39], supporting our finding for the PRS. The ER-negative weighted PRS was associated with basal-like breast cancer in our study thus showing potential as a complement or replacement to the overall PRS score for identification of women at risk of this aggressive disease. Collectively, these observations also indicate a role of germline variants beyond BRCA in distinct etiology of molecular subtypes which should be further investigated.

Our results confirm previous reports of the largest protective effect of breastfeeding on risk of basal-like breast cancer[5-8]. We additionally show that the protective effect for basal-like breast cancer stems from never-breastfeeding parous women having higher risk of the basal-like subtype than both nulliparous and breastfeeding women. The reason behind this increased risk, should it be causal, is not known. As basal-like cancers are thought to originate from undifferentiated luminal progenitor cells[33,34], the association may be related to higher numbers of progenitor cells in the absence of breastfeeding. Fully differentiated type 4 lobules do not form until the end of pregnancy and during lactation under the influence of prolactin[40], and prolactin, released throughout lactation[41] , has recently been identified as a central promotor of luminal progenitor cell maturation in vitro[42]. This hypothesis would agree with observations of BRCA1 mutations resulting in accumulation of luminal progenitor cells[30-33]and the high proportion of basal-like breast cancer among BRCA1 mutation carriers.

Epidemiological studies have additionally shown that BRCA1 carriers who breastfeed are more protected from developing breast cancer[43,44]. However, future studies should definitely consider possible confounding by lifestyle factors, preferably including data on reasons for not breastfeeding. The lack of association with breastfeeding for the HER2-overexpressing subtype distinguishes it from the other subtypes and merits further investigation, preferably with gene-expression based subtype definitions.

Point estimates for somatotype, menarche, age at first birth, parity and benign breast disease suggested heterogeneity but were not statistically different. We saw a protective effect of parity on all subtypes except for a null association of risk of basal-like breast cancer with parity, after adjusting for age at first birth and breastfeeding. This is in line with some of the published literature, whereas others have reported increased risks of the basal-like subtype with increasing parity. Differences may be partly due to underlying variations in prevalence of breastfeeding.

We found no evidence of subtype heterogeneity for mammographic density, or with ER-negative weighted PRS. These are important messages for prevention efforts using such exposures in risk prediction, as they would be expected to identify women at risk of breast cancer independent of subtype. Mammographic density is an especially promising tool to predict risk, as it is also among the stronger risk factors for the disease.

All results were robust albeit sometimes attenuated in sensitivity analysis using the St Gallen IHC proxy to define subtypes (Supplementary Tables 5-7). The only exception was seen for HRT use, which did not display subtype heterogeneity using the St Gallen IHC proxy. Future studies should ideally address this question using true gene-expression based subtype data, but our results suggest heterogeneity in subtype risk for HRT use.

A limitation of our study is that we only had true PAM50 subtype information for 237 of the cases and predicted subtype for the rest. As the assigned subtypes were largely latent for true subtype status, model estimates should be interpreted with caution. Although the overall accuracy of our classifier was higher compared to the St Gallen IHC proxy for defining subtypes, it still performed poorly in classifying luminal B and HER2 types and our results should be interpreted in light of this.

The retrospective nature of this study introduces some elements of caution in interpretation. Cases may recollect exposures differently than controls, but there is little reason to believe such differences would vary greatly by subtype. Controls were only available from one of the cohorts, which potentially could introduce bias. Reassuring is that the case-control odds ratios for generic breast cancer exhibited the expected strengths and directions of associations .Moreover, case-only analysis yielded the same conclusions regarding heterogeneity as case-control analysis. Strengths of the current work include the large cohort, the use of a subtype classifier with improved accuracy compared to a previously used IHC-based method, and the availability of genetic, lifestyle and reproductive exposures comprehensively assessed in the same study.

In conclusion, both rare and common inherited gene variants displayed subtype heterogeneity primarily between basal-like and non-basal like subtypes, suggesting separate etiology and implications for risk prediction. We additionally found subtype heterogeneity by breastfeeding status. Relative to nulliparous women, women who did not breastfeed postpartum were exclusively at an increased risk of basal-like breast cancer. Future research is needed to confirm or refute this finding, subsequently addressing reasons behind the observed increase in risk of the highly aggressive subtype basal-like breast cancer. Mammographic density did not display subtype heterogeneity, which is assuring for usage of mammographic density in risk prediction and prevention. Finally, although we did improve in overall

accuracy over available subtype surrogacy classifiers, more work remains to improve on accuracy for defining luminal B and HER2 subtypes by IHC markers for their future use in research.

