

# Prediction of Pathologic Complete Response to Neoadjuvant Chemotherapy Using Machine Learning Models in Patients with Breast Cancer

**Ji-Yeon Kim**

Samsung Medical Center

**Eunjoo Jeon**

Samsung SDS

**Soonhwan Kwon**

Samsung SDS

**Hyungsik Jung**

samsung SDS

**Sunghoon Joo**

Samsung SDS

**Youngmin Park**

Samsung SDS

**Se Kyung Lee**

Samsung Medical Center

**Jeong Eon Lee**

Samsung Medical Center

**Seok Jin Nam**

Samsung Medical Center

**Eun Yoon Cho**

Samsung Medical Center

**Yeon Hee Park**

Samsung Medical Center

**Jin Seok Ahn**

Samsung Medical Center

**Young-Hyuck Im** (✉ [imyh00@skku.edu](mailto:imyh00@skku.edu))

Samsung Medical Center <https://orcid.org/0000-0001-6459-8118>

---

## Research Article

**Keywords:** Breast cancer, neoadjuvant chemotherapy, machine learning, pathologic complete response

**Posted Date:** February 12th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-217080/v1>

**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published at Breast Cancer Research and Treatment on July 5th, 2021. See the published version at <https://doi.org/10.1007/s10549-021-06310-8>.

# Abstract

## Background

The aim of this study was to develop a machine learning (ML) based model to accurately predict pathologic complete response (pCR) to neoadjuvant chemotherapy (NAC) using pretreatment clinical and pathological characteristics of electronic medical record (EMR) data in breast cancer (BC).

## Methods

The EMR data from patients diagnosed with early and locally advanced BC and who received NAC followed by curative surgery were reviewed. A total of 16 clinical and pathological characteristics was selected to develop ML model. We practiced six ML models using default settings for multivariate analysis with extracted variables.

## Results

In total, 2,065 patients were included in this analysis. Overall, 30.6% (n=632) of patients achieved pCR. Among six ML models, the LightGBM had the highest area under the curve (AUC) for pCR prediction. After hyper-parameter tuning with Bayesian optimization, AUC was 0.810. Performance of pCR prediction models in different histology-based subtypes was compared. The AUC was highest in HR+HER2- subgroup and lowest in HR-/HER2- subgroup (HR+/HER2- 0.841, HR+/HER2+ 0.716, HR-/HER2 0.753, HR-/HER2- 0.653).

## Conclusions

A ML based pCR prediction model using pre-treatment clinical and pathological characteristics provided useful information to predict pCR during NAC. This prediction model would help to determine treatment strategy in patients with BC planned NAC.

## Introduction

Breast cancer (BC) is the most common cancer in women worldwide, and 2.1 million women were newly diagnosed with BC in 2018 [1]. While BC specific mortality has gradually decreased, it still causes the highest portion of cancer mortality in females [1, 2]. Because most BC specific mortality occurs with metastatic breast cancer (MBC) [2], early detection of BC recurrence and effective treatment to eradicate metastasis are the most accurate strategies to reduce BC mortality.

Neoadjuvant chemotherapy (NAC) has been performed in patients with early and locally advanced BC (LABC). NAC is used to reduce tumor size in inoperable cases and eliminate micrometastasis [3]. In addition, response to NAC is one of the most powerful surrogate markers to predict BC prognosis [4]. Pathologic complete response (pCR), defined as no residual tumor cells in either breast or axilla after

NAC, represents long term survival without BC recurrence [5]. In addition, residual cancer burden (RCB) score, based on residual tumor volume after NAC, can accurately predict BC prognosis [6].

Clinical trials have indicated that human epidermal growth factor receptor 2 (HER2) positive BC had about 60–70% pCR for NAC with dual anti-HER2 blockade [7, 8], while triple negative breast cancer (TNBC) achieved about 30% pCR after anthracycline and taxane based NAC [9]. These clinical trials aid in determining overall patient prognosis but do not clarify the individual fates of patients with BC who receive NAC. There is also insufficient knowledge of baseline characteristics that affect NAC response at the time of BC diagnosis, regardless of BC subtype.

