Potential Predictive Factors for Breast Cancer Subtypes from a North Cyprus Cohort Analysis

Ayse Ulgen[1]*, Özlem Gürkut[2], Wentian Li[3†]

1. Girne American University, Faculty of Medicine, Kyrenia, Mersin-10-Turkey

2. Dr Burhan Nalbantoğlu State Hospital (BNSH), Nicosia, Mersin-10-Turkey

3. The Robert S. Boas Center for Genomics and Human Genetics

The Feinstein Institutes for Medical Research, Northwell Health, Manhasset, NY, USA

ABSTRACT

More than three hundreds North Cyprus breast cancer patients with subtype information are surveyed for their demographic, reproductive, genetic, epidemiological factors. Despite the fact that our cohort differs significantly from some larger cohorts (e.g., the Breast Cancer Family Registry (BCFR) with samples from USA/Canada/Australia) in age, menopause status, age of menarche, parity, education level, oral contraceptive use, breast feeding, the distribution of breast subtypes is not significantly different. Using regularized regressions, we show that the estrogen-receptor-positive (ER+) subtype is positively related to post-menopause and negatively associated with hormone therapy; the estrogen-receptor-positive and progesterone-receptor-positive (ER+/PR+) subtype is positively associated with breast feeding, and negatively associated with hormone therapy status. On the other hand, the human epidermal growth factor 2 positive (HER2+) subtype, which itself is negatively correlated with ER+ and ER+/PR+, is positively related to having first-degree-relative with cancer, and negatively associated with post-menopause. Single and multiple regression also identify older age to be positively correlated to ER+ and ER+/PR+ subtypes, and negatively correlated to HER2+ subtype. Assuming ER+ and ER+/PR+ subtypes to have better prognostic, then post-menopause and breast-feeding are beneficial, and hormone therapy treatment is detrimental.

keywords: breast cancer subtypes, epidemiology, predictive factors, estrogen receptor, progesterone receptor, human epidermal receptor 2, regularized regression.

*Correspondence: ayshe.ulgen@global.t-bird.edu

†wli@northwell.edu

Introduction

Breast cancer is the most common type of cancer diagnosed in the Western part of the world. In Europe, more than 523,000 women were diagnosed with breast cancer in 2018 and more than 138,000 women died from it (Ferlay et al., 2018). World-wide, close to 2 million women are diagnosed with breast cancer each year and approximately 30% die from this disease (Bray et al., 2018). Breast cancer is largely viewed as a disease predominantly influenced by risk factors related to lifestyle (Madigan et al., 1995; McPherson et al., 2000; Martin and Weber, 2000; Key et al., 2001; Singletary, 2003; Hulka and Moorman, 2008) though through the twin studies of heritability of breast cancer, genetic contribution can still be significant (Peto and Mack, 2003; Möller et al., 2016). Recent work to combine contribution from many genetic variants to breast cancer achieves an above 60% area-under-receive-operator-curve prediction rate (Mavaddat et al., 2019; Shieh et al. , 2019b), and 20% variance explained (Lee et al., 2019).

Female hormones may affect breast cancer, and their status have been used to classify breast cancers. In particular, estrogen receptor positive (ER+) or negative (ER-) (Knight et al., 1977; Hähnel et al., 1979), progesterone receptor positive (PR+) or negative (PR-) (Osborne et al., 1980; Clark et al., 1983), human epidermal growth factor 2 positive (HER2+) or negative (HER-) (Wolff et al., 2007) are the major classification schemes of breast cancer subtypes. It has been shown that ER+/-, PR+/-, HER+/- breast cancer subtypes have different clinical features (Richard et al., 1987; Fisher et al., 1980; Onitilo et al., 2009), the cancer etiology of these subtypes can be heterogeneous, and treatment strategies also diverge. In particular, hormone receptor-positive (ER+ or PR+) subtype should expect good prognosis, using drugs like Tamoxifen/Nolvadex (Jordan, 2003; Nasrazadani et al., 2018). Similarly, the more aggressive HER2+ subtype can be treated successfully with drugs like Trastuzumab/Herceptin (Pegram et al., 1998). On the other hand, triple-negative subtype (ER-PR-HER2-) faces challenges in treatment plan (Cleator et al., 2007; Lehmann et al., 2011).

There have been international and national studies of breast cancer with large sample sizes, such as BCFR (*www.bcfamilyregistry.org*), GICR (*gicr.iarc.fr*), BCSC (*www.bcsc-research.org*). However, there has never been a breast cancer survey on the subtype distributions, potentially

explanatory variables, and correlation between these variables and breast cancer subtypes, in North Cyprus (though there are some studies in Turkey (Kuzhan et al., 2013; Yildiz et al., 2014; Özmen, 2014; Özmen et al., 2019)). To fill this gap, we present a first epidemiological survey of close to 300 breast cancer patients from North Cyprus.

We collected and analyzed reproductive (age of menarche, number of children (zero for nulliparity), menopause status, hormone therapy or not, oral contraceptive use or not, breast feeding or not, left or right breast with cancer), demographic (age at diagnosis, education level, housewife or employed), genetic (whether a first relative has cancer), and epidemiological (smoking or not, whether the patient has other cancers) information. Most of these factors are known to be risk factors for breast cancer, e.g., early menarche, late menopause, nulliparity, long hormone replacement therapy, older age,family history of breast cancer, but it is unclear which factor is predictive for breast cancer subtypes. Some information are collected but not used as they lack diversified values. For example, even though there are three male samples, the extreme imbalanced sample size makes it unlikely to extract useful information. Therefore, we exclude male samples and discard the gender information. Another example if alcoholic use whose value is "No" for all samples, which would also be not useful for the analysis.

Our analysis strategy is the following: We separate ER, PR, HER2, the dependent variables, from other factors which are independent variables. Since we do not have control (non-breast cancer) samples, it is a case-only analysis or subtypes-with-case analysis (Martínez et al., 2010; Redondo et al., 2012). The first analysis is to compare our independent variables distribution with another major public breast cancer databases, and to compare the distributions of dependent variable (i.e., breast cancer subtypes) also. Even without the raw data from database we do not have access, summary statistics with sample size/mean/standard deviation are enough for statistical tests. Second, correlation between the cancer subtypes are determined. Third, uni-variate, multiple, and regularization logistic regression are performed to detect any factor-subtype association, i.e., to identify potential predictive factors for breast cancer subtypes. We will show that though there are some minor surprises, our cohort conforms with some other studies concerning predictive factors of breast cancer subtypes.

Results

Visual inspection of the data by t-SNE: The t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008) is a popular method to represent high-dimensional data in 2 or 3 dimensions. Its application ranges from handwriting recognition (Van der Maaten, 2009) to single-cell expression data analysis (Kobak and Berens, 2018), to genetics/genomics (Li et al., 2017; Gaspar and Breen, 2019), and other biological topics (Hirata et al., 2019; Li et al., 2019).