Acknowledgments

The authors wish to thank all participants and staff in the Libro1 and Karma studies for their effort and time, and research nurses Agneta Lönn and Cecilia Arnesson with colleagues for assistance with data collection in Stockholm and Lund respectively.

References

1. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747-752.

2. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci*. 2003;100(14):8418-8423.

3. Toft DJ, Cryns VL. Minireview: Basal-Like Breast Cancer: From Molecular Profiles to Targeted Therapies. *Mol Endocrinol*. 2011;25(2):199-211.

4. Barnard ME, Boeke CE, Tamimi RM. Established breast cancer risk factors and risk of intrinsic tumor subtypes. *Biochim Biophys Acta - Rev Cancer*. 2015;1856(1):73-85.

5. Millikan RC, Newman B, Tse C-K, Moorman PG, Conway K, Dressler LG, et al. Epidemiology of basal-like breast cancer. *Breast Cancer Res Treat*. 2008;109(1):123-139.

6. Kwan ML, Kushi LH, Weltzien E, Maring B, Kutner SE, Fulton RS et al. Epidemiology of breast cancer subtypes in two prospective cohort studies of breast cancer survivors. *Breast Cancer Res*. 2009;11(3):R31.

7. Redondo CM, Gago-Domínguez M, Ponte SM, Castelo ME, Jiang X, García AA et al. Breast feeding, parity and breast cancer subtypes in a Spanish cohort. *PLoS One*. 2012;7(7).

8. Shinde SS, Forman MR, Kuerer HM, Yan K, Peintinger F, Hunt KK et al. Higher parity and shorter breastfeeding duration: association with triple-negative phenotype of breast cancer. *Cancer*. 2010;116(21):4933-4943.

9. Brewster AM, Chavez-MacGregor M, Brown P. Epidemiology, biology, and treatment of triple-negative breast cancer in women of African ancestry. *Lancet Oncol*. 2014;15(13):e625-e634.

10. Boyle P. Triple-negative breast cancer: Epidemiological considerations and recommendations. *Ann Oncol*. 2012;23(SUPPL. 6).

11. Palmer JR, Viscidi E, Troester MA, Hong CC, Schedin P, Bethea TN et al. Parity, lactation, and breast cancer subtypes in African American women: results from the AMBER Consortium. *J Natl Cancer Inst*. 2014;106(10).

12. Trivers KF, Lund MJ, Porter PL, Liff JM, Flagg EW, Coates RJ et al. The epidemiology of triple-negative breast cancer, including race. *Cancer Causes Control*. 2009;20(7):1071-1082.

13. Ambrosone CB, Zirpoli G, Ruszczyk M, Shankar J, Hong CC, McIlwain D et al. Parity and breastfeeding among African-American women: Differential effects on breast cancer risk by estrogen receptor status in the Women's Circle of Health Study. *Cancer Causes Control*. 2014;25(2):259-265.

14. Bandera E V., Chandran U, Hong C-C, Troester MA, Bethea TN, Adams-Campbell LL, et al. Obesity, body fat distribution, and risk of breast cancer subtypes in African American women participating in the AMBER Consortium. *Breast Cancer Res Treat*. 2015;150(3):655-666.

15. Holm J, Humphreys K, Li J, Ploner A, Cheddad A, Eriksson M et al. Risk factors and tumor characteristics of interval cancers by mammographic density. *J Clin Oncol*. 2015;33(9):1030-1037.

16. Gabrielsson M, Eriksson M, Hammarström M, Borgquist S, Leifland K, Czene K et. al. Cohort profile: The Karolinska Mammography Project for Risk Prediction of Breast Cancer (KARMA). Accepted for publication in International Journal of Epidemiology 29-Nov-2016.

17. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat Med*. 2014;33(6):1057-1069.

18. Grabau D. *KVAST Dokument Brösttumörer*. Vol Edition 3.; 2014. http://svfp.se/files/docs/kvast/brostpatologi/Brostcancerdokument_godkant_maj_2014.pdf.

19. Barlow L, Westergren K, Holmberg L, Talbäck M. The completeness of the Swedish Cancer Register: a sample survey for year 1998. *Acta Oncol*. 2009;48(1):27-33.