Current machine learning (ML) methods improve the accuracy of BC diagnosis and help to predict BC recurrence and prognosis [10–12]. If precise prediction of NAC response is possible using a ML algorithm, it may also be used to predict NAC responses of individual patients. This may change treatment plans for LABC during NAC and maximize the pCR rate and improve patient outcome. Therefore, we aimed to develop ML based model to accurately predict pCR to NAC using pretreatment clinical and pathological characteristics of electronic medical record (EMR) data in a neoadjuvant BC cohort in this study.

## Methods

### Study population

In this retrospective study, clinical information of 2,168 patients with BC treated with NAC followed by curative surgery at Samsung Medical Center between January 2010 and September 2018 was collected. Among 2,168 patients, 103 were excluded: 49 who had bilateral synchronous BC, 52 who did not receive curative surgery due to patient refusal, 1 who did not have complete surgical pathology results, and 1 who did not have information regarding NAC regimen. The Institutional Review Board (IRB) of Samsung Medical Center, Seoul, Korea (IRB No. 2019-04-021) reviewed and approved this study with an informed consent waiver due to use of retrospective clinical data. This study was performed in accordance with the Declaration of Helsinki.

### Measurements

Detailed information of 16 baseline characteristics of patients with BC was obtained. Age, body mass index (BMI), menopausal status, smoking status, comorbidities at BC diagnosis, baseline CEA and CA-15-3, clinical and pathologic T and N stages, NAC regimen, and information regarding pathologic diagnosis, including immunohistochemistry (IHC) for estrogen receptor (ER), progesterone receptor (PR), HER2, and Ki-67.

BMI was calculated as weight in kilograms divided by height in meters squared at the time of BC diagnosis. Women were considered post-menopausal if they were age <60 and amenorrheic for  $\geq 12$  months in the absence of chemotherapy, tamoxifen, or ovarian suppression; had undergone prior bilateral oophorectomy; or were aged  $\geq 60$  years [13]. Pathologic and clinical stages were based on criteria of the

American Joint Committee on Cancer, 7<sup>th</sup> Edition [14]. Two experienced pathologists reviewed and determined primary tumor characteristics based on biopsy and surgical specimens. Primary tumor characteristics based on size, axillary nodal status, and resection margin were determined using the surgical specimen. Morphologic diagnosis and receptor status (ER, PR, and HER2) were determined by IHC staining. ER positivity and PR positivity were defined as Allred scores of 3–8 based on IHC staining with antibodies against ER (Immunotech, France) and PR (Novocastra, UK), respectively. Hormone receptor (HR) positivity was defined as ER or PR positivity. HER2 status was evaluated using the appropriate antibody (Dako, CA) and/or silver in situ hybridization (SISH). HER2 grades 0 and 1 indicated a negative result, while grade 3 indicated a positive result. Amplification of HER2 was confirmed by SISH for results of 2+.

BC was further categorized into four subtypes according to IHC results: HR+/HER2-, HR+HER2+, HR-HER2+, and TNBC, defined as BC with negative ER and PR expression and lack of HER2 overexpression [15].

### **Data Pre-processing**

Menopause status was divided into premenopausal or postmenopausal, and smoking status was into current/ex-smoker or never-smoker. Serum CEA and CA-15-3 were used as continuous variables. Morphology was defined as invasive ductal carcinoma (IDC) or non-IDC. ER and PR ranges were between 0 and 8, and Ki-67 range was between +1 and +4. HER2 was handled as a binary variable with negative or positive results. The chemotherapy regimen of each patient was reconstructed into six subcategories and reviewed by an oncologist: 1) anthracycline based chemotherapy (AC: doxorubicin plus cyclophosphamide) followed by taxane (T), 2) chemotherapy with trastuzumab as an anti-HER2 targeted agent, 3) chemotherapy with dual anti-HER2 blockade (combination of trastuzumab and pertuzumab), 4) AC followed by T with platinum, 5) AC alone, and 6) others. pCR was defined as absence of residual invasive cancer on pathologic evaluation of the resected breast specimen and all sampled regional lymph nodes (ypT0/isN0).

