Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status

SUPPLEMENTARY MATERIAL

Corresponding author:

Anika Hüsing,

Im Neuenheimer Feld 581

69120 Heidelberg

Germany

Phone: 49-(0)6221-42-2219
Fax: 49-(0)6221-42-2203
e-mail: a.huesing@dkfz.de

Table S1: Characteristics of subjects with complete genotypes, within the different cohorts contributing to BPC3

	CPS2		EPIC		MEC		PLCO		NHS		Total	
count (%)	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
Total	547	744	2213	2017	508	531	753	930	1988	3605	6009	7827
Age at baseline mean (sd)	62.23 (6.16)	62.01 (6.03)	54.53 (7.69)	54.77 (7.77)	60.56 (8.31)	59.25 (8.58)	62.33 (5.07)	62.26 (5.01)	58.17 (9.86)	57.17 (10.68)	57.9 (8.69)	57.8 (9.28)
Hormone Receptor status												
ER +*	413 (93)		1149 (68)		362 (84)		514 (87)		1482 (81)		3920 *(78)*	
ER -*	33 (7)		530 (32)		69 (16)		75 (13)		352 (19)		1059 *(21)*	
ER not classified	101 (18)		534 (24)		77 (15)		164 (22)		154 (8)		1030 (17)	
PR +*	335 (79)		609 (54)		296 (74)		456 (78)		1257 (69)		2953 *(68)*	
PR -*	89 (21)		510 (46)		103 (26)		125 (22)		554 (31)		1381 *(32)*	
PR not classified	123 (22)		1094 (49)		109 (21)		172 (23)		177 (9)		1675 (28)	
Age at diagnosis												
mean	70.28		58.66		65.29		66.31		61.38		62.14	
(sd)	6.49		7.84		8.46		5.69		10.43		9.28	
< 55 years	4 (1)		682 (31)		65 (13)		0		565 (28)		1316 (22)	
≥ 60 years	519 (95)		1018 (46)		359 (71)		668 (89)		1183 (60)		3747 (62)	
Disease onset missing			2 (0)								2 (0)	
Menopausal status at baseline												
premenopausal	23 (4)	32 (4)	484 (22)	453 (22)	45 (9)	85 (16)	0	0	506 (25)	1026 (28)	1058 (18)	1596 (20)
postmenopausal	518 (95)	700 (94)	1410 (64)	1295 (64)	449 (88)	438 (82)	744 (99)	922 (99)	1429 (72)	2471 (69)	4550 (76)	5826 (74)
perimenopausal / unknown	6 (1)	12 (2)	319 (14)	269 (13)	14 (3)	8 (2)	9 (1)	8 (1)	53 (3)	108 (3)	401 (7)	405 (5)

* percentages relate to subjects with classified receptor status

Table S2: Completeness of covariate and genetic information in all eligible cases and controls.

Covariate		% of missing values in all subjects
Height		0.12
Weight		1.73
Age at baseline		0
Menopausal status at baseline		0
Age at menarche		2.25
Age at menopause		7.99
Ever full term pregnancy		2.28
Number of full term pregnancies		2.75
Age at first full term pregnancy		2.23
Ever use of oral contraceptives		1.76
Ever use of hormone replacement therapy		1.87
Family history of breast cancer		21.33
Smoking status		0.66
Diabetes		1.31
Alcohol consumption		4.39
		% of all subjects
Count of variables missing per person	0	83.98
	1	12.66
	2	10.66
	3	0.54
	4 to 11	1.39
Count of SNPs missing per person	0	45.39
	1	32.53
	2	12.18
	3	4.13
	4-8	5.77

Table S3: Estimated effects of classical covariates in the covariate model in terms of odds ratios (OR) (with 95% confidence intervals) as derived from the training data, overall and by ER-status.

