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

Ayse Ulgen[1,2] ⓘ, Özlem Gürkut[3] ⓘ, Wentian Li[4] ⓘ

[1]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, USA
[2]Girne American University, School of Medicine, Kyrenia, Cyprus
[3]Department of Oncology, Dr Burhan Nalbantoğlu State Hospital, Nicosia, Cyprus
[4]The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institutes for Medical Research, Northwell Health, NY, USA

ORCID iDs of the authors: A.U. 0000-0002-0872-667X; Ö.G. 0000-0002-3906-1092; W.L. 0000-0003-1155-110X.

**BACKGROUND/AIMS**
We present the first epidemiological survey from North Cyprus to determine the predictive factors for breast cancer subtypes.

**MATERIAL and METHODS**
More than 300 patients with breast cancer, with 90% of them having the cancer subtype information, were examined at the State Hospital in Nicosia between 2006 and 2015 for their demographic, reproductive, genetic, and epidemiological factors. The breast cancer subtypes and the estrogen receptor (ER) +/- progesterone receptor (PR) +/- and human epidermal growth factor 2 (HER2) +/- status were determined. Single and multiple variable regularized regressions, with predictive factors as independent variables and breast cancer subtypes as dependent variables, were conducted.

**RESULTS**
Our cohort differed significantly from larger cohorts (e.g., the Breast Cancer Family Registry) in terms of age, menopause status, age at menarche, parity, education, oral contraceptive use, and breastfeeding, but the distribution of breast cancer subtypes was not significantly different. The subtype distribution in our cohort was also not different from that of another Turkish cohort. We found that the ER+ subtype was positively related to age/postmenopause, ER+/PR+ subtype was associated positively with age but negatively with cancer stage, and HER2+ subtype that negatively correlated with ER+ and ER+/PR+ was associated positively with cancer stage but negatively with age/postmenopause.

**CONCLUSION**
Assuming ER+ and ER+/PR+ to have better prognostic, HER+ to have worse prognostic, then older age and postmenopause seem to be beneficial, smoking and family history of cancer seem to be detrimental. Further steps include exploring potential biomarkers and using cure models to determine long-term breast cancer survivors.

**Keywords:** Breast cancer subtypes, predictive factors, estrogen, progesterone, human epidermal receptors, regularized regression, LASSO, ridge, elastic nets

## INTRODUCTION

Breast cancer is the most common type of cancer diagnosed in the western part of the world. In Europe, there were more than 523,000 breast cancer diagnoses and more than 138,000 deaths among women in 2018 (1). Worldwide, almost 2 million women are diagnosed with breast cancer each year, and approximately 30% of them die from this disease. Breast cancer is largely considered as a disease predominantly influenced by lifestyle-related risk factors (2), although twin studies of heritability of breast cancer have shown that genetic contribution could be significant (3). Recent research on the combined contribution from several genetic variants to breast cancer reports a >60% area-under-receiver-operator-curve prediction rate (4, 5), explaining 20% variance (6).

Ulgen et al. Breast Cancer Subtypes and Predictive Factors

Cyprus J Med Sci 2020; 5(4): 339-49

Female hormones may affect breast cancer, and their status has been used to classify breast cancer types. In particular, estrogen receptor-positive (ER+) or -negative (ER-), progesterone receptor-positive (PR+) or -negative (PR-), and human epidermal growth factor 2-positive (HER2+) or -negative (HER-) are the major classification schemes of breast cancer subtypes. It has been demonstrated that ER+/-, PR+/- and HER+/- breast cancer subtypes have different clinical characteristics (7); the cancer etiology of these subtypes can be heterogeneous, and treatment strategies are also divergent. Especially, hormone receptor-positive (ER+ or PR+) subtype might have good prognosis using drugs such as tamoxifen/Nolvadex. Similarly, the more aggressive HER2+ subtype can be treated successfully using drugs such as trastuzumab/Herceptin. In contrast, the triple-negative subtype (ER-/PR-/HER2-) poses challenges in treatment strategies (8).

Although international and national studies of breast cancer have been conducted with large sample sizes, such as the Breast Cancer Family Registry (BCFR) (*www.bcfamilyregistry.org*), the Global Initiative for Cancer Registry Development (GICR) (*gicr.iarc.fr*), and the Breast Cancer Surveillance Consortium (BCSC) (*www.bcsc-research.org*), there has never been a breast cancer survey investigating the subtype distributions, potentially explanatory variables, and the correlation between these variables and breast cancer subtypes in North Cyprus (except some studies in Turkey) (9). To fill this research gap, we conducted the first epidemiological survey of approximately 300 patients with breast cancer from North Cyprus, among whom approximately 230 patients were Turkish Cypriots.

We collected and analyzed data regarding reproductive (age at menarche, number of children (0 for nulliparity), menopause status, hormone therapy, oral contraceptive use, breastfeeding, and left/right breast with cancer), demographic (age at diagnosis, education level, and housewife/employed), genetic (first-degree relative having cancer), and epidemiological (smoking and other cancer types) characteristics. The majority of these characteristics are known to be risk factors for breast cancer, e.g., early menarche, late menopause, nulliparity, long hormone replacement therapy, older age, and family history of breast cancer; however, it is unclear which factor is predictive for breast cancer subtypes.

Our analysis strategy is as follows: We considered ER, PR, and HER2 as dependent variables and others as independent variables. As we did not have control (non-cancer) samples, this survey was a case-only analysis or subtypes-with-case analysis (10, 11). The first analysis was conducted to compare the distribution of our independent and dependent variables with that reported in major public breast cancer databases. Second, we determined the correlation between the cancer subtypes. Third, univariate, multiple, and regularized logistic regressions were performed to detect any factor–subtype association, i.e., to identify potential predictive factors for breast cancer subtypes. Although we observed some minor surprising results, our study cohort conforms to some other studies concerning predictive factors for breast cancer subtypes.

**MATERIAL and METHODS**

**Sample collection**: We included the data of 324 subjects (321 women; 3 men) collected retrospectively from Dr. Burhan Nalbantoğlu State Hospital (BNSH) in Nicosia, North Cyprus, during 2006–2015, largely from the years 2011–2015 (93%). This sample represented approximately 40% of total breast cancer cases that existed in the archives during this period. The data consisted of reproductive factors, histology, and biomarker information such as the status of ER, PR, and HER2. Permission was obtained from the Ministry of Health from the Turkish Republic of Cyprus for the scientific use of the data. In addition, ethical approval to conduct the study was obtained from the Eastern Mediterranean University Ethics Committee in Famagusta, with the IRB approval number AAAP8950. Patient consent forms were not required. Telephone interviews were conducted when necessary to collect information from patients to fill in the missing factor values.

For the study sample, pathologists from the BNSH ascertained the ER and PR status based on the immunohistochemistry (IHC) and/or pathology reports of the patients' tumor tissues using a standardized protocol and pathology reporting forms. For all cases, the available HER2 status (approximately 290 cases) was provided from patient medical reports. In cases where tumor tissue was available, pathologists used IHC testing for ER and PR and categorized tumors as ER and PR positive if ≥10% of tumor cells stained positive. When the ER or PR +/- status is not labeled, but with a specific percentage, we considered it as unknown. Menopause status and other information were extracted either from the medical records (with guidance/approval from an oncologist) or by telephone interviews.