20. Emilsson L, Lindahl B, Köster M, Lambe M, Ludvigsson JF. Review of 103 Swedish Healthcare Quality Registries. *J Intern Med*. 2015;277(1):94-136.

21. Ludvigsson JF, Otterblad-Olausson P, Pettersson BU, Ekbom A. The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research. *Eur J Epidemiol*. 2009;24(11):659-667.

22. Li J, Szekely L, Eriksson L, Heddson B, Sundbom A, Czene K et al. High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer. *Breast Cancer Res*. 2012;14(4):R114.

23. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet*. 2013;45(4):353-361, 361e1-e2.

24. Mavaddat N, Pharoah PDP, Michailidou K, Tyrer J, Brook MN, Bolla MK et al. Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst*. 2015;107(5).

25. Li J, Holm J, Bergh J, Eriksson M, Darabi H, Lindström LS et al. Breast cancer genetic risk profile is differentially associated with interval and screen-detected breast cancers. *Ann Oncol*. 2015;26(3):517-522.

26. Kuhn M. caret: Classification and Regression Training. *R Packag version 60-58*. 2015. http://cran.r-project.org/package=caret.

27. Wang M, Klevebring D, Lindberg J, Czene K, Grönberg H, Rantalainen M. Determining breast cancer histological grade from RNA-sequencing data. *Breast Cancer Res*. 2016;18(1):48.

28. Goldhirsch A, Winer EP, Coates AS, Gelber RD, Piccart-Gebhart M, Thürlimann B et al. Personalizing the treatment of women with early breast cancer: Highlights of the st

gallen international expert consensus on the primary therapy of early breast Cancer 2013. *Ann Oncol*. 2013;24(9):2206-2223.

29. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/. Published 2015.

30. Foulkes WD. BRCA1 functions as a breast stem cell regulator. *J Med Genet*. 2004;41(1):1-5.

31. Burga LN, Tung NM, Troyan SL, Bostina M, Konstantinopoulos PA, Fountzilas H et al. Altered Proliferation and Differentiation Properties of Primary Mammary Epithelial Cells from BRCA1 Mutation Carriers. *Cancer Res*. 2009;69(4):1273-1278.

32. Furuta S, Jiang X, Gu B, Cheng E, Chen P-L, Lee W-H. Depletion of BRCA1 impairs differentiation but enhances proliferation of mammary epithelial cells. *Proc Natl Acad Sci U S A*. 2005;102(26):9176-9181. 53.

33. Liu S, Ginestier C, Charafe-Jauffret E, Foco H, Kleer CG, Merajver SD et al. BRCA1 regulates human mammary stem/progenitor cell fate. *Proc Natl Acad Sci*. 2008;105(5):1680-1685.

34. Molyneux G, Smalley MJ. The Cell of Origin of BRCA1 Mutation-associated Breast Cancer: A Cautionary Tale of Gene Expression Profiling. *J Mammary Gland Biol Neoplasia*. 2011;16(1):51-55.

35. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med*. 2009;15(8):907-913.

36. Figueroa JD, Garcia-Closas M, Humphreys M, Platte R, Hopper JL, Southey MC et al. Associations of common variants at 1p11.2 and 14q24.1 (RAD51l1) with breast cancer risk and heterogeneity by tumor subtype: Findings from the Breast Cancer Association Consortium. *Hum Mol Genet*. 2011;20(23):4693-4706.

37. Liang H, Yang X, Chen L, Li H, Zhu A, Sun M et al. Heterogeneity of Breast Cancer Associations with Common Genetic Variants in FGFR2 according to the Intrinsic Subtypes in Southern Han Chinese Women. *Biomed Res Int*. 2015;2015.

38. O'Brien KM, Cole SR, Engel LS, Bensen JT, Poole C, Herring AH et al. Breast cancer subtypes and previously established genetic risk factors: a bayesian approach. *Cancer Epidemiol Biomarkers Prev*. 2014;23(1):84-97.

39. Purrington KS, Slager S, Eccles D, Yannoukakos D, Fasching PA, Miron P et al. Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. *Carcinogenesis*. 2014;35(5):1012-1019.

40. Russo J, Moral R, Balogh GA, Mailo D, Russo IH. The protective role of pregnancy in breast cancer. *Breast Cancer Res*. 2005;7(3):131-142.

41. Crowley WR. Neuroendocrine Regulation of Lactation and Milk Production. In: *Comprehensive Physiology*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2014:255-291.