We use 3 dependent variables (ER, PR, HER2), 5 quantitative independent variables (age of diagnosis, age of menarche, number of children, education level (0-3), cancer grade (1-4)), and 10 binary independent variables (left or right breast, menopause or not, first relative with cancer or not, having other cancer or not, smoker or not, hormone therapy or not, oral contraceptive use or not, breast feeding or not, housewife or not, invasive cancer or not). The quantitative variables are standardized to have zero-mean and unit-variance (z-transformation).

Since there are a lot of missing data for age of menarche (missing rate = 33%), hormone therapy (31%), oral contraceptive use (32%), breast feeding (33%), we only keep samples who have information on these factors. This reduced the sample size from 321 to 211. For these 211 patients, other missing data (of much lower missing rate) are imputed.

Fig.1 shows one run of t-SNE (different runs would lead to different layout of the points but similar cluster patterns). Because ER, PR, HER2 are part of the variables used in the input, it is not surprising that their values are partitioned in the plot (e.g., ER+ and ER- samples). It can be seen that ER+ samples tend to PR+, and HER2-, ER- samples tend to be PR- and HER2+. The 7 samples with other cancers (including metastasis) form a distinct cluster from the rest of the samples. While ER, PR, HER2 values separate in up-down direction in Fig.1, other factors, such as menopause status, breast feeding, age, etc. seem to be separated in (not completely) orthogonal direction.

Distribution of patient's factors: Table 1 shows that our cohort is distinct from the Breast Cancer Family Registry (BCFR) samples, which are mostly of USA/Canada/Australia origin, in several demographic or reproductive factors. The north Cyprus cohort is older, more post-menopause, lesser number of young ($\leq$ 11) age at menarche, lesser nulliparous, lesser

factor	North Cyprus				BC Fam Registry		NC vs BCFR pv		
	whole n=321	ER+PR+ n=204	ER-PR- n=64	++vs– pv	ER+PR+ n=2486	ER-PR- n=920	all	++	- -
age	57.4 ± 12.8	58.7 ± 12.5	55.3 ± 13	0.07	47.1 ± 9.3	44.5 ± 9.8		**6E-29**	**6E-4**
menopause	201 (64.8%)	136 (68.3%)	36 (57.1%)	0.13	951(40%)	310(35%)	**8E-19**	**1E-14**	**2E-6**
not	109 (35.2%)	63 (31.7%)	27 (42/9%)		1431 (60%)	574 (65%)			
NA	11	5	1						
menarche	13.15 ± 1.29	13.11 ± 1.30	13.32 ± 1.25	0.3					
age≤ 11	19 (8.8%)	12 (8.5%)	3 (5.7%)	0.43	528(22%)	183(21%)	**6E-6**	**3E-4**	6E-3
12	54 (25.1%)	40 (28.4%)	11 (20.8%)		590(24%)	215(24%)			
≥ 13	142 (66%)	89 (63.1%)	39 (73.6%)		1317(54%)	482(55%)			
NA	106	63	11						
parity	2.34 ± 1.42	2.47 ± 1.5	2.17 ± 1.18	0.1					
no.kid=0	31 (10%)	21 (10.5%)	6 (9.4%)	0.44	565 (23%)	191(21%)	**2E-7**	**8E-5**	0.04
1-2	159 (51.1%)	91 (45.5%)	35 (54.7%)		1015(41%)	391(42%)			
≥ 3	121 (38.9%)	88 (44%)	23 (35.9%)		906(36%)	338(37%)			
NA	10	4	0						
edu: < HS	174(56.5%)	113(57.1%)	52 (81.3%)	**5E-4**	710(29%)	289(32%)	**4E-20**	**4E-15**	**1E-14**
≥ HS	134(43.5%)	85 (42.9%)	12 (18.8%)		1740(71%)	602(68%)			
NA	13	6	0						
OC use	76(35%)	46 (32.4%)	19(35.8%)	0.73	1795 (73%)	680 (77%)	**7E-32**	**6E-23**	**5E-10**
no	141(65%)	96 (67.6%)	34(64.2%)		648 (27%)	198 (23%)			
NA	104	62	11						
breast feed	168(78.5%)	112(80.6%)	38 (71.7%)	0.24	1359 (55%)	454 (50%)	**5E-13**	**1E-9**	0.003
no	46 (21.5%)	27 (19.4%)	15 (28.3%)		1105 (45%)	448(50%)			
NA	107	65	11						

Table 1: A list of factors that are significantly different between North Cyprus cohort and BCFR (Breast Cancer Family Registry) cohort (data taken from (Work et al., 2014)), either for the whole dataset or for ER/PR subtypes: age at diagnosis, post or pre menopause status, age at menarche (first occurrence of menstruation), parity (number of births), education level (HS: high school), oral contraceptive use, and beast feeding. pv(++ vs - -) is the Fisher's test p-value comparing the North Cyprus ER+PR+ vs ER-PR- group. pv (NC vs BCFR) is the Fisher's test p-value comparing North Cyprus and Breast Cancer Family Registry group. Missing data (NA) are not counted in calculating percentage and not used in Fisher's test. All p-values smaller than 0.001 (this threshold is recommended in (Colquhoun, 2014)) are marked by boldface.

factor	North Cyprus				BC Fam Registry		NC vs BCFR pv		
	whole	ER+PR+	ER-PR-	++ vs - -	ER+PR+	ER-PR-	all	++	- -
	n=321	n=204	n=64	pv	n=2486	n=920			
HT	24 (19.5%)	14 (9.7%)	7 (14%)	0.43	424 (18%)	111 (13%)	0.03	**0.0093**	0.83
no	198 (80.5%)	131 (90.3%)	43 (86%)		1955[1] (82%)	758[1] (87%)			
NA	99	59	10						
fam-hist	85 (26.5%)	55 (27%)	19 (29.7%)	0.75	714 (9%)	244 (27%)	0.89	0.63	0.56
no	236 (73.5%)	149 (73%)	45 (70.3%)		1761 (29%)	673 (73%)			
grade: 1	60 (23.4%)	41 (24.4%)	10 (18.2%)	0.55					
2	135 (52.7%)	89 (53%)	30 (54.5%)		1546[2](74%)	628[2](80%)			
3	51 (19.9%)	31 (18.5%)	14 (25.5%)		554 (26%)	154 (20%)	0.11[3]	0.027[3]	0.3[3]
4	10 (3.9%)	7 (4.2%)	1 (1.8%)		NA	NA			
NA	65	36	9						
otherC	17 (5.3%)	11 (5.4%)	4 (6.3%)	0.76					
no	304(94.7%)	193 (94.6%)	60 (93.8%)						
smoke	84 (26.2%)	53 (26%)	20 (31.2%)	0.42					
no	237 (73.8%)	151 (74%)	44 (68.8%)						
breast:L	162 (50.6%)	107 (52.5%)	31 (48.4%)	0.27					
R	149 (46.6%)	90 (44.1%)	33 (51.6%)						
both	9(2.8%)	7 (3.4%)	0 (0%)						
NA	1	0	0						
housewife	118 (42.8%)	83 (45.9%)	21 (36.8%)	0.28					
employed	158 (57.2%)	98 (54.1%)	36 (63.2%)						
NA	45	33	7						
invasive	237 (75.5%)	155(76.4%)	48 (75%)	0.87					
no	77 (24.5%)	48 (23.6%)	16(25%)						
NA	7	1	0						