**Preprocessing of data**: We excluded the three male samples, reducing the sample size from 324 to 321. Regarding the hormone receptor status, if the left/right breast had a different value, it was labeled as NA (unknown). Moreover, if the hormone receptor status was not binarized but represented by a percentage, it was labeled as NA.

Other recoding of data included (smoking) seldom=0, quit=1, x-number-pocket=1; (family history) first-degree relatives are parents, children, and siblings; (other cancer) anything not "no" was considered as yes (including metastasis); (education) 0, 1, 2, and 3 for no school, primary/middle school, high school, and college or more, respectively; (housewife/employed) retired was considered as the same as employed; (tumor stage) "high stage" was considered as 3, inoperable was considered as 4, A/B/C were ignored; (invasive cancer) invasive ductal carcinoma/invasive cribriform cancer/invasive secretory cancer were considered as invasive, and all others were considered as not invasive.

---

**Main Points:**

- A Turkish Cypriot cohort of breast cancer patients show a similar distribution of subtypes based on hormone receptor status as other international cohorts.
- Older age and postmemopause status tend to have the ER+ or PR+ subtype with a better prognosis.
- Although having not achieved statistical significance, breastfeeding seems to be positively associated with ER+PR+ subtype, hormone therapy negatively being associated with ER+ subtype, and family history of cancer being positively associated with HER2+ subtype.

Of the 321 samples, 300 had ER or PR status information, 291 had HER2 status information, all samples had age information, but only 214, 215, 217, and 222 patients had details regarding breastfeeding, age at menarche, use of oral contraceptives, and use of hormone therapy. The amount of missing data for other factors is given in Tables 1 and 2. An independent variable was removed when the missing rate was too high (e.g., >0.2). The remaining missing values were imputed from the values of known variables (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]).*

**Turkish Cypriot and nonTurkish Cypriot patients:** Of the 321 patients, 314 reported their birth country, which indicated that the majority of them were born in Cyprus (n=233), 53 were born in Turkey, and the remaining 28 were born in other countries, including UK, Turkmenistan, and Bulgaria. Although our analysis did not focus on the genetic or ethnicity contribution to the breast cancer subtypes, and foreign-born does not automatically imply non-Turkish Cypriots, we performed all analyses twice, one using 321 samples and the other using 233 Turkish Cypriot-only samples.

**Statistical Analysis**

All statistical analyses were conducted using either R 3.5.1 (*www.r-project.org*, released July 2018) or the Statistical Package for Social Sciences software version 17.0 (SPSS Inc.; Chicago, IL, USA). The *Rtsne* R package was used for the t-SNE analysis (*github.com/jkrijthe/Rtsne*), with default parameter settings (e.g., perplexity=30, dims=2). The *glmnet* R package (*web.stanford.edu/~hastie/glmnet*) (12) was used for the following regularized regressions: least absolute shrinkage and selection operator (LASSO; alpha=1, family="binomial"), elastic net (alpha=0.5), and ridge (alpha=1). The logistic regression was conducted using the standard R function *glm(... family=binomial(link="logit"),* and Fisher's test was conducted using the R function *fisher test*. The independent two-sample t-test between age distribution, one from raw data and the other from summary statistics, was conducted using our customized R script.

**RESULTS**

**Visual inspection of data by t-SNE:** The t-distributed stochastic neighbor embedding (t-SNE) (13) is a commonly used method

| | **North Cyprus** | | | | **BC Fam Registry** | | **NC vs BCFR pv** | | |
|---|---|---|---|---|---|---|---|---|---|
| factor | 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 | 0.06 | 47.1±9.3 | 44.5±9.8 | | **6E-29** | **1E-8** |
| menopause | 201 (64.8%) | 136 (68.3%) | 36 (57.1%) | 0.13 | 951 (40%) | 310 (35%) | **5E-18** | **1E-14** | **6E-4** |
| Not | 109 (35.2%) | 63 (31.7%) | 27 (42.9%) | | 1431 (60%) | 574 (65%) | | | |
| NA | 11 | 5 | 1 | | | | | | |
| menarche | 13.14±1.31 | 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%) | **1E-5** | **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%) | **9E-8** | **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%) | 33 (51.6%) | 0.47 | 710 (29%) | 289 (32%) | **1E-20** | **4E-15** | **0.002** |
| ≥ HS | 134 (43.5%) | 85 (42.9%) | 31 (48.4%) | | 1740 (71%) | 602 (68%) | | | |
| NA | 13 | 6 | 0 | | | | | | |
| OC use | 76 (35%) | 46 (32.4%) | 19 (35.8%) | 0.73 | 1795(73%) | 680 (77%) | **3E-31** | **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%) | **2E-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.** Factors that are distributed differently between North Cyprus cohort and BCFR

Factors that are significantly different between the North Cyprus cohort and the BCFR cohort: age at diagnosis, postmenopause or premenopause status, age at menarche (first occurrence of menstruation), parity (number of births), education level (HS: high school), oral contraceptive use, breastfeeding. 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 the North Cyprus and BCFR groups. Missing data (NA) are not counted in calculating the percentage and not used in Fisher's test. All p values smaller than 0.001 (this threshold is recommended in (37)) are marked by boldface.
Abbr.: ER; Estrogen Receptor, PR; Progesterone Receptor, HER2; Human Epidermal Growth Factor 2, BCFR; Breast Cancer Family Registry, NC; North Cyprus, edu; Education, HS; High School, OC; Oral Contraceptives

Ulgen et al. Breast Cancer Subtypes and Predictive Factors

Cyprus J Med Sci 2020; 5(4): 339-49

**TABLE 2.** Other Factors

| 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 | ++ | -- |
| HT | 24 (19.5%) | 14 (9.7%) | 7 (13%) | 0.6 | 424 (18%) | 111 (13%) | 0.03 | **0.0093** | 1 |
| no | 198 (80.5%) | 131 (90.3%) | 47 (87%) | | 19551 (82%) | 7581 (87%) | | | |
| NA | 99 | 59 | 10 | | | | | | |
| fam-hist | 85 (26.5%) | 55 (27%) | 19 (29.7%) | 0.75 | 714 (9%) | 244 (27%) | 0.65 | 0.63 | 0.56 |
| no | 236 (73.5%) | 149 (73%) | 45 (70.3%) | | 1761(29%) | 673 (73%) | | | |
| stage : 1 | 60 (23.4%) | 41 (24.4%) | 10 (18.2%) | 0.55 | | | | | |
| 2 | 135 (52.7%) | 89 (53%) | 30 (54.5%) | | | | | | |
| 3 | 51 (19.9%) | 31 (18.5%) | 14 (25.5%) | | | | | | |
| 4 | 10 (3.9%) | 7 (4.2%) | 1 (1.8%) | | | | | | |
| 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%) | 112 (80.6%) | 38 (71.7%) | 0.27 | | | | | |
| R | 149 (46.6%) | 27 (19.4%) | 15 (28.3%) | | | | | | |
| 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 | | | | | | |

Similar to Table 1, a list of factors that are not significantly different between the North Cyprus cohort and the BCFR cohort: hormone therapy (HT) and family breast cancer history. Other factors do not have available data in BCFR: tumor stage, whether the patient has other cancers, smoking, left (L) or right (R) breast or both with cancer, housewife or employed, invasive cancer. (note 1) never or former menopausal HT use.
Abbr.: ER; Estrogen Receptor, PR; Progesterone Receptor, HER2; Human Epidermal Growth Factor 2, BCFR; Breast Cancer Family Registry, NC; North Cyprus, BC; Breast Cancer, HT; Hormone Replacement, otherC; other Cancers

to represent high-dimensional data in two or three dimensions. We had previously used this technique in other applications in biology/genomics (14, 15).