		All		ER+		ER -	
Variable		OR	(95%-CI)	OR	(95%-CI)	OR	(95%-CI)
Body mass index	per 1 increment change	0.99	(0.97-1.01)	0.99	(0.96-1.01)	1.00	(0.97-1.04)
	additional BMI-effect for postmenopausal women at baseline	1.03	(1.00-1.05)	1.03	(1.00-1.05)	1.00	(0.96-1.04)
	additional BMI-effect for women who were perimenopausal or undefined at baseline	1.04	(0.99-1.08)	1.03	(0.98-1.09)	1.08	(1.00-1.16)
Menopausal status at baseline*	premenopausal (reference)	1.00		1.00		1.00	
	postmenopausal	0.68	(0.25-1.81)	0.62	(0.21-1.83)	1.45	(0.09-23.1)
	perimenopausal/ undefined	0.63	(0.16-2.48)	0.71	(0.14-3.64)	0.22	(0.01-5.61)
Age at menopause	early (before 45, ref)	1.00		1.00		1.00	
	medial (45-49)	1.09	(0.94-1.27)	1.05	(0.90-1.24)	1.24	(0.90-1.70)
	late (50+)	1.22	(1.06-1.40)	1.21	(1.05-1.39)	1.22	(0.91-1.63)
Age at first parity	early (before 21)	0.83	(0.67-1.03)	0.75	(0.60-0.94)	0.82	(0.54-1.26)
	medial (21-29)	0.83	(0.67-1.03)	0.75	(0.60-0.94)	0.82	(0.54-1.26)
	late (30+)	1.00	(0.83-1.22)	0.94	(0.77-1.15)	1.02	(0.68-1.51)
	nulliparous (ref)	1.00		1.00		1.00	
Number of full term pregnancies	per birth	0.95	(0.92-0.98)	0.93	(0.90-0.97)	0.98	(0.92-1.05)
Age at menarche**	early (before 12, ref)	1.00		1.00		1.00	
	medial (12,13)	0.94	(0.83-1.06)	0.99	(0.88-1.12)	0.91	(0.72-1.15)
	late (14+)	0.95	(0.83-1.09)	0.97	(0.84-1.12)	0.99	(0.76-1.29)
Alcohol consumption	none (<1g/day, ref)	1.00		1.00		1.00	
	moderate (<14 g/day)	1.00	(0.91-1.11)	1.00	(0.90-1.11)	1.01	(0.84-1.23)
	regular (≥14 g/day)	1.32	(1.16-1.51)	1.38	(1.20-1.58)	1.17	(0.91-1.51)
Hormone Replacement therapy	premenopausal	1.00		1.00		1.00	
	Ever	0.79	(0.36-1.72)	0.86	(0.36-2.08)	0.61	(0.05-7.60)
	never	0.62	(0.28-1.36)	0.64	(0.27-1.57)	0.63	(0.05-7.87)
Smoking status at recruitment	never (ref.)	1.00		1.00		1.00	
	former	1.06	(0.96-1.16)	1.11	(1.00-1.22)	1.07	(0.88-1.29)
	current	1.09	(0.95-1.24)	1.10	(0.95-1.27)	1.24	(0.96-1.60)

* This effect is biased due to matching of cases and controls

** slightly different limits for MEC, where median category was 11 to 14 years, with early and late menarche shifted accordingly.

Table S4: Size of substrata according to disease onset and estrogen-receptor-status and model discrimination (*AUROC*$_a$ with 95% confidence interval) within these strata.

	Early onset		Late onset	
genetic effect	ER -	ER +	ER -	ER +
# cases in total	343	719	532	2590
w/o covariates				
+ 32 SNPs	0.560 (0.491 - 0.629)	0.638 (0.588 - 0.689)	0.510 (0.460 - 0.561)	0.586 (0.560 - 0.612)
+ 18 SNPs	0.562 (0.493 - 0.631)	0.634 (0.584 - 0.684)	0.510 0.460 - 0.560)	0.585 (0.560 - 0.611)
covariates alone	0.546 (0.475 - 0.618)	0.514 (0.461 - 0.568)	0.521 (0.471 - 0.572)	0.578 (0.551 - 0.605)
+ 32 SNPs	0.580 (0.510 - 0.649)	0.631 (0.581 - 0.681)	0.524 (0.474 - 0.574)	0.613 (0.587 - 0.639)
+ 18 SNPs	0.579 (0.510 - 0.648)	0.626 (0.576 - 0.676)	0.526 (0.475 - 0.576)	0.613 (0.587 - 0.639)

Conversion of relative risk-scores from multivariate logistic models to absolute risk-levels

Relative risk levels were computed from fitting multivariate logistic regression models to our case-control study population, where all subjects belonged to a specific cohort- and age-group, where specification of "cohort" included country within EPIC and study-phase within NHS.