Table 2: Similar to Table 1, a list of factors that are not significantly different between the North Cyprus cohort and BCFR: hormone therapy (HT), if any first-degree relative has any cancer ("family history"), tumor grade, whether the patient has other cancers, smoking, left (L) or right (R) breast or both with cancer, housewife or employed, invasive cancer or not. Notes: 1: never or former menopausal hormone therapy use. 2: grade-1 plus grade-2. 3: Fisher test p-value for comparing the grade=3 and non-grade-3 counts in North Cyprus and BCFR.

education, less use of oral contraceptives, and more breast feeding. There are two explanations for these significant differences. The first is due to cultural and customary differences between countries (e.g. use of oral contraceptives). The second explanation is that our samples are collected from the state hospital, and a higher percentage of well-to-do patients may select to be treated at private hospitals, or hospitals overseas. The differences remain even for ER+/PR+ subgroup, and for ER-/PR- subgroup (though less significant due to smaller sample sizes).

Within our North Cyprus cohort, when the ER+/PR+ and ER-/PR- groups are compared on these factors, only the education level is significantly different (ER-/PR- are less educated), and ER-/PR- is slightly younger (t-test p-value = 0.07) (Table 1). Differences in other factors could become significant if the sample size is larger (e.g., there is a trend that ER-PR- patients breast feed less, or more likely to have the first menstruation at age older than 13, or more likely to be in pre-menopause stage), but not significant with our limited sample size.

On the other hand, some other demographic and other factors are not very different between our cohort and BCFR, as summarized in Table 2. These include hormone therapy usage, having a first-degree relative with cancer, and tumor grade. We also list factors which do not have the corresponding information in BCFR: having other cancers, smoking status, left or right breast with cancer, housewife or employed, invasive cancer or not. Only for ER+/PR+ subtype, the North Cyprus cohort is significantly less likely to have hormone therapy than the BCFR samples.

We also examine correlation between factors. Using all breast cancer patients without considering the subtypes, these correlations are observed: (1) patients who breast feed are less likely to go through hormone therapy (OR=7.1, Fisher p-value 9×10^{-5}); (2) patients who work are more likely to smoke than housewives (OR=3.1, p-value= 1.3×10^{-4}); (3) patients who work are more likely to be pre-menopause than housewives (OR=2.8, p-value= 1.3×10^{-4}).

Distribution of breast cancer subtypes: There are n=290 samples with all ER, PR, HER2 subtype information available. The distribution of ER, PR and HER2 values of these samples are listed in Table 3, their marginal counts are listed. ER and PR are strongly positively correlated (Fisher p-value = 1.6×10^{-35}, Odds-Ratio (OR) = 63). The majority (68% (198/290)) samples are ER+/PR+. ER and HER2 are negatively correlated (Fisher p-value =0.018, OR=0.47). PR and HER2 are also negatively correlated (Fisher p-value = 2.3

$\times 10^{-4}$, OR=0.34). There are 40 triple negative patients (ER-/PR-/HER2-) or 13.8% of the total.

If we ignore HER2 status, there are four ER/PR groups with the following distribution: ER+/PR+ n=198 (68.3%), ER+/PR- n=18 (6.2%), ER-/PR+ n=11 (3.8%), ER-/PR- n=63 (21.7%). This distribution can be compared to that from the Breast Cancer Family Registry (Work et al., 2014): ER+/PR+ n=2486 (62%), ER+/PR- n=397(9.9%), ER-/PR+ n=208(5.2%), ER-/PR- n=920 (22.9%). A Fisher's test comparing these two distributions has p-value 0.08 which is not significant at 0.05 or 0.01 level.

The highly significant correlation between ER and PR may make PR measurement redundant. In fact, it is argued that added value of PR is questionable (Hefti et al., 2013). More specifically, ER-/PR+ subtype is rare and may not be reproducible (i.e., can be reclassified to another subtype by another method) (Hefti et al., 2013). If we ignore PR and HER2, the ER-, ER+ frequencies in North Cyprus (25.5% and 74.5%) are not significantly different from those in the Breast Cancer Family Registry (28.1% and 71.9%) (Fisher test p-value is 0.38).

If we ignore ER+/PR- and ER-/PR+ subtypes, and only compare the ER+/PR+ and ER-/PR- subtype frequency between North Cyprus (75.9% and 24.1%) and BCFR (73.0% and 27.0%), the Fisher test p-value is 0.35. If we compare HER2+ and HER2- frequency in the two groups, that in North Cyprus is 24.1% and 75.9%, in BCRF is 23.4% and 76.6%, again no significant difference (Fisher p-value is 0.81).

The similarity between breast cancer subtype frequencies in North Cyprus and BCRF is in strong contrast with the dissimilarity of many demographic and reproductive factors.

The breast cancer subtype distribution in our cohort is also striking similar to another mostly European/Caucasian database: the Breast Cancer Association Consortium (BCAC) in UK (Breast Cancer Association Consortium, 2006). From Table 3, the luminal A/B (the difference between them is determined by lower/higher protein level of Ki-67, whose information is not included in the table) or simply hormone receptor positive), HER2+ but hormone receptor negative (or luminal HER2), HER2+only, triple-negative have frequency of 62%, 16%, 8%, and 14%, as compared to those of BCAC (Brouckaert et al., 2017), 66%, 13%, 7%, 13%, with Fisher p-value of 0.44. Due to limited information provided in (Brouckaert et al., 2017), we cannot carry out a more systematic comparison of factors. But we do observe some difference:

total sample	HER2- (n=220, 75.9%)		HER2+ (n=70, 24.1%)	
N=290	PR-	PR+	PR-	PR+
ER- (n=74, 25.5%)	40 (triple negative)	8	23 (HER2+only)	3
ER+ (n=216, 74.5%)	9	163	9	35
	PR- (n=81,27.9%), PR+ (n=209, 72.1%)			

Table 3: Breast cancer subtype counts in the North Cyprus cohort. Hormone receptor positive (including luminal A and luminal B) consists of 8+9+163=180 counts (62.1%). HER2+ and hormone receptor positive consists of 3+9+35=47 counts (16.2%)

e.g., percentage of patients without child is around 10% in our cohort, but more than 15% in BCAC (Brouckaert et al., 2017).

Predictive factors for breast cancer subtypes: The comparison of factor values between ER+PR+ and ER-PR- samples can also be cast into a regression of ER/PR (dependent variable) over individual factors (independent variables). Table 4 shows all results which are significant at 0.1 level from regressing ER/PR, or ER, or HER2, over either single factor by univariate logistic regression, or all factors by a multiple logistic regression.

Table 4 shows age being positively correlated with ER+/PR+, and ER+, but negatively correlated with HER2+. These results are similar to what is shown in Table 1 that ER+PR+ patients are older. Due to the positive correlation between ER and PR, ER+ patients and HER2- patients are older.