In this study, we used three dependent variables (ER, PR, and HER2), five quantitative independent variables (age at diagnosis, age at menarche, number of children, education level, and cancer stage), and ten binary independent variables (left/right breast, menopause, first-degree relative with cancer, other cancer, smoking, hormone therapy, oral contraceptive use, breastfeeding, housewife/employed, and cancer invasiveness). The quantitative variables were standardized to have zero-mean and unit-variance (z-transformation).

Due to high missing rates for age at menarche (33%), hormone therapy (31%), oral contraceptive use (32%), and breastfeeding (33%), we retained only those samples that had information on these factors. This reduced the sample size from 321 to 211 for the t-SNE plot. Other missing data (of much lower missing rate) were imputed for these 211 patients.

Figure 1 shows one run of t-SNE. Because ER, PR, and HER2 are a component of the input, it is not surprising that their values are well partitioned in the plot (e.g., ER+ and ER- samples). It was observed that ER+ samples tended to be PR+, and HER2- and ER- samples tended to be PR- and HER2+. The seven samples with other cancers (including metastasis) formed a distinct cluster from the remaining samples. Although ER, PR, and HER2 values separated in an up-down direction in Figure 1, other factors such as menopause status, breastfeeding, and age appeared to be separated in (not completely) an orthogonal direction.

The nonTurkish Cypriots are marked with different symbols (Turkey-born patients in circles, other foreign-country-born patients in crosses) in the top two rows of Figure 1. There was no evidence indicating that the location of these points in the plot, or their collective features, are highly different from the remaining Turkish-Cypriot samples.

**Distribution of patient factors**: Table 1 shows that our study cohort (n=321) is distinct from the BCFR sample, which consisted of

Cyprus J Med Sci 2020; 5(4): 339-49

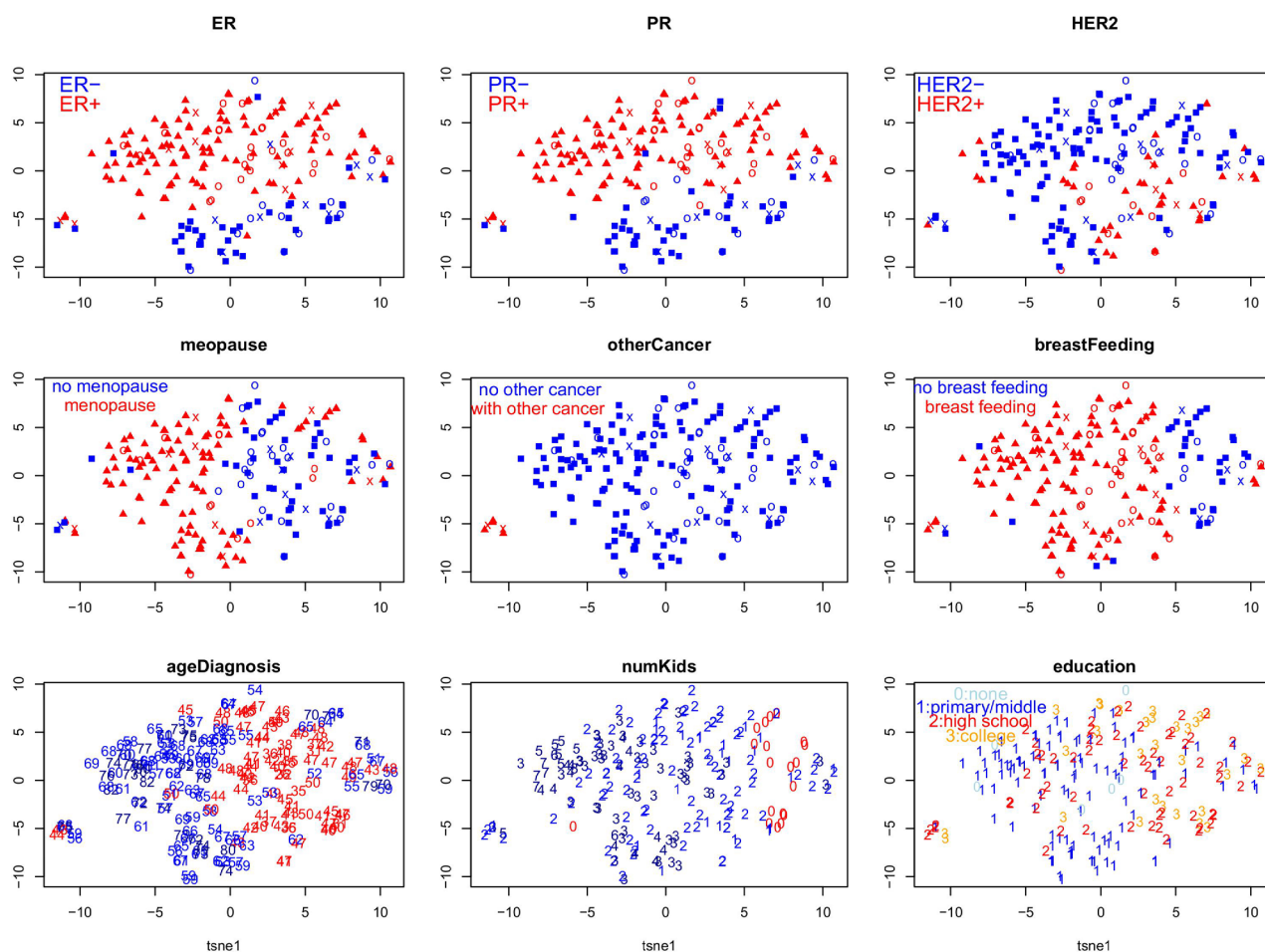Ulgen et al. Breast Cancer Subtypes and Predictive Factors



**FIGURE 1.** t-SNE plot of North Cyprus samples. 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 to 50 years old), parity/number of children, education level (0 for none, 1 for primary or middle school, 2 for high school, 3 for college or higher). The Turkey-born samples are marked with circle, and other foreign born samples are in crosses
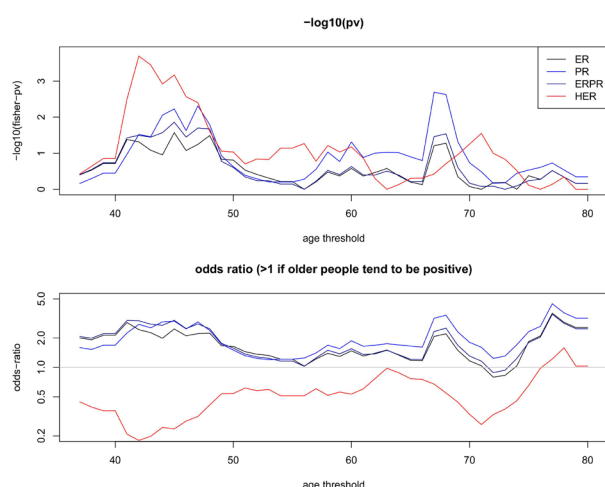


**FIGURE 2.** Tuning age threshold to convert age to age-group. (top) Fisher's p-value (in minus log with base 10) of age-group vs breast cancer subtype (ER, or PR, or ER/PR, or HER2) as a function of the age threshold. (bottom) odds-ratio as a function of the age threshold

subjects of primarily USA/Canada/Australia origin, in terms of several demographic or reproductive factors. The North Cyprus cohort was older in age, had lower education level and less use of oral contraceptives, and had greater number of postmenopausal subjects, lesser number of subjects with young (≤11 years) age at menarche, fewer nulliparous subjects, and greater number of breastfeeding subjects.