These were the same age-groups as provided by cancer-registries for estimated age-specific incidence rates for breast cancer, summarized on a (usually nation-wide) general population level. Assuming the risk factor distribution within our cohorts, and thus in our control subjects, to be equivalent to the general population, for each cohort the controls would be on average under the same risk as the general population (in that country- and age-group). Thus we computed for each prediction model the average relative risk $\overline{RR}_{cohort,age}$ specific for a study- and age-group, calculated from the corresponding controls as

$$\overline{RR}_{cohort,age} = \frac{1}{N_{control,age}} \sum_{i=1}^{Ncontrol,age} \exp(X_i \cdot \hat{\beta}),$$

where X_i represents the vector of individual risk-factor combinations and $\hat{\beta}$ the vector of parameter estimates. This was combined with the incidence rate (AR) retrieved from the respective general population (cancer registry):

$$AR_{GP,age} = k_{cohort,age} \cdot \overline{RR}_{cohort,age}$$

This allowed the calculation of a study- and age-group specific baseline-risk $k_{cohort,age}$, associated with a relative risk of 1, which represents the absence of all risk factors that were

considered in the model, or rather the occurrence of all reference categories combined (this very special situation is usually true for only a small proportion of subjects). Combination of this calculated baseline-risk $k_{cohort,age}$ with the individual relative risk scores estimated from the logistic model produced absolute risk levels for all cases and controls within that study- and age-group.

$$AR_{individual} = k_{cohort,age} \cdot RR_{individual}$$

Age-specific incidence rates for areas corresponding to the study regions were only available on an annual basis. Approximating the increase in incidence rates up to the age of 60 according to a regression analysis of these rates, we derived 5-year risk levels by multiplying the annual rates by 5.3. From age 60 onwards, incidence rates were found to vary little, so factor 5 was applied. The incidence rates underlying these calculations are summarized in **Figure A1**.

Figure A1: Annual age-specific incidence rates for breast cancer per 100,000 women.

Assessment of hormone-receptor status

CPS-II

Information on estrogen receptor status was obtained from abstracted medical records or state cancer registries. Data from medical records for cases diagnosed from 1992-2003 were examined for information on ER testing by certified tumor registrars, and prognostic groups were defined based on qualitative descriptions from the medical record. These qualitative prognostic groups were collapsed to ER positive, negative, and unknown status. For cases diagnosed in 2004 and later, ER status from medical records was coded to the collaborative staging (CS) variable CS Site-Specific Factor 1. The estrogen receptor grouping from medical records was divided into three groups: negative values were negative, unfavorable or low, positive values were positive, favorable or elevated, all other values were considered unknown (borderline, equivocal, not classified, test not done, missing or unknown). Data from state cancer registries for case diagnosed from 1992-2003 came from the Tumor Marker 1 field that indicates ER status in breast cancer cases, if it was available. For cases diagnosed in 2004 and later, data came from CS Site-Specific Factor 1, and/or Tumor Marker1 as available from each registry. The estrogen receptor groupings from cancer registries were divided as follows: negative values were considered negative, positive values were considered positive, all other missing/unknown, borderline, test not done were classified as unknown.

EPIC

Data regarding ER/PR status have been received from 20 centers in 10 countries. Approximately, 80% of the tumors were ER-positive and 64% of the tumors were PR-positive. Deviations from expected frequencies occurred when centre-specific numbers were small.

Laboratory methods to ascertain ER and PR status included ligand-binding dextran-coated charcoal (DCC), enzyme immunoassays, immunohistochemistry (IHC), c-erb2 serum levels (ECD - extracellular domain), multi-parameter flow cytometry. For ER and PR receptors, the percentage of cell staining was the most common reported quantification method, followed by immunoreactive score (IRS) and Allred Score. If explicit score values were reported, which was the case for 23% of the tumors, a common threshold was applied to define positive receptors as follows:

- ≥ 10% cells stained or
- ≥ 20fmol/mg or
- Allred Score ≥ 3 or
- IRS ≥ 2 or
- H-Score ≥ 10 or
- Plus-system “+”.

For cases without further information on scoring and ER/PR quantification methods, the positive and negative status as reported by the respective centre were used. The heterogeneity of practices and classification of receptor status reported in the literature are reflected in the EPIC data. Previous validation studies have shown an overall robustness of the various methods and therefore appearing to compensate for the lack of precision (1-4). This was also evident in

the EPIC data, because the observed overall receptor expression frequencies corresponded to the expected distributions.