42. Liu F, Pawliwec A, Feng Z, Yasruel Z, Lebrun J-J, Ali S. Prolactin/Jak2 directs apical/basal polarization and luminal linage maturation of mammary epithelial cells through regulation of the Erk1/2 pathway. *Stem Cell Res*. 2015;15(2):376-383.

43. Jernström H, Lubinski J, Lynch HT, Ghadirian P, Neuhausen S, Isaacs C et al. Breast-feeding and the risk of breast cancer in BRCA1 and BRCA2 mutation carriers. *J Natl Cancer Inst*. 2004;96(14):1094-1098.

44. Gronwald J, Byrski T, Huzarski T, Cybulski C, Sun P, Tulman A et al. Influence of selected lifestyle factors on breast and ovarian cancer risk in BRCA1 mutation carriers from Poland. *Breast Cancer Res Treat*. 2006;95(2):105-109.

Table 1. Demographics and covariates by case status and breast cancer subtype.

	Controls	Cases	Luminal A	Luminal B	HER2-overexpressing	Basal-like
Age at enrollment, Range	25 - 88	27 - 88	29 - 87	30 - 82	33 - 88	27 - 81
Mean (SD)	58 (9.7)	61 (10.3)	62 (9.7)	60 (11.5)	59 (10.4)	55 (12.4)
Age at diagnosis, Range		25 - 84	26 - 84	28 - 79	28 - 82	25 - 78
Mean (SD)		58 (10.4)	59 (9.9)	57 (11.4)	55 (10.6)	52 (12.3)
Education						
<10 years	2,437 (15%)	307 (12%)	231 (13%)	31 (12%)	23 (8%)	22 (14%)
10-12 years	3,926 (25%)	590 (24%)	419 (23%)	54 (22%)	78 (28%)	39 (26%)
University	6,840 (43%)	1,135 (46%)	811 (45%)	112 (45%)	143 (52%)	69 (45%)
Other	2,681 (17%)	448 (18%)	341 (19%)	53 (21%)	31 (11%)	23 (15%)
Country of birth						
Sweden	14,361 (90%)	2,063 (83%)	1,517 (84%)	205 (82%)	224 (81%)	117 (77%)
Other	1,570 (10%)	429 (17%)	298 (16%)	45 (18%)	50 (19%)	36 (24%)
Mother or sister with breast cancer						
No	13,379 (86%)	1,906 (80%)	1,397 (80%)	182 (76%)	211 (80%)	116 (80%)
Yes	2,083 (14%)	481 (20%)	341 (20%)	57 (24%)	53 (20%)	30 (20%)
Parous						
Yes	14,008 (88%)	2,086 (84%)	1,521 (84%)	200 (81%)	232 (84%)	133 (86%)
No	1,931 (12%)	405 (16%)	292 (16%)	48 (19%)	44 (16%)	21 (14%)
Breastfed[a]						
Yes	13,583 (97%)	1,981 (96%)	1,444(96%)	192 (96%)	226 (98%)	119 (91%)
No	367 (3%)	88 (4%)	64 (4%)	8 (4%)	5 (2%)	11 (9%)
BRCA mutation						
Yes		41 (2%)	17 (1%)	3 (1%)	3 (1%)	18 (15%)
No		2,047 (98%)	1,505 (99%)	203 (99%)	234 (99%)	105 (85%)
IHC markers						
Ki67 mean (SD)		22.6 (20.7)	13.5 (10.1)	45.5 (16.9)	35.8 (17.5)	69.3 (17.7)
ER+ PR+ HER2-		1,668 (64%)	1,372 (80%)	129 (54%)	4 (0%)	0 (0%)
ER+ PR+ HER2+		111 (4%)	33 (2%)	38 (16%)	31 (12%)	0 (0%)
ER+ PR- HER2-		392 (15%)	281 (16%)	57 (24%)	1 (<0.5%)	16 (11%)
ER+ PR- HER2+		69 (3%)	18 (1%)	14 (6%)	31 (12%)	0 (0%)
ER- PR- HER2+		123 (5%)	0 (0%)	0 (0%)	114 (45%)	3 (2%)
ER- PR- HER2-		238 (9%)	15 (1%)	0 (0%)	74 (29%)	119 (85%)
ER- PR+ HER2-		10 (<0.5%)	6 (<0.5%)	0 (0%)	1 (<0.5%)	3 (2%)

a = Among parous women only

Table 2. Risk of breast cancer, overall and by subtype, by genetic background. Odds ratios with 95% confidence intervals for controls as reference group. Adjusted for born in Sweden or not, education level and age.