The same analysis was conducted on the 233 Turkish-Cypriot-only patients, and the same results were obtained (data not shown).

There are two explanations for these significant differences. The first is due to the cultural and customary differences between countries (e.g.,, use of oral contraceptives). The second explanation is that our study sample was collected from the state hospital, and a higher percentage of affluent patients may opt for treatment at private hospitals or hospitals overseas. The differences remained even for the ER+/PR+ subgroup and for the ER-/PR- subgroup (although less significant due to smaller sample sizes).

Within our North Cyprus cohort, when the ER+/PR+ and ER-/PR- groups were compared in terms of these factors, only the ER-/PR- group was slightly younger (t-test p value=0.06) (Table 1). The differences in other factors were not yet significant, probably because either they were actually not different or there was no sufficient sample size to confirm the difference.

Ulgen et al. Breast Cancer Subtypes and Predictive Factors

Cyprus J Med Sci 2020; 5(4): 339-49

**TABLE 3.** Count and testing of ER, PR, HER2 distributions

**(A) Distribution of breast cancer hormone receptor subtypes in North Cyprus cohort**

| total sample/ Turkish Cypriot N=290/212 | HER2- (n=220/166, 75.9%/78.3%) | | HER2+ (n=70/46, 24.1%/21.7%) | |
|---|---|---|---|---|
| | PR- | PR+ | PR- | PR+ |
| ER- (n=74, 25.5% /n=52, 24.5%) | 40 (triple negative)/32 | 8/5 | 23 (HER2+ only)/14 | 3/1 |
| ER+ (n=216, 74.5% /n=160, 75.5%) | 9/6 | 163/123 | 9/6 | 35/25 |
| | PR- (n=81, 27.9%/n=58, 24.5%) | | | |
| | PR+ (n=209, 72.1%/n=154, 72.6%) | | | |

**(B) Distribution of ER and PR subtypes in BCFR**

| total sample (N=4011) | PR- (n=1317, 32.8%) | PR+ (n=2694, 67.2%) |
|---|---|---|
| ER- (n=1128, 28.1%) | 920 | 208 |
| ER+ (n=2883, 71.9%) | 397 | 2486 |

**(C) Distribution of HER2 subtypes in BCFR**

| total sample (N=792) | HER2- (n=607, 76.6%) | HER2+ (n=185, 23.4%) |
|---|---|---|
| ER- and PR- (n=206, 26%) | 139 (triple negative) | 67 |
| ER+ and/or PR+ (n=586, 74%) | 468 | 118 |

**(D) Fisher test p-value between Cyprus cohort and BCFR/BCAC**

| subtype | BCFR vs n=321 set | BCFR vs Turkish Cypriot n=233 |
|---|---|---|
| ER+ vs ER- | 0.38 | 0.27 |
| PR+ vs PR- | 0.09 | 0.1 |
| ER+/PR+ vs ER-/PR- | 0.81 | 0.61 |
| 4 ER/PR group | 0.08 | 0.04 |
| HER2+ vs HER2 | 0.81 | 0.65 |
| triple-negative vs not | 0.17 | 0.47 |
| HER/(ER-PR- or not) | 0.49 | 0.65 |
| triple-negative vs not (BCAC) | 0.79 | 0.48 |
| HER/(ER-PR- or not) (BCAC) | 0.44 | 0.69 |

(A) Breast cancer subtype counts in the North Cyprus cohort (using either all samples or Turkish-Cypriot-only samples). Hormone receptor-positive (including luminal A and luminal B) consists of 8+9+163=180 counts (62.1%) if all samples are used; and (5 + 6 + 123) / 212=63.2% if Turkish Cypriot samples are used. HER2+ and hormone receptor-positive consists of 3 + 9 + 35=47 counts (16.2%) or (1 + 6 + 25) / 212=15.1% if only Turkish Cypriot samples are used. (B) ER and PR subtype distribution in BCFR (data taken from (17). (C) HER2 subtype distribution in BCFR (data taken from (17). (D) Fisher's test p value of subtype distribution difference between BCFR (or BCAC for the last two rows) and Cyprus cohorts (all and Turkish-Cypriot-only sample).
Abbr.: ER; Estrogen Receptor, PR; Progesterone Receptor, HER2; Human Epidermal Growth Factor 2, BCFR; Breast Cancer Family Registry, BCAC; Breast Cancer Association Consortium

There were also some other factors distributed not very differently between our cohort and the BCFR cohort, as summarized in Table 2. These factors included hormone therapy usage and having a first-degree relative with cancer. The remaining factors shown in Table 2 did not have the information corresponding to the BCFR, including other cancers, smoking status, left/right breast with cancer, housewife/employed, and cancer invasiveness. Only for the ER+/PR+ subtype, the North Cyprus cohort was significantly less likely to have hormone therapy than the BCFR samples.

We also examined the correlation between factors. Using the data of all patients with breast cancer without considering the subtypes, the following correlations were observed: (1) patients who breastfeed are less likely to undergo hormone therapy (OR=7.1, Fisher's p value $9 \times 10^{-5}$), (2) patients who are employed are more likely to smoke than housewives (OR=3.1, p value $1.3 \times 10^{-4}$), and (3) patients who are employed are more likely to be in premenopause than housewives (OR=2.8, p value $1.3 \times 10^{-4}$).

**Conversion of age into age group:** Age is a special factor different from others because it is a continuous variable spanning a wide range of values. Discretizing or categorizing a continuous variable is an involved topic by itself. Age is a well-known target for such categorization (16). We categorized the study sample into younger and older age groups based on an age threshold. When an age threshold is chosen, a 2-by-2 count table can be constructed according to the binary age group and binary breast cancer subtype. Age group versus breast cancer subtype association can be judged using Fisher's test.

Figure 2 shows (-log) Fisher's p value (top) and odds ratio (bottom) as a function of age threshold for converting age into age group. For HER2, the best p value was 0.0002 when the age threshold was 42, and there was a broad range of age threshold (41–47) where the Fisher's test was significant at 0.01 level. For ER, PR, and ER/PR, this age threshold range also led to some significant test results, indicating that patients younger than mid-40s may form a distinct group, which tended to be ER, PR, and ER/PR negative and HER2 positive. At the age threshold of 67–68, there was a second peak, indicating that patients older than that age tended to be ER, PR, and ER/PR positive.