MEC

In the cohort, incident cancer cases are identified annually through cohort linkage to population-based cancer Surveillance, Epidemiology, and End Results (SEER) registries in Hawaii and Los Angeles County as well as to the California State cancer registry. Information on estrogen receptor status is also obtained through these registries.

PLCO

Incident cancers are primarily ascertained by questionnaires mailed annually to the study participants, with additional information provided by physicians, next-of-kin, and state cancer registries. Mortality is obtained through the National Death Registry. Hospital medical records and pathology reports are requested for all cancers reported. Hormone receptor status and other tumor characteristics are abstracted from hospital records for all confirmed breast cancer cases. Within the screening arm, 1147 invasive White non-Hispanic breast cancer cases with buffy coat or whole blood available for genotyping and informed consent had been identified by September 30, 2009; 150 were ER-negative, 758 were ER-positive, and 239 had an unknown ER status. Within the control arm, 1052 invasive White non-Hispanic breast cancer cases with buccal cells available for genotyping and informed consent had been identified by September 30, 2009; 134 were ER-negative, 529 were ER-positive, and 389 had an unknown ER status. The 284 ER-negative breast cancer cases from both arms of the Trial included 220 women with 0% ER-positive staining on a quantitative test, 29 women with ER-positive readings between 1% and 9%, and 35 women determined to be ER-negative based on a qualitative test.

A total of 482 controls were identified from White non-Hispanic women with no history of breast cancer who had served as controls for genome-wide association studies of lung (338) or pancreatic cancer (144). Of these controls, 459 were from the screening arm of PLCO; and 23, the control arm. Attained ages were assigned to the 482 controls so that they matched the frequency distribution of age at diagnosis of the cases. Specifically, within each of four five-year age at entry groups (55-59, 60-64, 65-69, 70-74 y), the distribution for time elapsed from age at cohort entry to age at diagnosis (one-year groupings) for the ER-negative breast cancer cases was used to assign a similar distribution of elapsed time from age at cohort entry to attained age for the controls.

NHS1 and NHS2

For these breast cancer cases, pathology reports were reviewed to obtain information on ER and PR status. Receptor status was determined by either biochemical or immunoperoxidase assay, with the immunoperoxidase assay more commonly used than the biochemical assay on the more recent breast cancer cases (5). ER results from pathology reports and central laboratory testing were in agreement for 87.3% of specimens and for 92.3% of specimens when the results were restricted to specimens originally tested with an immunohistochemical test (6). For NHS1, a total of 181 ER- incident cases of breast cancer were identified among the 1,145 women originally scanned. For NHS2, a total of 45 ER- incident cases of breast cancer were identified among the 310 women originally scanned.

Reference List

(1) Layfield LJ, Gupta D, Mooney EE. Assessment of Tissue Estrogen and Progesterone Receptor Levels: A Survey of Current Practice, Techniques, and Quantitation Methods. Breast J 2000;6(3):189-96.

(2) Magne N, Toillon RA, Castadot P, Ramaioli A, Namer M. Different clinical impact of estradiol receptor determination according to the analytical method: A study on 1940 breast cancer patients over a period of 16 consecutive years. Breast Cancer Res Treat 2006;95(2):179-84.

(3) von WR, Mengel M, Wiese B, Rudiger T, Muller-Hermelink HK, Kreipe H. Tissue array technology for testing interlaboratory and interobserver reproducibility of immunohistochemical estrogen receptor analysis in a large multicenter trial. Am J Clin Pathol 2002;118(5):675-82.

(4) Chebil G, Bendahl PO, Idvall I, Ferno M. Comparison of immunohistochemical and biochemical assay of steroid receptors in primary breast cancer--clinical associations and reasons for discrepancies. Acta Oncol 2003;42(7):719-25.

(5) Colditz GA, Rosner BA, Chen WY, Holmes MD, Hankinson SE. Risk factors for breast cancer according to estrogen and progesterone receptor status. J Natl Cancer Inst 2004;96(3):218-28.

(6) Collins LC, Marotti JD, Baer HJ, Tamimi RM. Comparison of estrogen receptor results from pathology reports with results from central laboratory testing. J Natl Cancer Inst 2008;100(3):218-21.