Table 4 also shows that post-menopause is positively correlated with ER+/PR+, and ER+, but negatively correlated with HER2+. Since post-menopause implies a more curable ER+ type, it is said that menopause plays a protective role for less chance to be in ER- type (Tarone and Chu, 2002). The positive correlation between menopause status and age is self-explanatory, and the association between menopause status and ER or HER2 is also easily explained by the age. Finally, Table 4 shows that HER2+ patients are more likely to have a first relative with cancer.

Between univariate and multiple regression, we also applied a regularized regression, LASSO (least absolute shrinkage and selection operator) (Tibshirani, 1996), to study the situation with a few explanatory variables. LASSO accomplished the task of variable selection (e.g., (Halinski

pheno	factor	what regression	p-value	direction
ER/PR	age	univar	0.056	+
ER	age	univar	0.024	+
ER	menopause	univar	0.0076	+
ER	menopause	multiple	0.08	+
HER2	age	univar	0.0018	-
HER2	menopause	univar	0.0016	-
HER2	menopause	multiple	0.08	-
HER2	fam histo	univar	0.0078	+
HER2	fam histo	multiple	0.02	+

Table 4: Factors that are significantly correlated with breast cancer subtypes according to either single-variable or multiple logistic regressions. The last column indicates the sign of the correlation. "family history" refers to the presence of any cancer (not necessarily breast cancer) in any first relative.

and Feldt, 1970; Li and Yang, 2002)) by imposing constraint on the sum of absolute value of all fitting coefficients, effectively setting many coefficients to be zero, thus removing the contribution from these variables. Fig.2 shows how the coefficient of each explanatory variable increases, from left to right, when the number of non-zero-coefficient variables increases, for the dependent variables of ER, ER/PR, HER2.

For ER subtype, Fig.2(A) shows the dominant contribution from menopause status, consistent with Table 4. Fig.2(A) also shows that after menopause, hormone therapy is the second most important contributing factor, though the contribution is negative. Fig.2(B) shows breast-feeding being positively contributes to the ER/PR+ status, a result not prominent in the single-variable and all-variable regression, though the trend can already be seen in Table 1. This result is consistent with reports in the literature that breast-feeding is beneficial in reducing the probability in acquiring poor prognostic breast cancers, such as triple negative subtypes (Islami et al., 2015; Fortner et al., 2019). Fig.2(C) confirms the positive contribution from first-degree-relative cancer history, and negative contribution from the menopause status to HER2+, with the latter result already seen in Table 4.

Materials and Methods

Sample collection: We included 324 samples collected retrospectively from the Dr. Burhan Nalbantoğlu State Hospital (BNSH) in Nicosia, North Cyprus between 2006-2015, with the majority from years 2011-2015 (93%). This represented around 40% of total beast cancer cases that exist in the archives during this period. The data consists of reproductive factors, histology and biomarker information such as the Estrogen receptor (ER), Progesterone receptor (PR), and human epidermal growth factor 2 (HER2) status. Permission was obtained from the Ministry of Health from the Turkish Republic of Cyprus for scientific use of the data. Additionally, ethical approval from the Eastern Mediterranean University (EMU) Ethics Committee in Famagusta was granted to conduct the study. Telephone interviews were made when necessary to collect information from patients to fill in the missing factor values.

Tumor marker data collection: For the 324 cases, pathologists from the BNSH ascertained ER and PR status from patient tumor tissue using immunohistochemistry (IHC) and/or pathology reports using a standardized protocol and pathology reporting forms. For all cases, HER2 status available (300) was provided from patient medical reports. Where tumor tissue was available, pathologists used IHC testing for ER and PR, and categorized tumors as ER and PR positive if $\geq 10\%$ of tumor cells stained positive. When the ER or PR +/- status is not labeled, but with a specific percentage, we treat it as unknown. When the left and right breast are labeled with different ER, PR, or HER2 status, it is treated as unknown. Menopausal and other information were extracted either from the medical records (with guidance/approval from an oncologist) or by telephone interviews.

Pre-processing of data: We remove the three male samples, reducing the sample size from 324 to 321. For hormone receptor status, if the left and right breast has different value, it is labeled as NA (unknown). Also, if the hormone receptor status is not binarized but represented by a percentage, it is labeled as NA.

Other re-coding of the data include: (smoking) seldom=0, quit=1, x-number-pocket=1; (menopause status) "not clear" is considered as unknown; (family history) first degree relatives are parents, children, and siblings; (other cancer) anything not "no" is considered as yes (including metastasis); (education) 0,1,2,3 are for no school, primary/middle school, high

school, college or more; (housewife/employment) "did not work" is considered as housewife, retied is considered as the same as employed; (tumor grade) "high grade" is considered as 3, inoperable is considered as 4, A/B/C are ignored; (invasive cancer) IDC(invasive ductal carcinoma)/ICC(invasive cribriform cancer)/ISC(invasive secretory cancer) are invasive, everything else is not invasive.

For regression analysis when a sub-table of the whole dataset is created, we alway do not use samples where the dependent variable is unknown. An independent variable is removed if the missing rate is too high (e.g. > 0.2). For an independent variable with low missing rate, the missing rate is imputed from the known variable value (e.g., if x is the independent variable, two values are missing, they are replaced by (R code): *sample(x[!is.na(x)][1:2])* .

Statistical programs used: All statistical analysis either used R 3.5.1 (*www.r-project.org*, released July 2018) or SPSS 17.0 (released 2008, Chicago: SPSS Inv.). The *Rtsne* R package is used for the t-SNE analysis (*github.com/jkrijthe/Rtsne*), with the default parameter setting (e.g., perplexity=30, dims=2). The *glmnet* R package (Friedman et al., 2010) is used for the LASSO analysis (alpha=1, family="binomial"). The logistic regression is carried out by the standard R function: *glm(... family=binomial(link="logit"))*, the Fisher's test by R function: *fisher.test*.

Discussion

Without control samples, we carried out a case-only analysis of potential predictive factors of different subtypes of breast cancer. Case-only design has been implemented in breast cancer studies before, and it is "an important initial step in understanding the extent of etiologic heterogeneity between tumor subtypes" (Martínez et al., 2010; Redondo et al., 2012). Since different subtypes of breast cancer have different prognostics, it is important to assess their distribution.

One of the striking results we obtained is that our North Cyprus has very similar ER+, ER+/PR+, and HER2+ as BCFR, even though our cohort is much older, more post-menopause, less educated, less hormone therapy use, more breast feeding, etc. Since we also show a correlation between menopause status and ER subtype, it might seem to be paradoxal that the

higher proportion of post-menopause samples in our data does not lead to a significantly higher ER+ frequency. In fact, the ER+ proportion (74.5%) in indeed higher than that in BCFR (71.9%), only that the difference is not statistically significant.