**Distribution of breast cancer subtypes:** We compared the hormone receptor-defined subtype distribution between the North Cyprus cohort and the BCFR cohort as shown in Table 3. The count in each of the eight ER/PR/HER2 subtypes in the n=321 set and the n=233 Turkish-Cypriot-only set is listed in Table 3(A). These counts were not available for the BCFR cohort (17), but the distribution according to ER/PR subtypes and that according to HER2/(ER-PR- and not) subtypes were available, which are reproduced in Table 3(B) and (C).

The Fisher's test for the following subtype groupings was conducted: ER, PR, ER+/PR+ vs ER-/PR-, four ER/PR groups, HER2, triple-negative vs remaining, and the p values are listed in Table 3(D). The lack of a significant difference in subtype distribution between the North Cyprus cohort and the BCFR cohort, indicating certain similarity, is in strong contrast to the dissimilarity of several demographic and reproductive factors as shown in Table 1.

The largest difference was probably in the proportion of PR subtypes (higher PR+ proportion in the North Cyprus cohort than in the BCFR cohort), which may also cause a relatively larger difference for the four ER/PR groups. The highly significant correlation between ER and PR may make PR measurement redundant. In fact, it has been argued that the added value of PR is questionable (18). More specifically, the ER-/PR+ subtype is rare and may not be reproducible (i.e., it can be reclassified into another subtype by another method) (18).

Cyprus J Med Sci 2020; 5(4): 339-49

Ulgen et al. Breast Cancer Subtypes and Predictive Factors

**TABLE 4.** Significant factors in univariate or multiple logistic regression (and the corresponding p-values)

| factor | ER | PR | HER2 | ER | PR | HER2 |
|---|---|---|---|---|---|---|
| | | all-samples | | | Turkish Cypriot | |
| age | 0.024 | 0.025 | 0.0018(-) | | 0.034 | 0.019 |
| menopause | 0.0076 | | 0.0016 (-) | 0.065 | | 0.0023 |
| | 0.08 (multiple) | | 0.08 (multiple) (-) | | | 0.017 (multiple) |
| numkids | | 0.037 | | | | |
| famhist | | | 0.0078 | | | |
| | | | 0.018 (multiple) | | | 0.017 (multiple) |
| smoking | | | | | | 0.017 |

Factors that are significantly (at 0.05 level) related to breast cancer subtypes according to either single variable or multiple logistic regressions using either all samples or Turkish-Cypriot-only samples. The values are single variable logistic regression p values (or multiple variate if marked). "famhist" refers to the presence of any cancer (not necessarily breast cancer) in any first-degree relative.
Abbr. ER;
Abbr.: ER; Estrogen Receptor, PR; Progesterone Receptor, HER2; Human Epidermal Growth Factor 2, famhist; Presence of any cancer (not necessarily breast cancer) in any first-degree relative
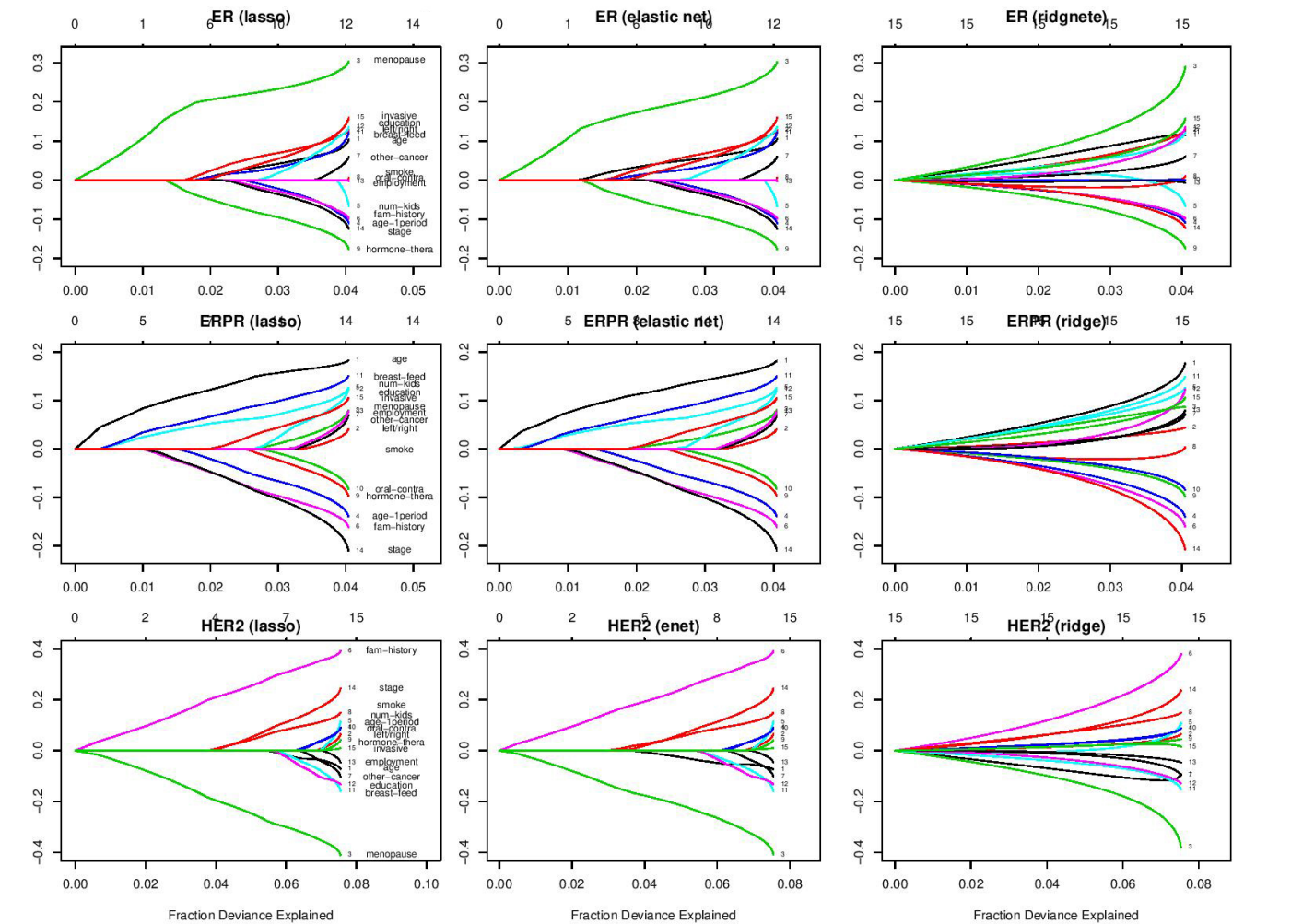


**FIGURE 3.** Regularized regression on ER, ER/PR and HER2. Variable tracing/selecting plot of LASSO, elastic net, ridge logistic regressions (columns 1-3) for ER, ER/PR, and HER2 (rows 1-3). Each line is a factor, and positive x direction represents a more relaxed constraint, allowing more variables. 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+, HER2+ are 1's, ER-, PR-, HER2- are 0's). The x axis is deviance explained by the (regularized) logistic regression