Exposure		Controls (n, %)	Cases (n, %)	OR (95% CI) Any breast cancer	Luminal A (n, %)	OR (95% CI) Luminal A	Luminal B (n, %)	OR (95% CI) Luminal B	HER2-overexpressing (n, %)	OR (95% CI) HER2	Basal-like (n, %)	OR (95% CI) Basal-like	P heterogeneity
Mother or Sister with Breast Cancer	No (Ref.)	13,379 (86%)	1,906 (80%)	1.00 (ref)	1,397 (80%)	1.00 (ref)	182 (76%)	1.00 (ref)	211 (80%)	1.00 (ref)	116 (79%)	1.00 (ref)	
	Yes	2,083 (14%)	481 (20%)	**1.60 (1.43-1.79)**	341 (20%)	**1.53 (1.35-1.75)**	57 (24%)	**1.99 (1.47-2.69)**	53 (20%)	**1.65 (1.21-2.24)**	30 (21%)	**1.71 (1.14-2.57)**	0.45
Polygenic Risk Score	1st quartile	1,521 (28%)	288(16%)	**0.79 (0.66-0.95)**	190 (15%)	**0.77 (0.62-0.96)**	31 (18%)	0.88 (0.54-1.44)	41 (18%)	0.70 (0.46-1.05)	26 (24%)	1.03 (0.59-1.80)	0.68
	2nd quartile (Ref.)	1,449 (27%)	362 (20%)	1.00 (ref)	244 (19%)	1.00 (ref)	35 (21%)	1.00 (ref)	48 (26%)	1.00 (ref)	25 (23%)	1.00 (ref)	
	3rd quartile	1326(24%)	477 (27%)	**1.49 (1.26-1.76)**	334 (26%)	**1.56 (1.28-1.89)**	51 (30%)	**1.63 (1.05-2.53)**	63 (28%)	1.21 (0.84-1.75)	29 (27%)	1.28 (0.74-2.20)	0.56
	4th quartile	1,129 (21%)	663 (37%)	**2.52 (2.14-2.97)**	520 (40%)	**3.01 (2.50-3.62)**	52 (31%)	**2.06 (1.32-3.20)**	64 (28%)	**1.49 (1.03-2.15)**	27 (25%)	1.43 (0.82-2.48)	**0.0003**
	Linear, per SD increase			**1.61 (1.51-1.71)**		**1.74 (1.63-1.87)**		**1.43 (1.62-1.28)**		**1.36 (1.18-1.55)**		1.15 (0.94-1.40)	**<0.0001**
Polygenic Risk Score Weighted on ER -	1st quartile	1,453 (27%)	357 (20%)	0.87 (0.73-1.03)	257 (20%)	0.85 (0.70-1.04)	41 (24%)	1.11 (0.71-1.75)	40 (18%)	0.75 (0.49-1.14)	19 (18%)	0.94 (0.50-1.77)	0.61
	2nd quartile (Ref.)	1,392(26%)	413 (23%)	1.00 (ref)	303 (24%)	1.00 (ref)	37 (22%)	1.00 (ref)	53 (24%)	1.00 (ref)	20 (19%)	1.00 (ref)	
	3rd quartile	1,336 (25%)	467 (26%)	**1.19 (1.01-1.41)**	337 (26%)	1.17 (0.97-1.41)	41 (24%)	1.17 (0.74-1.84)	57 (25%)	1.13 (0.77-1.66)	32 (30%)	1.67 (0.95-2.93)	0.68
	4th quartile	1,244 (23%)	553 (31%)	**1.53 (1.30-1.78)**	391 (30%)	**1.47 (1.23-1.76)**	50 (30%)	1.54 (0.99-2.39)	76 (34%)	**1.62 (1.13-2.33)**	36 (34%)	**2.03 (1.17-3.53)**	0.71
	Linear, per SD increase			**1.28 (1.21-1.36)**		**1.25 (1.17-1.34)**		**1.21 (1.03-1.42)**		**1.36 (1.19-1.56)**		**1.41 (1.16-1.70)**	0.43

Table 3. Risk of breast cancer overall and by subtype, by reproductive risk factors. Odds ratios with 95% confidence intervals, for controls as reference group.