### **Feature selection**

Features were selected by univariate and correlation analyses. Univariate analysis was conducted to identify significant differences between the pCR and non-pCR groups using t-test after testing for homogeneity of variance (Levene's test) or chi-square test. If similar features were associated with Pearson correlation and CramerV > 0.5, the feature with the higher effect size (only) was included to minimize collinearity among features. If variables were not significant in univariate analysis and the correlations were high, other features were additionally selected after expert review of the hospital workflow.

### **Model selection**

Six ML models were compared using default settings for multivariate analysis with extracted variables: 1) Logistic regression (Logistic Regression function in Python Scikit-Learn library (version 0.21.2)), 2) Lasso regression (Lasso function in Scikit-Learn), 3) SVM-linear (SVC function in Scikit-Learn), 4) SVM-rbf (SVC function in Scikit-Learn), 5) Random Forest (RandomForestClassifier function in Scikit-Learn), and 6) LightGBM (lgb function in lightgbm).

### **Model development and statistical analysis**

Since missing data comprised less than 10% of the total data, a simple imputation technique was used. Missing data were imputed using mean for continuous features and median for categorical features. Stratified 10-fold cross-validation (StratifiedKFold function in Python Scikit-Learn) was used to prevent overfitting and generate comparison results with lower bias and lower variance [16]. In 10-fold cross-validation, the entire dataset was divided into 10 mutually exclusive subsets with approximately the same class distribution as the original dataset (stratified). The dataset (2065 records) was randomly divided into 10 disjoint subsets (folds), with each fold containing approximately the same number of records (208 records). Sampling was stratified by class label to ensure the subset proportions were the same as the original set. For each fold, a classifier was trained using 9 of the 10 folds and was tested on the 10<sup>th</sup> one.

In the training phase, Bayesian optimization using BayesOpt [17] was employed with 100 iterations to select hyperparameters of LightGBM: number of leaves, feature fraction, bagging fraction, max depth, lambda1, lambda2, min split gain, and min child weight. After choosing the best parameters, the LightGBM model was trained at learning rate 0.01 for 2000 iterations using an early stopping round of 100. The performance of classification models was evaluated by four standard metrics, which provide better results than accuracy evaluation for unbalanced data sets. The four standard metrics were area under the receiver operation characteristics (ROC) curve (AUC), accuracy, sensitivity, and specificity.

To avoid feature importance bias, a weakness of the mean decrease in importance, the permutation importance of features was calculated. This indicates the difference between baseline and the decrease in overall accuracy caused by permuting the column [18].

## **Results**

### **Baseline characteristics**

In total, this study included 2065 patients diagnosed with early or locally advanced BC treated with NAC followed by curative surgery (Table S1). Data on every characteristic were collected for all patients except baseline serum CEA (null counts: n=35, 1.7%), CA-15-3 (n=30, 1.4%), and ER score (n=1, 0.1%). Median age at BC diagnosis was 46.6 years, and 65.5% of patients were premenopausal. In terms of subtypes, there were 643 (31.1%) cases of HR+/HER2- BCs, 343 (16.6 %) cases of HR+/HER2+ BCs, 412 (19.9%) cases of HR-/HER2+ BCs, and 667 (32.3%) cases of TNBC. Clinical stage I and II comprised 35.8% of

cases, while stage III comprised 66.2% of cases. The pCR rates were 30.6% in the entire population, 10.6% in HR+HER2- group, 39.9% in HR+HER+ group, 45.4% in HR-HER2+ group, and 30.3% in TNBC group.

## **Feature selection**

Univariate analyses of the relationship between baseline patient characteristics and pathologic response to NAC were performed (Table 1). Baseline patient characteristics of age, menopausal status, baseline CA-15-3, Allred scores of ER and PR, HER2 status, expression level of Ki-67, clinical stage, T and N stages, and NAC regimens significantly affected pCR status ( $p < 0.05$ , respectively). Among these factors, age and clinical stage were excluded due to interference with other factors: age and menopausal status (correlation coefficient: 0.75) and clinical stage and N stage (correlation coefficient: 0.92) (Figure S1). Both ER and PR (correlation coefficient: 0.84) were clinically important and were included for further analysis. Lastly, CEA and histology were added for further analysis after literature review, even though these two factors were not significantly associated with pCR [19-21]. Therefore, 11 features were selected: menopausal status, CEA, CA-15-3, histology, ER score, PR score, HER2 status, Ki-67, T stage, N stage, and NAC regimen.