The distribution of breast cancer subtype, defined as the combination of HER2 and ER-PR- (see, e.g.,., Table 3 (C)), was also strikingly similar between our cohort and the cohort of the Breast Cancer Association Consortium (BCAC) conducted in the UK (19), which primarily consisted of Europeans/Caucasians. The subtype information was obtained from (20), and the Fisher's p value for testing triple-negative-only proportion in the two cohorts was not significant (Table 3 (D)). The testing result for the
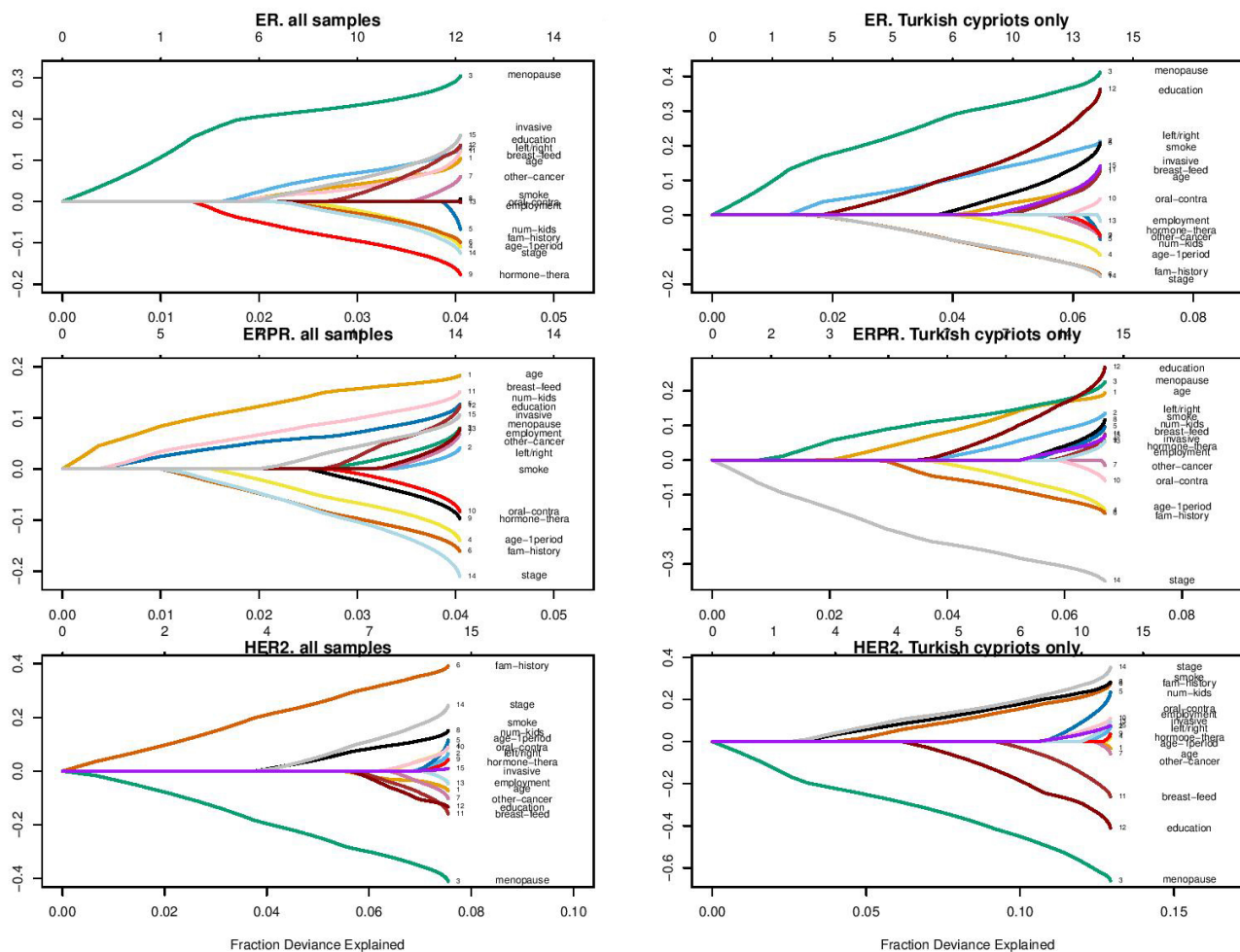
Ulgen et al. Breast Cancer Subtypes and Predictive Factors

Cyprus J Med Sci 2020; 5(4): 339-49



**FIGURE 4.** Comparison of LASSO regression between n=32l and n=233 sets. LASSO regression results for all n=32l patients (left) and for n=233 Turkish Cypriot patients (right), for ER, ER/PR, HER2 (rows 1-3). See Fig. 3 for more explanation of the y and x axes

proportions of four subtypes (triple-negative, HER2+ only, ER+ and/or PR+ and HER2+, and ER+ and/or PR+ and HER2-) was also not significant.

**Predictive factors for breast cancer subtypes**: The comparison of factor values between ER+ or PR+ and ER- or PR- samples can also be cast into a regression of ER/PR (dependent variable) over individual factors (independent variables). Table 4 shows all the results that are significant at the 0.1 level from regressing ER or PR or HER2 by either single factor using univariate logistic regression or all factors using multiple logistic regression, and for the n=32l sample and for the n=233 Turkish-Cypriot-only sample.

As shown in Table 4, age correlated positively with ER+ and PR+ but negatively with HER2+. These results are similar to those shown in Table I as well as Figure 2, where ER+PR+ patients are older in age. Due to the positive correlation between ER and PR, ER+ patients and HER2- patients are older in age.

Table 4 also shows that postmenopause status correlated positively with ER+ but negatively with HER2+. It can be stated that menopause plays a protective role as postmenopause patients are more likely to be in the more curable ER+ subtype and less likely to have the worse prognosis ER- type (2I). The positive correlation between menopause status and age is self-explan-

atory, and the association between menopause status and ER or HER2 can also be easily explained by age. The last three minor conclusions are that HER2+ patients are more likely to have a first-degree relative with cancer, PR+ patients tend to have more children, and HER2+ patients tend to be smokers.

Between univariate and multiple regressions, we also applied three closely related regularized regressions, LASSO, elastic net, and ridge (22), to analyze the situation with a few explanatory variables. The regularized regressions accomplished the task of variable selection [e.g., (23, 24)] by imposing constraint on the sum of the absolute value of all fitting coefficients, effectively setting several coefficients to be zero, thus removing the contribution from these variables. Figure 3 shows how the coefficient of each explanatory variable increases, from left to right, when the number of nonzero-coefficient variables increases, for (rows) the dependent variables of ER, ER/PR, and HER2, and for (columns) LASSO, elastic net, and ridge.

First, we observed that LASSO, elastic net, and ridge regression showed a similar trend as that of the breast cancer subtype. Therefore, we might focus only on the LASSO plot, which is shown in the first column in Figure 3. Second, to observe how the n=32l dataset may differ from the n=233 Turkish-Cypriot-only dataset, we plotted the LASSO results side by side in Figure 4.