This topic can be discussed in a general term: can correlation at one level be translated to correlation at another level? In our case, we examine the potential similarity/dissimilarity of distribution of a factor in two data sets (low-level), and wonder whether it can be translated to the similarity/dissimilarity of distribution of a subtype affected by these predictive factor (high-level) in those two datasets. In our previous investigation of a very different issue, we did observe that correlation may not be transferable from one level to another. It is the example of genetic linkage/association analysis of multiple correlated phenotypes (Ulgen et al., 2003). One might guess that simply because these phenotypes are correlated, their risk genes should be located in the same chromosome regions. But our work on traits/phenotypes in a typical lipid panel shows that genetic linkage results are not necessarily correlated even though the phenotypes are (Ulgen et al., 2003).

Another explanation is that causal link between the two levels are not strong enough to transfer correlation from one level to another. In our LASSO analysis (Fig.2), it can be seen that the fraction of deviation explained (range of x axis) of ER, ER/PR, HER2 is at most a few percent, even using all factors. Random forest run on the same data also show that the classification rate on ER, or ER/PR, or HER2 status is not high: on average barely over 50% (results not shown). It highlights the fact that many true predictive factors of breast cancer subtypes are not yet included in our data, and also the known genetic causes of breast cancer is not part of the analysis.

In a recent systematic meta analysis of African breast cancer subtypes (Eng et al., 2014), it is found that proportion of ER+ and PR+ samples fluctuate greatly from study to study. There are also data showing the triple-negative subtype rate is much higher in African women than European/Caucasian (Huo et al., 2009; Zheng et al., 2018). To double check whether breast cancer subtype distribution in our North Cyprus cohort is still the same with another study, we picked a published summary statistics from a southeastern Turkish cohort (Kuzhan et al., 2013). The ER+, ER+/PR+, HER2+ proportions in the Turkish cohort are 73.5%, 81.8%, 30.4%, compared to our 74.5%, 75.9% and 24.1%, leading to Fisher test p-values of 0.8,

0.086, and 0.076 (number of sample in Turkish cohort with the subtype information are 438, 437, 434). These differences are within ranges, and are not significant.

Our regularized regression (LASSO) (Fig.2) reveal several potential predictive factors to breast cancer subtypes. In order to compare our results with other studies, we note (1) since we do not have normal samples, a risk factor with positive correlation means its larger value may lead to a higher risk for one breast cancer subtype vs. another, not breast cancer vs cancer-free (Key et al., 2001); (2) some predictive factor for breast cancer subtype, e.g. body-mass-index (BMI) (Phipps et al., 2011) and mammographic density (Shieh et al., 2019a) are not included in our survey; (3) our analysis treated breast cancer subtype as dependent variable and all other factors as independent variable, in a regression framework, whereas some studies are stratified with some factors fixed.

In (Kerlikowske et al., 2016), benign disease proliferation risk is higher in ER+ subtype patients than in ER- group. This can be compared our positive contribution from other cancer (including metastasis) and ER+, or ER+/PR+ subtype (Fig.2(A,B). In (Work et al., 2014), not breast feeding is associated with ER-/PR- subtype, which can be compared with our result that breast feeding is positively correlated with the ER+/PR+ subtype. In (Tarone and Chu, 2002), ER- cancer rate stop to increase at certain age, whereas ER+ rate continue to increase. This can be compared to our result that post-menopause is is positively correlated with the ER+ subtype (Fig.2(A)). In (Colditz et al. , 2004), significant difference of age, menopause status, past use of hormone therapy is observed in four ER/PR group. In (Yang et al., 2010), early age at menarche ($\leq$ 12 years) is less common in PR- group than PR+ group, and it is also true in our data comparing ER-/PR- and ER+/PR+ groups. To summarize, many of our observed predictive factors for breast cancer subtypes are consistent with the literature. The positive correlation between cancer family history and HER2+ subtype (Fig.2(C)) remains intriguing.

In conclusion, we use a unique cohort of breast cancer in a under-studied population to survey the breast cancer subtypes and related factors. We use a simplified analysis framework: keeping breast cancer subtypes at one level, and all factors at another level. Distribution of many factors are extremely different from that of another large breast cancer registry, while the subtype distribution is similar. This indirectly shows that we have not exhaustively

measured all predictive factors of breast cancer subtypes. The relationship between the two levels is investigated by regression, with one variable, all variable, or subset of variables. These regression analyses show post-menopause and/or older breast cancer patients are more likely to have the ER+ subtype and HER2- subtype; hormone therapy is positively correlated with ER- or ER-/PR- subtype; breast-feeding and/or older breast cancer patients are more like to have ER+/PR+ subtype; and family history is observed more frequently in HER2+ subtypes.

Acknowledgements

This study was supported through the Fulbright Visiting Research Scholarship Grant by the US Department of State. AU would like to thank her advisor Prof. Mary Terry Beth and her colleagues at the BCFR at the Department of Epidemiology at Columbia University in New York for their help and support. AU would also like to thank to the Ministry of Health at the Turkish Republic of Northern Cyprus for access to the breast cancer archives at the Burhan Nalbantoglu State Hospital in Nicosia, and Dr. Nilay Acar, Dr. Mehmet Ali Alpdoğan, Dr. Fuat Ağlarcan, and Dr.Whitney A. Onuorah from the Faculty of Medicine, Eastern Mediterranean University for their help and guidance for the data collection. WT would like to thank the support from Robert S Boas Center for Genomics and Human Genetics.

References

F Bray, J Ferlay, I Soerjomataram, RL Siegel, LA Torre, A Jemal (2018), Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: A Cancer J. for Clinicians*, 68:394-424.

Breast Cancer Association Consortium (2006), Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium, *N. Natl. Cancer Inst.*, 98:1382-1396.

O Brouckaert, A Rudolph, A Laenen, R Keeman, MK Bolla, Q Wang, A Soubry, H Wildiers, IL Andrulis, V Arndt, et al. (2017), Reproductive profiles and risk of breast cancer subtypes: a multi-center case-only study, *Breast Cancer Res.*, 19:119.

GM Clark, WL McGuire, CA Hubay, OH Pearson, JS Marshall (1983), Progesterone receptors as a prognostic factor in stage II breast cancer, *New Eng. J. Med.*, 309:1343-1347.

S Cleator, W heller, RC Coombes (2007), Triple-negative breast cancer: therapeutic options, *Lancet Oncology*, 8:235-244.

GA Colditz, BA Rosner, WY Chen, MD Holmes, SE Hankinson (2004), Risk factors for breast cancer according to estrogen and progesterone receptor status, *J. Natl. Cancer Inst.*, 96:218-228.

D Colquhoun (2014), An investigation of the false discovery rate and the misinterpretation of p-values, *Royal Soc. Open Sci.*, 1:140216.

A Eng, V McCormack, I Dos-Santos-Silva (2014), Receptor-defined subtypes of breast cancer in indigenous populations in Africa: a systematic review and meta-analysis, *PLoS Med.*, 11:e1001720.

J Ferlay, M Colombet, I Soerjomataram, T Dyba, G Randi, M Bettio, A Gavin, O Visser, F Bray (2018), Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018, *Euro. J. Cancer*, 103:356-287.