Exposure		Controls (n, %)	Cases (n, %)	OR (95% CI) Any breast cancer	Luminal A (n, %)	OR (95% CI) Luminal A	Luminal B (n, %)	OR (95% CI) Luminal B	HER2-overexpressing (n, %)	OR (95% CI) HER2	Basal-like (n, %)	OR (95% CI) Basal-like	P heterogeneity
Parity[a]	Nulliparous (ref)	1,931 (12%)	405 (16%)	1.00 (ref)	292 (16%)	1.00 (ref)	48 (19%)	1.00 (ref)	44 (16%)	1.00 (ref)	21 (14%)	1.00 (ref)	
	1-2 children	10,094 (63%)	1,532 (62%)	**0.68 (0.60-0.79)**	1,118 (62%)	**0.68 (0.58-0.80)**	142 (57%)	**0.56 (0.38-0.82)**	173 (63%)	0.70 (0.48-1.04)	99 (64%)	0.97 (0.57-1.66)	0.42
	>2 children	3,914(25%)	554 (22%)	**0.63 (0.55-0.74)**	403 (22%)	**0.62 (0.52-0.74)**	58 (23%)	**0.61 (0.40-0.92)**	59 (21%)	**0.61 (0.40-0.94)**	34 (22%)	0.95 (0.63-1.58)	0.58
	Linear, per child increase			**<0.0001**		**<0.0001**		0.12		**0.04**		0.6	
Age at first birth[b], parous women only	<30 (ref)	9,851 (74%)	1,436 (71%)	1.00 (ref)	1,078 (72%)	1.00 (ref)	138 (69%)	1.00 (ref)	158 (68%)	1.00 (ref)	89 (68%)	1.00 (ref)	
	>= 30	3,448 (26%)	594 (29%)	**1.32 (1.17-1.47)**	419 (28%)	**1.32 (1.16-1.50)**	61 (31%)	**1.42 (1.02-1.47)**	73 (32%)	1.32 (0.97-1.79)	41 (32%)	1.16 (0.77-1.75)	0.91
Breastfeeding[c], parous women only	Ever (ref)	13,583 (97%)	1,981 (96%)	1.00 (ref)	1,444 (96%)	1.00 (ref)	192 (96%)	1.00 (ref)	226 (98%)	1.00 (ref)	119 (92%)	1.00 (ref)	
	Never	367 (3%)	88 (4%)	**1.59 (1.23-2.03)**	64 (4%)	**1.49 (1.12-1.98)**	8 (4%)	1.71 (0.81-3.53)	5 (2%)	0.90 (0.37-2.22)	11 (8%)	**4.20 (2.20-7.99)**	**0.01**
Breastfeeding[c], including all women	Nulliparous (ref)	1,931(12%)	405 (16%)	1.00 (ref)	292 (16%)	1.00 (ref)	48 (19%)	1.00 (ref)	44 (16%)	1.00 (ref)	21 (14%)	1.00 (ref)	
	Parous, never breastfed	367 (2%)	88 (4%)	1.09 (0.82-1.43)	64 (4%)	1.01 (0.74-1.39)	8 (3%)	0.95 (0.43-2.09)	5 (1%)	0.64 (0.24-1.67)	11 (7%)	**4.17 (1.89–9.21)**	**0.005**
	Parous, breastfed >0-1.5 years	9,148 (58%)	1,406 (57%)	**0.70 (0.61-0.80)**	1,039 (57%)	**0.69 (0.59-0.82)**	128 (52%)	**0.55 (0.37-0.81)**	157 (57%)	0.72 (0.49-1.07)	82 (54%)	1.02 (0.59-1.76)	0.33
	Parous, breastfed >1.5 years	4,435 (28%)	575 (23%)	**0.63 (0.54-0.75)**	405 (25%)	**0.63 (0.52-0.76)**	64 (26%)	**0.59 (0.37-0.95)**	69 (25%)	0.64 (0.40-1.02)	37 (25%)	0.81 (0.43-1.60)	0.87

a = Parity adjusted for born in Sweden or not, age, education level, breastfeeding, age at first birth and BMI.

b = Age at first birth adjusted for born in Sweden or not, age, education level, breastfeeding, parity and BMI.

c = Breastfeeding adjusted for born in Sweden or not, age, education level, parity, age at first birth and BMI.

Table 4. Risk for breast cancer overall and by subtype: Ever hormone replacement therapy (HRT use), age at menarche, somatotype at age 18, mammographic density, benign breast disease (BBD). Odds ratios with 95% confidence intervals, for controls as reference group. SD = standard deviation.