Cyprus J Med Sci 2020; 5(4): 339-49

Ulgen et al. Breast Cancer Subtypes and Predictive Factors

For the ER subtype, the dominant contribution from menopause status was not only consistently observed for both the n=321 and n=233 datasets, but it was also consistent with Table 4. Hormone therapy was a promising signal for the n=321 dataset but not for the n=233 dataset. For ER/PR, the cancer stage was a consistent signal that negatively correlated with the ER+/PR+ status. This signal was also emphasized in the BCFR article (17). Breastfeeding was a promising signal for the n=321 dataset, but it appeared to be less important for the n=233 dataset. Further discussion on the benefits of breastfeeding in reducing the probability of acquiring poor prognostic breast cancers, such as triple-negative subtypes, has been described previously (25, 26). For the HER2 subtype, similar to univariate and multiple regressions (Table 4), menopause status was a dominant factor with negative correlation. Interestingly, there was a signal for the n=321 dataset from first-degree-relative cancer history, but that signal became weaker for the n=233 dataset.

## DISCUSSION
Without using control samples, we conducted a case-only analysis of potential predictive factors for different subtypes of breast cancer. The case-only design has been implemented in previous breast cancer studies, and it is "an important initial step in understanding the extent of etiologic heterogeneity between tumor subtypes" (10, 11). Because different subtypes of breast cancer have different prognoses, it is important to evaluate their distribution.

One of the striking results we obtained in this study was that our North Cyprus cohort had highly similar ER+, ER+/PR+, and HER2+ status as that of the BCFR cohort, even though the our cohort was much older in age and had greater number of postmenopausal subjects, lower education status, less use of hormone therapy, and greater number of breastfeeding subjects. If older age/postmenopausal patients tend to have ER+, our older cohort should have a higher proportion of the ER+ subtype than the BCFR cohort. Although it was in fact the case (74.5% ER+ in the n=321 North Cyprus dataset and 71.9% ER+ in the BCFR dataset), the difference in the underlying factors (age or menopause status) was highly significant between the two cohorts, but it was not significant in the ER distributions.

This aspect can be discussed in general terms, i.e., can the correlation at one level be translated into correlation at another level? In our study, we examined the potential similarity/dissimilarity of the distribution of a factor in two datasets (low-level), and contemplated whether it can be translated into the similarity/dissimilarity of the distribution of a subtype affected by these predictive factors (high-level) in those two datasets. In our previous investigation of a very different issue, i.e., the linkage/association analysis of multiple correlated phenotypes in a lipid panel, we had observed that the correlation at the high-level (phenotype) does not necessarily translate into a correlation at the low-level (linkage disequilibrium or colocalization between genetic variants) (27).

The causal link between the two levels could also be not sufficiently strong to translate correlation from one level to another. In our LASSO analysis (Figure 3, 4), it can be observed that the fraction of the explained deviation (range of x-axis) of ER, ER/PR, and HER2 is at the most a few percentage, even using all

factors. Random forest run on the same data also showed that the classification rate in terms of the ER or ER/PR or HER2 status was not high, i.e., on average, it was scarcely >50% (results not shown). This highlights the fact that several true predictive factors for breast cancer subtypes are not yet included in our data (e.g.,, body mass index (BMI)), and moreover, the known genetic causes of breast cancer (e.g.,, BRCA1 and BRCA2) are not part of the analysis.

In a recent systematic meta-analysis of African breast cancer subtypes (28), it was found that the proportion of ER+ and PR+ samples fluctuated significantly from study to study. There are also data indicating that the triple-negative subtype rate is much higher in African women than in European/Caucasian women (29, 30). To double-check whether the breast cancer subtype distribution in our North Cyprus cohort was still the same as in another study, we selected a published summary statistics from a southeastern Turkish cohort (31). The ER+, ER+/PR+, and HER2+ proportions in the Turkish cohort were 73.5%, 81.8%, and 30.4% compared to the proportions of 74.5%, 75.9%, and 24.1% in our North Cyprus cohort, respectively, leading to Fisher's test p values of 0.8, 0.086, and 0.076, respectively (number of samples in the Turkish cohort with the subtype information: 438, 437, and 434, respectively). These differences are within the ranges and are not significant.

It could be of great interest to compare our breast cancer subtypes statistics with those of a Greek Cypriot cohort. We found two surveys on breast cancer in Greek Cypriots, one with 1100 patients conducted from 1999 to 2005 (32) and another with more than 4000 patients conducted from 2005 to 2013 (33). Unfortunately, there was no hormone receptor subtype information for the Greek Cypriot cohorts in both time periods. However, we could compare the distribution of other factors when the comparable data are available. We found that for the 1995–2005 period, the distribution of children was almost identical between the Greek and Turkish Cypriot cohorts; age and smoking status were not significantly different; and Greek Cypriot patients had higher education level (p value=$1 \times 10^{-8}$), less incidence of family breast cancer history (p value=$5 \times 10^{-7}$), more early (age ≤11 years) menarche (p=0.04), less breastfeeding (p=0.02), and less use of oral contraceptives (p=0.004). The Greek Cypriots 2005-2013 cohort had more patients with invasive type of breast cancer compared to the Turkish Cypriots (p-value=$1E-14$), but level of smoking status was comparable (not significant).

Our regularized regressions (LASSO, elastic net, and ridge) (Figure 3, 4) revealed potential predictive factors for breast cancer subtypes. However, these weak signals can only be considered as a "trend" that has not yet been confirmed by statistical tests, as those shown in Table 4. In a previous study (34), the risk of benign disease proliferation was found to be higher in patients with the ER+ subtype than in patients with the ER- subtype. This finding can be compared to the positive contribution observed in our study from other cancers (including metastasis) and the ER+ or ER+/PR+ subtype (Figure 4). In another previous research (17), breastfeeding was not associated with the ER-/PR- subtype, which can be compared with our result that breastfeeding positively correlated with the ER+/PR+ subtype. In the study conducted by (21), it was observed that the ER- cancer rate stopped increasing at a certain age, whereas the ER+

Ulgen et al. Breast Cancer Subtypes and Predictive Factors

Cyprus J Med Sci 2020; 5(4): 339-49

rate continued to increase. This observation can be compared to our result that postmenopause positively correlated with the ER+ subtype (Figure 4). Colditz et al. (35) reported significant differences in age, menopause status, and past use of hormone therapy in four ER/PR groups. Yang et al. (36) found that early age at menarche (≤12 years) was less common in the PR- group than in the PR+ group, and this was also true in our data comparing the ER-/PR- and ER+/PR+ groups. To summarize, many of the predictive factors for breast cancer subtypes observed in our study are consistent with those reported in the literature. The positive correlation observed between cancer family history and HER2+ subtype (Figure 4) remains intriguing.

In conclusion, we used a unique cohort of breast cancer in an understudied population to examine the breast cancer subtypes and related factors. A simplified analysis framework was used, keeping the breast cancer subtypes at one level and all factors at another level. The distribution of several factors was extremely different from that of another large breast cancer registry, whereas the subtype distribution was similar. This indirectly shows that we have not exhaustively measured all the predictive factors for breast cancer subtypes. The relationship between the two levels was investigated by regression using one variable, all variables, or a subset of variables. These regression analyses indicated that postmenopause and/or older age patients with breast cancer are more likely to have the ER+ subtype and the HER2- subtype. We also observed several other trends that need to be validated in a larger cohort.