ER Fisher, CK Redmond, H Liu, H Rockette, B Fisher (1980), Correlation of estrogen receptor and pathologic characteristics of invasive breast cancer, *Cancer*, 45:349-353.

RT Fortner, J Sisti, B Chai, LC Collins, B Rosner, SE Hankinson, RM Tamimi, AH Eliassen (2019), Parity, breastfeeding, and breast cancer risk by hormone receptor status and molecular phenotype: results from the Nurses Health Studies, *Breast Cancer Res.*, 21:40.

J Friedman, T Hastie, R Tibshirani (2010), Regularization paths for generalized linear models via coordinate descent, *J. Stat. Software*, 33:1-22.

HA Gaspar and G Breen (2019), Probabilistic ancestry maps: a method to assess and visualize population substructures in genetics, *BMC Bioinfo.*, 20:116.

R Hähnel, T Woodings, AB Vivian (1979), Prognostic value of estrogen receptors in promary breast cancer, *Cancer*, 44:671-675.

RS Halinski and LS Feldt (1970), The selection of variables in multiple regression analysis, *J. Edu. Measurement*, 7:151-157.

MM Hefti, R Hu, NW Knoblauch, LC Collins, B Haibe-Kains, RM Tamimi, AH Beck (2013), Estrogen receptor negative/progesterone receptor positive breast cancer is not a reproducible subtype, *Breast Cancer Res.*, 15:R68.

J Hirata, K Hosomichi, S Sakaue, M Kanai, H Nakaoka, K Ishigaki, K Suzuki, M Akiyama, T Kishikawa, K Ogawa, T Masuda, K Yamamoto, M Hirata, K Matsuda, Y Momozawa, I Inoue, M Kubo, Y Kamatani, Y Okada (2019), Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population, *Nature Genet.*, 51:470-480.

D Huo, F Ikpatt, A Khramtsov, J-M Dangou, R Nanda, J Dignam, B Zhang, T Grushko, C Zhang, O Oluwasola, D Malaka, S Malami, A Odetunde, AO Adeoye, F Iyare, A Falusi, CM Perou, OI Olopade (2009), Population differences in breast cancer: survey in indigenous African women reveals over-representation of triple-negative breast cancer, *J. Clin. Oncol.*, 27:4515-4521.

BS Hulka and PG Moorman (2008), Reprint of Breast cancer: hormones and other risk factors, *Maturitas*, 61:203-213.

F Islami, Y Liu, A Jemal, J Zhou, E Weiderpass, G Colditz, P Boffetta, M Weiss (2015), Breastfeeding and breast cancer risk by receptor status - - a systematic review and meta-analysis, *Ann. Oncol.*, 26:2398-2407.

VC Jordan (2013), Tamoxifen: a most unlikely pioneering medicine, *Nature Rev. Drug Dis.*, 2:205-213.

K Kerlikowske, CC Gard, JA Tice, E Ziv, SR Cummings, DL Miglioretti, Breast Cancer Surveillance Consortium (2016), Risk factors that increase risk of estrogen receptor-positive and -negative breast cancer, *J. Natl. Cancer Inst.*, 109:djw276.

T Key, PK Verkasalo, E Banks (2001), Epidemiology of breast cancer, *Lancet Oncology*, 2:133-140.

WA Knight III, RB Livingston, EJ Gregory, WL McGuire (1977), Estrogen receptor as an independent prognostic factor for early recurrence in breast cancer, *Cancer Res.*, 37:4669-4671.

D Kobak and P Berens (2018), The art of using t-SNE for single-cell transcriptomics, *bioRxiv* preprint 453449. DOI: *10.1101/453449*

A Kuzhan, M Adli, H Eryigit Alkis, D Caglayan (2013), Hormone receptor and HER2 status in patients with breast cancer by races in southeastern Turkey, *J. Balkan Union Oncol.*, 18:619-622.

A Lee, N Mavaddat, AN Wilcox, AP Cunningham, T Carver, S Hartley, CB de Villiers, A Izquierdo, J Simard, MK Schmidt, FM Walter, N Chatterjee, M Garcia-Closas, M Tischkowitz, P Pharoah, DF Easton, AC Antoniou (2019), BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors, *Genet. in Med.*, in press.

BD Lehmann, JA Bauer, X Chen, ME Sanders, AB Chakravarthy, Y Shyr, JA Pietenpol (2011), Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies, *J. Clin. Invest.*, 121:2750-2767.

W Li, JE Cerise, Y Yang, H Han (2017), Application of t-SNE to human genetic data, *J. Bioinfo. and Comp. Biol.*, 15:1750017.

W Li, J Freudenberg, J Freudenberg (2019), Alignment-free approaches for predicting novel Nuclear Mitochondrial Segments (NUMTs) in the human genome, *Gene*, 691:141-152.

W Li and Y Yang (2002), How many genes are needed for a discriminant microarray data analysis, in *Methods of Microarray Data Analysis*, eds. SM Lin, KF Johnson, (Kluwer Acvademic Publishers), pp.137-149.

MP Madigan, RG Ziegler, J Benichou, C Byrne, RN Hoover (1995), Proportion of breast cancer cases in the United States explained by well-established risk factors, *JNCI: J. Natl. Cancer Inst.*, 22:1681-1685.

AM Martin and BL Weber (2000) Genetic and hormonal risk factors in breast cancer, *JNCI: J. Natl. Cancer Inst.*, 92:1126-1135.

ME Martínez, GI Cruz, AM Brewster, ML Bondy, PA Thompson (2010), What can we learn about disease etiology from case-case analyses? lessons from breast cancer, *Cancer Epid. Biomarker & Prev.*, 19:2710-2714.

N Mavaddat, K Michailidou, J Dennis, M Lush, L Fachal, A Lee, JP Tyrer, TH Chen, Q Wang, MK Bolla, et al. (2019), Polygenic risk scores for prediction of breast cancer and breast cancer subtypes, *Am. J. Hum. Genet.*, 104:21-34.

K McPherson, CM Steel, JM Dixon (2000), Breast cancerepidemiology, risk factors, and genetics, *BMJ*, 321:624-628.

S Möller, LA Mucci, JR Harris, T Scheike, K Holst, U Halekoh, HO Adami, K Czene, K Christensen, NV Holm, E Pukkala, A Skytthe, J Kaprio, JB Hjelmborg (2016), The heritability of breast cancer among women in the Nordic twin study of cancer, *Cancer Epidemiol. Biomarkers & Prev.*, 25:145-150.

A Nasrazadani, RA Thomas, S Oesterreich, AV Lee (2018), Precision medicine in hormone receptor-positive breast cancer, *Front. Oncol.*, 8:144.

AA Onitilo, JM Engel, RT Greenlee, BN Mukesh (2009), Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival, *Clin. Med. & Res.*, 7:4-13.

CK Osborne, MG Yochmowitz, WA Knight III, WL McGuire (1980), The value of estrogen and progesterone receptors in the treatment of breast cancer, *Cancer*, 46:2884-2888.