## **Multivariate analysis with machine learning model**

Six ML models were tested using 11 selected features (Figure 1A). Logistic regression (LR), LR with L1 penalty, linear support vector machine (SVM-L), radial basis function SVM, random forest, and LightGBM were used for these tests. Among these six models, LightGBM had the highest performance with an AUC of 0.78 compared with other models for pCR prediction in one-fold evaluation. Therefore, further multivariate analysis was performed using the LightGBM model. In multivariate analysis, AUC increased from 0.7845 to 0.810 as an average of 10-fold results after hyper-parameter tuning with Bayesian optimization (Figure 1B, Table 2). Detailed hyper-parameters of LightGBM were described in Supplementary Table S2.

## **Weight of feature importance affecting the machine learning model**

Permutation feature importance was performed to identify features that significantly affect pCR. This analysis identified seven features that significantly affected pCR, and NAC regimen was the most contributing feature with 0.26 AUC drop (-0.26 0.033). Other features affecting pCR were ER score (-0.04 0.010), N stage (-0.02 0.010), T stage (-0.02 0.011), and Ki-67 (-0.01 0.007). Menopausal status, histology, PR score, and HER2 did not change AUC in this ML model.

## **Machine learning model for pCR prediction according to BC subtype**

Further analyses to establish a pCR prediction model according to BC subtype were performed. The BC cohort consisted of 643 (31.1%) HR+/HER2- BCs, 343 (16.6%) HR+/HER2+ BCs, 412 (19.9%) HR-/HER2+ BCs, and 667 (32.3%) TNBCs. The pCR rate according to subtype was 10.6% for HR+HER2- BCs, 39.9% for HR+HER+ BCs, 45.4% for HR-HER2+ BCs, and 30.3% for TNBCs. AUC, sensitivity, and specificity of pCR according to BC subtype are described in Table S4. Among the four subtypes, HR+/HER2- BCs had the

highest AUC (0.841) (0.716 for HR+/HER2+ BCs, 0.753 for HR-HER2+ BCs, and 0.653 for TNBCs) (Figure 3).

Permutation feature importance was also performed according to BC subtype (Figure 4). In HR+HER2- BCs, AUC changes were observed for CEA (-0.08 0.053), ER score (-0.07 0.017), CA-15-3 (-0.05 0.033), Ki-67(-0.03 0.020), and PR score (-0.03 0.016). For HR+HER2+ BCs, PR score intensely decreased AUC (-0.20 0.054), while NAC regimen (-0. 0.025), menopausal status (-0.0 0.031), N stage (-0.05 0.044), Ki-67 (-0.0 0.021), CEA (-0.03 0.010), and ER score (-0.0 0.031) affected pCR status. Interestingly, N stage was the most contributing feature to pCR (AUC change: -0.19 0.027) in HR-HER2+ BCs, followed by NAC regimen (-0.08 0.024), serum CA-15-3(-0.02 0.012), T stage (-0.02 0.015), and serum CEA (-0.01 0.005). For TNBC, NAC regimen, T stage, and Ki-67 intensely affected pCR (AUC change: -0.07 0.037, -0.07 0.054, and -0.05 0.016, respectively).

## Discussion

We developed a ML model to predict pCR after NAC in patients with BC using clinicopathologic characteristics. Moreover, a pCR prediction model was also developed according to BC subtype. This ML model suggested that NAC regimen was the most significantly contributing feature to predict pCR regardless of BC subtype. Features affecting pCR differed by BC subtype. Baseline serum CEA in HR + HER2- BCs, PR score in HR + HER2 + BCs, clinical N stage in HR-HER2 + BCs, and NAC regimen in TNBC were the most significantly contributing factors among all analyzed clinical features.