## REFERENCES

1. Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, et al. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. Euro J Cancer 2018; 103: 356-87. [Crossref]

2. McPherson K, Steel CM, Dixon JM. Breast cancer-epidemiology, risk factors, and genetics. BMJ 2000; 321(7261): 624-8. [Crossref]

3. Möller S, Mucci LA, Harris JR, Scheike T, Holst K, Halekoh U, et al. The heritability of breast cancer among women in the Nordic twin study of cancer. Cancer Epidemiol Biomarkers Prev 2016; 25(1): 145-50. [Crossref]

4. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. Am J of Hum Genet 2019; 104(1): 21-34.

5. Shieh Y, Scott CG, Jensen MR, Norman AD, Bertrand KA, Pankratz VS, et al. Body Mass Index, mammographic density, and breast cancer risk by estrogen receptor subtype. Breast Cancer Res 2019; 21(1): 48. [Crossref]

6. Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genet Med 2019; 21(8): 1708-18. [Crossref]

7. Fisher ER, Redmond CK, Liu H, Rockette H, Fisher B. Correlation of estrogen receptor and pathologic characteristics of invasive breast cancer. Cancer 1980; 45(2): 349-53. [Crossref]

8. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest 2011; 121(7): 2750-67. [Crossref]

9. Özmen V, Özmen T, Doğru V. Breast cancer in Turkey; an analysis of 20,000 patients with breast cancer. Euro J Breast Health 2019; 15(3): 141-6. [Crossref]

10. Martinez ME, Cruz GI, Brewster AM, Bondy ML, Thompson PA. What can we learn about disease etiology from case-case analyses? Lessons from breast cancer. Cancer Epidemiol Biomarkers Prev 2010; 19(11): 2710-4. [Crossref]

11. Redondo CM, Gago-Dominguez 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): e40543. [Crossref]

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

13. Van der Maaten LJ, Hinton GE. Visualizing data using t-SNE. J. Machine Learning Res 2008; 9: 2575-605.

14. Li W, Cerise JE, Yang Y, Han H. Application of t-SNE to human genetic data. J Bioinform Comput Biol 2017; 15(4): 1750017. [Crossref]

15. Li W, Freudenberg J, Freudenberg J. Alignment-free approaches for predicting novel Nuclear Mitochondrial Segments (NUMTs) in the human genome. Gene 2019; 691: 141-52. [Crossref]

16. Switf HJ, Abrams D, Drury L, Lamont RA. Categorization by age. Encyclopedia of Evolutionary Psychological Science. Shackelford T, Weekes-Shackelford V, editors. Springer; 2018. [Crossref]

17. Work ME, John EM, Andrulis IL, Knight JA, Liao Y, Mulligan AM, et al. Reproductive risk factors and oestrogen/progesterone receptor-negative breast cancer in the Breast Cancer Family Registry. Br J Cancer 2014; 110(5): 1367-77. [Crossref]

18. Hefti MM, Hu R, Knoblauch NW, Collins LC, Haibe-Kains B, Tamimi RM, et al. Estrogen receptor negative/progesterone receptor positive breast cancer is not a reproducible subtype. Breast Cancer Res 2013; 15(4): R68. [Crossref]

19. Breast Cancer Association Consortium, Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium. J Natl Cancer Inst 2006; 98(19): 1382-96. [Crossref]

20. Brouckaert O, Rudolph A, Laenen A, Keeman R, Bolla MK, Wang Q, et al. Reproductive profiles and risk of breast cancer subtypes: a multi-center case-only study. Breast Cancer Res 2017; 19(1): 119. [Crossref]

21. Tarone RE, Chu KC. The greater impact of menopause on ER- than ER+ breast cancer incidence: a possible explanation. Cancer Causes Control 2002; 13(1): 7-14. [Crossref]

22. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Second Edition; 2009.

Cyprus J Med Sci 2020; 5(4): 339-49

Ulgen et al. Breast Cancer Subtypes and Predictive Factors

23. Halinski RS, Feldt LS. The selection of variables in multiple regression analysis. J Edu Measurement 1970; 7: 151-7. **[Crossref]**

24. Li W, Yang Y. How many genes are needed for a discriminant microarray data analysis. Lin SM, Johnson KF, editors. Methods of Microarray Data Analysis. Kluwer Academic Publishers, 2002; pp.137-149. **[Crossref]**

25. Islami F, Liu Y, Jemal A, Zhou J, Weiderpass E, Colditz G, et al. Breastfeeding and breast cancer risk by receptor status -- a systematic review and meta-analysis. Ann Oncol 2015; 26(12): 2398-407. **[Crossref]**

26. Fortner RT, Sisti J, Chai B, Collins LC, Rosner B, Hankinson SE, et al. Parity, breastfeeding, and breast cancer risk by hormone receptor status and molecular phenotype: results from the Nurses' Health Studies. Breast Cancer Res 2019; 21(1): 40. **[Crossref]**

27. Ulgen A, Han Z, Li W. Correlation between quantitative traits and correlation between corresponding LOD scores: detection of pleiotropic effects. BMC Genet 2003; 4(Suppl 1): S60. **[Crossref]**

28. Eng A, McCormack V, Dos-Santos-Silva I. Receptor-defined subtypes of breast cancer in indigenous populations in Africa: a systematic review and meta-analysis. PLoS Med 2014; 11(9): e1001720. **[Crossref]**

29. Huo D, Ikpatt F, Khramtsov A, Dangou JM, Nanda R, Dignam J, et al. Population differences in breast cancer: survey in indigenous African women reveals over-representation of triple-negative breast cancer. J Clin Oncol 2009; 27(27): 4515-21. **[Crossref]**

30. Zheng Y, Walsh T, Gulsuner S, Casadei S, Lee MK, Ogundiran TO, et al. Inherited breast cancer in Nigerian women. J Clin Oncol 2018; 36(28): 2820-5. **[Crossref]**

31. Kuzhan A, Adli M, Eryigit Alkis H, Caglayan D. Hormone receptor and HER2 status in patients with breast cancer by races in southeastern Turkey. J BUON 2013; 18(3): 619-22.

32. Hadjisavvas A, Loizidou MA, Middleton N, Michael T, Papachristoforou R, Kakouri E, et al. An investigation of breast cancer risk factors in Cyprus: a case control study. BMC Cancer 2010; 10: 447. **[Crossref]**

33. Pilavaki P, Giallouros G, Yiallourou AI, Pantavou K, Marcou Y, Demetriou A, et al. Epidemiology of breast cancer in Cyprus: Data on newly diagnosed cases and survival rates. Data Brief 2018; 19: 353-69. **[Crossref]**

34. Kerlikowske K, Gard CC, Tice JA, Ziv E, Cummings SR, Miglioretti DL. Breast Cancer Surveillance Consortium. Risk factors that increase risk of estrogen receptor-positive and -negative breast cancer. J Natl Cancer Inst 2016; 109(5): djw276. **[Crossref]**

35. 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. **[Crossref]**

36. Yang XR, Chang-Claude J, Goode EL, Couch FJ, Nevanlinna H, Milne RL, et al. Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. J Natl Cancer Inst 2011; 103(3): 250-63. **[Crossref]**

37. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci 2014; 1(3): 140216. **[Crossref]**