Exposure		Controls (n, %)	Cases (n, %)	OR (95% CI) Any breast cancer	Luminal A (n, %)	OR (95% CI) Luminal A	Luminal B (n, %)	OR (95% CI) Luminal B	HER2-overexpressing (n, %)	OR (95% CI) HER2	Basal-like (n, %)	OR (95% CI) Basal-like	P heterogeneity
HRT use[a]	Never	10,922 (75%)	1,414 (66%)	1.00 (ref)	972 (62%)	1.00 (ref)	162 (74%)	1.00 (ref)	172 (70%)	1.00 (ref)	108 (79%)	1.00 (ref)	
	Ever	3,703 (25%)	743 (34%)	**1.33 (1.20-1.48)**	587 (38%)	**1.43 (1.28-1.61)**	56 (26%)	0.96 (0.69-1.33)	72 (30%)	1.19 (0.89-1.62)	28 (29%)	1.01 (0.64-1.57)	**0.05**
Menarche[a]	Linear, per year increase	15,465	2,389	**0.95 (0.92-0.98)**	1,736	**0.93 (0.90-0.97)**	239	0.94 (0.86-1.03)	265	0.99 (0.92-1.09)	149	1.02 (0.92-1.14)	0.23
Absolute Mammographic Density[a]	Linear, per SD increase	14,814	1,666	**1.69 (1.62-1.78)**	1,240	**1.71 (1.62-1.80)**	160	**1.71 (1.52-1.93)**	171	**1.65 (1.42-1.86)**	95	**1.58 (1.34-1.87)**	0.80
Somatotype at age 18[b]	Linear, Increasingly endomorph	15,478	2,395	**0.93 (0.89-0.97)**	1,739	**0.93 (0.89-0.97)**	242	0.93 (0.82-1.04)	265	0.90 (0.80-1.00)	149	1.00 (0.86-1.15)	0.74
BBD, Non-proliferative lesions[c]	No	14,922 (94%)	2,461 (94%)	1.00 (ref)	1,792 (94%)	1.00 (ref)	248 (94%)	1.00 (ref)	262 (91%)	1.00 (ref)	159 (96%)	1.00 (ref)	
	Yes	1,023 (6%)	171 (6%)	0.99 (0.83-1.18)	122 (6%)	0.93 (0.76-1.13)	17 (6%)	1.11 (0.67-1.82)	25 (9%)	1.48 (0.98-2.26)	7 (4%)	0.77 (0.36-1.65)	0.19
BBD, Proliferative lesions non-atypic[c]	No	15,566 (98%)	2,544 (97%)	1.00 (ref)	1,847 (97%)	1.00 (ref)	259 (98%)	1.00 (ref)	278 (97%)	1.00 (ref)	160 (96%)	1.00 (ref)	
	Yes	379 (2%)	88 (3%)	**1.56 (1.22-1.98)**	67 (3%)	**1.67 (1.27-2.19)**	6 (2%)	1.09 (0.48-2.47)	9 (3%)	1.44 (0.73-2.82)	6 (4%)	1.40 (0.57-3.44)	0.77