Advances of chemotherapeutic agents was the most important factor affecting NAC response irrespective of BC subtype. In the era of cytotoxic agents, patients with BC treated with anthracycline and taxane have high pCR rate and survival outcome compared to those treated with anthracycline alone [22, 23]. While the effect of additional carboplatin in the NAC regimen remains controversial, some patients with TNBC experience a survival benefit from platinum based therapeutic agents [24]. Our study suggested that NAC regimen affected pCR rate except in HR + HER2- BC subtype. In particular, HR-HER2 + BCs were most influenced by NAC regimen. Recent advances in anti-HER2 antibodies have remarkably improved the clinical outcomes of patients with HER2 + BCs. The addition of trastuzumab, an anti-HER2 antibody, to chemotherapy increased pCR rate by about two times compared to chemotherapy alone in HER2 + BCs [25]. Moreover, dual anti-HER2 blockade with chemotherapy achieved an about 60% pCR rate in HER2 + BCs [8]. Therefore, the result of our research would reflect the impact of advances of NAC regimen on pCR rate in real world practice.

Most studies of associations between baseline characteristics and NAC response showed BC subtypes were mostly associated pathologic characteristics to pCR. Among BC subtypes, HER2 + BCs and TNBCs had higher pCR rates compared with HR+/HER2- BC subtype [26]. Despite a pCR rate of about 10%, we have treated patients with HR+/HER2- LABC with NAC in order to decrease tumor size being operable and eliminate micrometastasis [27]. In addition, pCR is a surrogate marker of survival regardless of BC subtype.

Therefore, another approach for predicting NAC response by BC subtype is urgently needed to enable precise treatment of BC patients with NAC.

We developed a ML model for pCR prediction according to BC subtype. This pCR prediction model showed that accuracy and baseline clinical factors affecting pCR varied by BC subtype. In terms of TNBC, slightly low accuracy of our prediction model was observed. TNBC consists of several TNBC subtypes based on microarray data with different fates but this information could not be integrated into our model [28]. Therefore, further deep learning models should be developed using additional pathologic information to precisely predict NAC response, especially TNBC group.

Baseline serum CEA and CA-15-3 affected pCR status in HR + HER2-BCs. Baseline serum CEA and CA-15-3 are considered prognostic factors in patients with BC who undergo curative surgery, irrespective of BC subtype [20, 29]. A previous study showed that the prognosis of ER + or HER2- BCs was related to levels of both CEA and CA-15-3, while prognosis of ER- or HER2 + BCs was related to the level of CA-15-3 [29]. In our study, CEA level was associated with HR+/HER2- and HR+/HER2 + subtypes, while CA-15-3 was associated with HR-HER2 + subtype.

Currently, the American Society of Clinical Oncology (ASCO) does not recommend the use of CEA and CA-15-3 to screen, diagnose, stage, or monitor treatment in BC due to low sensitivity and specificity [30]. However, we suggested that pCR might be associated with baseline tumor markers in different BC subtypes, and serum tumor markers would be tested at baseline.

ER and PR IHC scores affected pCR in HR + BCs regardless of HER2 status. ER score had more significant influence than PR score in HR+/HER2- BCs, but PR score had more significant influence on pCR in HR+/HER2 + BCs. A previous study analyzing NAC response suggested that BCs with low ER score had similar pCR rate to ER negative BCs [31]. In addition, IHC4 score determined for ER, HER2, and Ki-67 status in pre-treated biopsy tissue was associated with NAC response in ER positive BCs [32]. In terms of PR score, lower PR score was associated with higher pCR rate regardless of HER2 status but was also associated with worse survival outcome in patients with BC who did not achieve pCR after NAC [33]. Other studies revealed that high growth factor activity may be associated with decreased PR score, and one-quarter of HER2 + BCs were ER+/PR- compared with 10% of ER+/PR + in HER2- BCs [34, 35]. Our study also suggested different contributions of ER and PR according to HER2 status, as demonstrated in previous studies in a neoadjuvant setting.

Interestingly, menopausal status influenced NAC response in HR+/HER + BCs. We might suggest that serum estrogen level influenced ER and HER2 pathways to affect pCR status.