V Özmen (2014), Breast cancer in Turkey: clinical and histopathological characteristics (analysis of 13,240 patients), *J. Breast Health*, 10:98-105

V Özmen, T Özmen, V Doğru (2019), Breast cancer in Turkey; an analysis of 20,000 patients with breast cancer, *Euro. J. Breast Health*, 15:141-146.

MD Pegram, A Lipton, DF Hayes, B L Weber, JM Baselga, D Tripathy, D Baly, SA Baughman, T Twaddell, JA Glaspy, DJ Slamon (1998), Phase II study of receptor-enhanced chemosensitivity using recombinant humanized anti-p185HER2/neu monoclonal antibody plus cisplatin in patients with HER2/neu-overexpressing metastatic breast cancer refractory to chemotherapy treatment, *J. Clin. Oncol.*, 16:2659-2671.

J Peto and TM Mack (2000), High constant incidence in twins and other relatives of women with breast cancer, *Nature Genet.*, 26:411-414.

AI Phipps, RT Chlebowski, R Prentice, A McTiernan, ML Stefanick, J Wactawski-Wende, LH Kuller, LL Adams-Campbell, D Lane, M Vitolins, GC Kabat, TE Rohan, CI Li (2011), Body size, physical activity, and risk of triple-negative and estrogen receptorpositive breast cancer, *Cancer Epid. Biomarkers & Prev.*, 20:454-463.

CM Redondo, M Gago-Domínguez, SM Ponte, ME Castelo, X Jiang, AA Garca, MP Fernández, MA Tomé, M Fraga, F Gude, ME Martnez, VM Garzn, Carracedo, JE Castelao (2012), Breast feeding, parity and breast cancer subtypes in a Spanish cohort, *PLoS ONE*, 7:e40543.

J Richard, C Sainsbury, GK Needham, JR Farndon, AJ Malcolm, AL Harris (1987), Epidermal-growth-factor receprtor status as predictor of early recurrence of and death from breast cancer, *Lancet*, 329:1398-1402.

Y Shieh, L Fejerman, PC Lott, K Marker, SD Sawyer, D Hu, S Huntsman, J Torres, M Echeverry, ME Bohorquez, wt al. (2019), A polygenic risk score for breast cancer in U.S. Latinas and Latin-American women, *bioRxiv* preprint 598730. DOI: *10.1101/598730*

Y Shieh, CG Scott, MR Jensen, AD Norman, KA Bertrand, VS Pankratz, KR Brandt, DW Visscher, JA Shepherd, RM Tamimi, CM Vachon, K Kerlikowske (2019), Body mass index, mammographic density, and breast cancer risk by estrogen receptor subtype, *Breast Cancer Res.*, 21:48.

SE Singletary (2003), Rating the risk factors for breast cancer, *Ann. Surg.*, 237:474-482.

RE Tarone and KC Chu (2002), The greater impact of meonopause on ER- than ER+ breast cancer incidence: a possible explanation, *Cacner Causes & Control*, 13:7-14.

R Tibshirani (1996), Regression shrinkage and selection via the lasso, *J. Royal Stat. Soc. B*, 58:267-288.

A Ulgen, Z Han, W Li (2003), Correlation between quantitative traits and correlation between corresponding LOD scores: detection of pleiotropic effects, *BMC Genet.*, 4:S60.

LJ Van der Maaten and GE Hinton (2008), Visualizing data using t-SNE, *J. Machine Learning Res.*, 9:2575-2605.

LJ Van der Maaten (2009), A new benchmark dataset for handwritten character recognition, Technical Report TR 2009-002, Yilburh Centre for Creative Computing, Tilburg University.

AC Wolff, ME Hammond, JN Schwartz, KL Hagerty, DC Allred, RJ Cote, M Dowsett, PL Fitzgibbons, WM Hanna, A Langer, LM McShane, S Paik, MD Pegram, EA Perez, MF Press, A Rhodes, C Sturgeon, SE Taube, R Tubbs, GH Vance, M van de Vijver, TM Wheeler, DF Hayes, American Society of Clinical Oncology/College of American Pathologists (2007), American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer, *J. Clin. Oncol.*, 25:118145.

ME Work, EM John, IL Andrulis, JA Knight, Y Liao, AM Mulligan, MC Southey, GG Giles, GS Dite, C Apicella, H Hibshoosh, JL Hopper, MB Terry (2014), Reproductive risk factors and oestrogen/progesterone receptor-negative breast cancer in the Breast Cancer Family Registry, *Brit. J. Cancer*, 110:1367-1377.

XR Yang, J Chang-Claude, EL Goode, FJ Couch, H Nevanlinna, RL Milne, M Gaudet, MK Schmidt, A Broeks, A Cox, et al. (2010), Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies, *J. Natl. Cancer Inst.*,103:250-263.

B Yildiz, E Fidan, F Ozdemir, O Sezen, H Kavgaci, F Aydin (2014), Clinicopathological

characteristics of triple-negative breast cancers in the northeast region of Turkey, *Balkan Med. J.*, 31:126-131.

Y Zheng, T Walsh, S Gulsuner, S Casadei, MK Lee, TO Ogundiran, A Ademola, AG Falusi, CA Adebamowo, AO Oluwasola, A Adeoye, A Odetunde, CP. Babalola, OA Ojengbede, S Odedina, I Anetor, S Wang, D Huo, TF Yoshimatsu, J Zhang, GES Felix, M-C King, OI Olopade (2018), Inherited breast cancer in Nigerian women, *J. Clin. Oncol.*, 36:2820-2825.