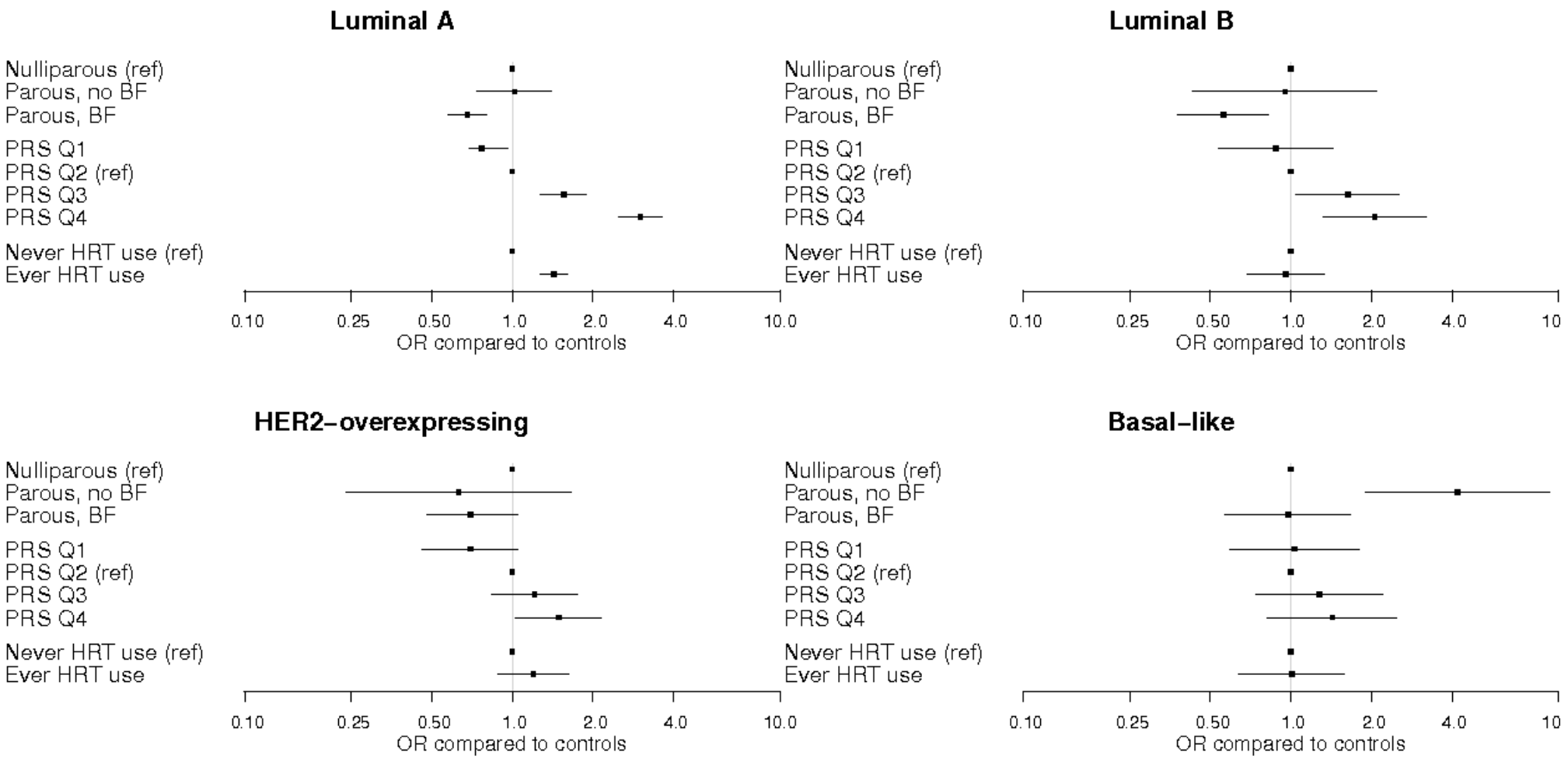
a = Adjusted for born in Sweden or not, age, education level, parity and BMI.

b = Adjusted for born in Sweden or not, age, age at menarche and education level.

c = Adjusted for born in Sweden or not, age, education level, parity and BMI.

Figure legend

Figure 1. Forest plots summarizing observed heterogeneity of results for exposures breastfeeding, PRS and HRT use across the subtypes. OR = Odds ratio. BF = Breastfeeding. PRS = Polygenic risk score. Q1-4 = Quartiles 1 to 4 of the PRS. HRT = Hormone replacement therapy.


Luminal A
Nulliparous (ref)
Parous, no BF
Parous, BF
PRS Q1
PRS Q2 (ref)
PRS Q3
PRS Q4
Never HRT use (ref)
Ever HRT use
0.10
0.25
0.50
1.0
2.0
4.0
10.0
OR compared to controls
Luminal B
Nulliparous (ref)
Parous, no BF
Parous, BF
PRS Q1
PRS Q2 (ref)
PRS Q3
PRS Q4
Never HRT use (ref)
Ever HRT use
0.10
0.25
0.50
1.0
2.0
4.0
10
OR compared to controls
HER2−overexpressing
Nulliparous (ref)
Parous, no BF
Parous, BF
PRS Q1
PRS Q2 (ref)
PRS Q3
PRS Q4
Never HRT use (ref)
Ever HRT use
0.10
0.25
0.50
1.0
2.0
4.0
10.0
OR compared to controls
Basal−like
Nulliparous (ref)
Parous, no BF
Parous, BF
PRS Q1
PRS Q2 (ref)
PRS Q3
PRS Q4
Never HRT use (ref)
Ever HRT use
0.10
0.25
0.50
1.0
2.0
4.0
10
OR compared to controls