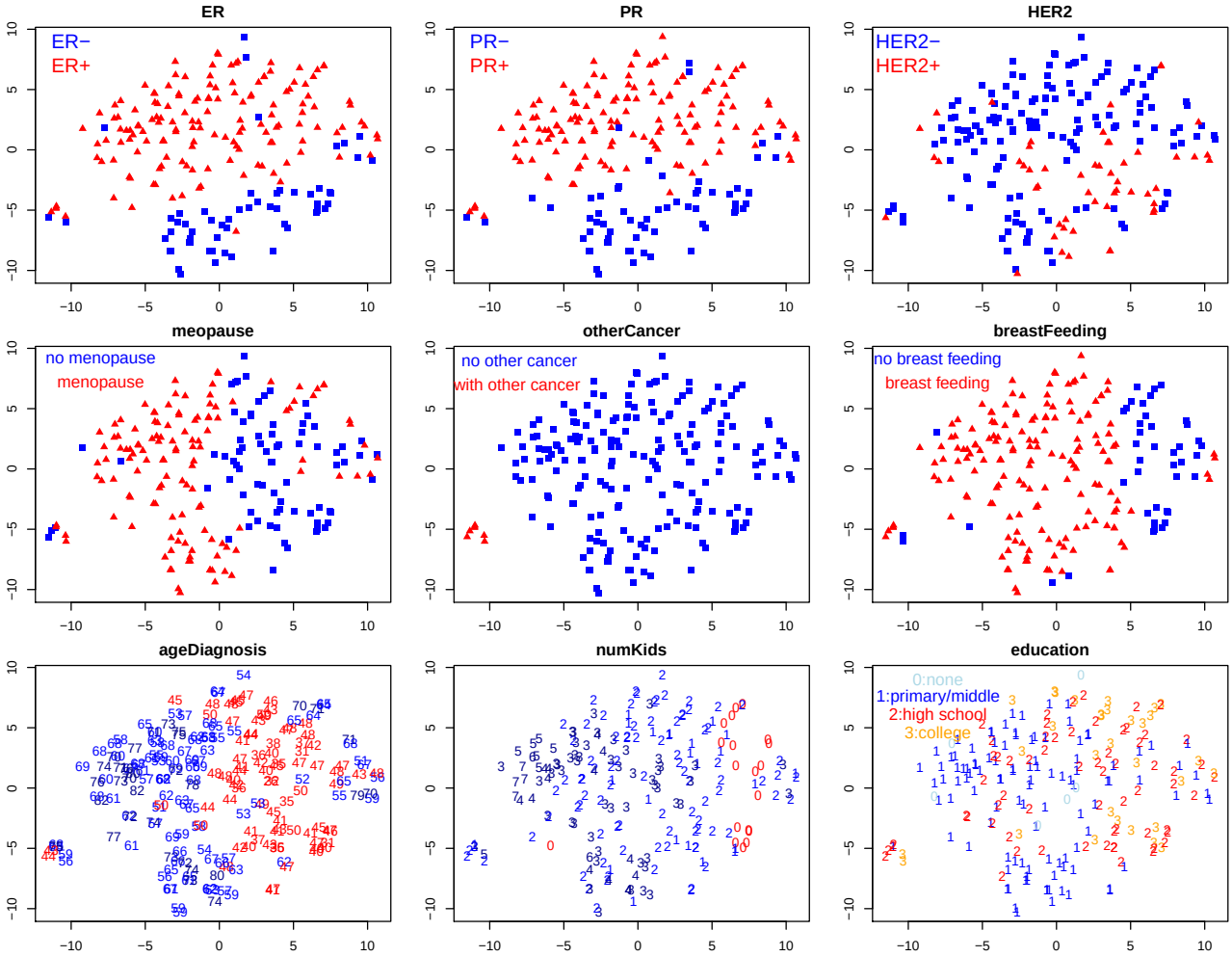

Figure 1: t-SNE plot of 211 breast cancer patients (out of 321 total) with enough non-missing factor values. The nine subplots are the same plot labeled with different information: ER subtype (red for ER+, blue for ER-), PR subtype, HER2 subtype, menopause status (post-menopause in red, pre-menopause in blue), if the patient has other cancer (red for yes, blue for no), breast feeding (red for yes, blue for no), age of diagnosis (red if younger or equal of 50 years ago), parity/number of children, education level (0 for none, 1 for primary or middle school, 2 for high school, 3 for college or higher).

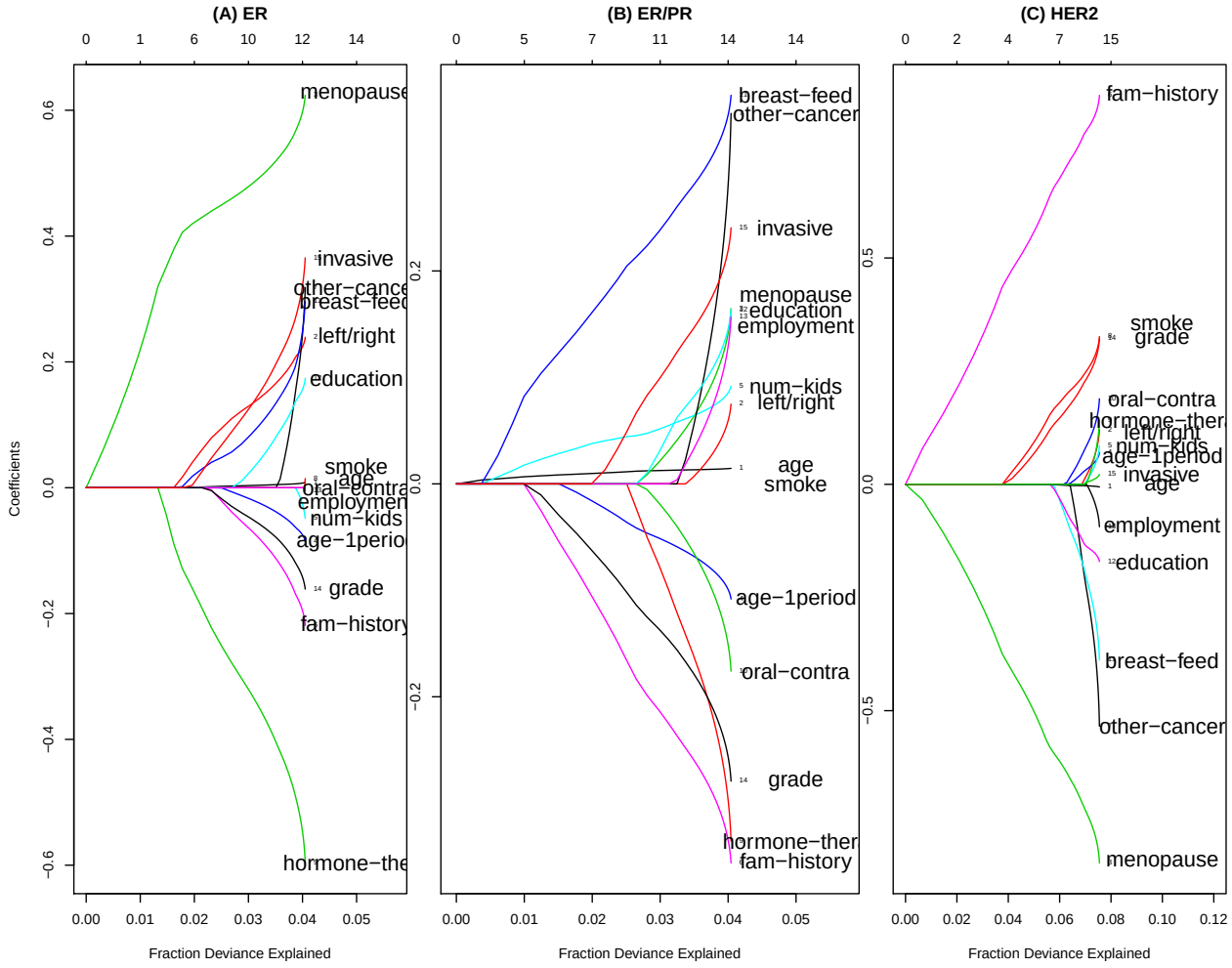

Figure 2: Variable tracing/selecting plot of LASSO logistic regression for ER, ER/PR, and HER2. Each line is a factor, and x direction represents a more relaxed constraint, allowing more variables. The y axis is the coefficient of a factor/variable: positive (negative) coefficient means a positive (negative) correlation between the factor and the subtype status (ER+, ER+PR+, HER+ are 1's, ER-, ER-PR-, HER2- are 0's). The x axis is deviance explained by the logistic regression.