Pathologic complete response is the most important clinical characteristic to predict relapse free survival and overall survival [4]. Therefore, pCR is the standard primary end point of studies in the setting of NAC [36]. Most studies for NAC response have focused on identifying biomarkers to predict NAC response [37–39]. According to these translational research studies, presence of ctDNA was associated with worse

prognosis. However, we cannot use results of translational research to predict pCR in real world practice at the present time.

We validated the prediction model internally, not externally using another cohort of BC treated with NAC. To improve the performance of our prediction model, external validation is warranted.

In conclusion, we developed a ML model of pCR prediction using baseline clinical characteristics in patients with early and locally advanced BC treated with NAC followed by curative surgery. This model indicated that different clinical characteristics with different weights affect pCR according to BC subtype. We suggest that our prediction model would help physicians to treat patients with BC in real world practice. Further external validation using another BC cohort would be warranted.

## **Declarations**

### **Acknowledgments**

### **Contributions**

*Y.I. and J.K. conceived and planned the experiments. J.K., E.J. and H.J. carried out analyses and experiments. J.K., J.E.L., S.J.N., Y.H.P., J.S.A., and Y.I. contributed to collection of samples and clinical data. J.K., E.J., H.J., S.K., S.J. and Y.P. contributed to the interpretation of the results. J.K., E.J., and Y.I. took the lead in writing the manuscript. Y.I. supervised the project. All authors reviewed and confirmed the manuscript.*

### **Ethical declarations**

### **Ethics approval and consent to participate**

This study was reviewed and approved by the Institutional Review Board (IRB) of Samsung Medical Center, Seoul, Korea (IRB No. 2019-04-021). The requirement for individual informed consent was waived because of the retrospective clinical data review. This study was performed in accordance with the Declaration of Helsinki.

### **Consent to publish**

All authors have agreed to the content of the manuscript and provided consent regarding its publication.

### **Data availability**

Data can be made available upon request; some restrictions will apply.

### **Competing interests**

The authors have no competing interests to declare.

## Funding information

This work was supported by Institute for Information and Communications Technology Promotion grant funded by the Korean government (2018-0-00861, Intelligent SW Technology Development for Medical Data Analysis) and grants from the National Research Foundation of Korea (NRF-2020R1F1A1072616).

## References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 68:394-424. <http://doi.org/10.3322/caac.21492>
2. Siegel RL, Miller KD, Jemal A (2019) Cancer statistics, 2019. *CA Cancer J Clin* 69:7-34. <http://doi.org/10.3322/caac.21551>
3. Bonadonna G (1992) Evolving concepts in the systemic adjuvant treatment of breast cancer. *Cancer Res* 52:2127-2137.
4. Cortazar P, Zhang L, Untch M, Mehta K, Costantino JP, Wolmark N, Bonnefoi H, Cameron D, Gianni L, Valagussa P, Swain SM, Prowell T, Loibl S, Wickerham DL, Bogaerts J, Baselga J, Perou C, Blumenthal G, Blohmer J, Mamounas EP, Bergh J, Semiglazov V, Justice R, Eidtmann H, Paik S, Piccart M, Sridhara R, Fasching PA, Slaets L, Tang S, Gerber B, Geyer CE, Jr., Pazdur R, Ditsch N, Rastogi P, Eiermann W, von Minckwitz G (2014) Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet* 384:164-172. [http://doi.org/10.1016/S0140-6736\(13\)62422-8](http://doi.org/10.1016/S0140-6736(13)62422-8)
5. von Minckwitz G, Untch M, Blohmer JU, Costa SD, Eidtmann H, Fasching PA, Gerber B, Eiermann W, Hilfrich J, Huober J, Jackisch C, Kaufmann M, Konecny GE, Denkert C, Nekljudova V, Mehta K, Loibl S (2012) Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. *J Clin Oncol* 30:1796-1804. <http://doi.org/10.1200/JCO.2011.38.8595>
6. Symmans WF, Peintinger F, Hatzis C, Rajan R, Kuerer H, Valero V, Assad L, Poniacka A, Hennessy B, Green M, Buzdar AU, Singletary SE, Hortobagyi GN, Pusztai L (2007) Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *J Clin Oncol* 25:4414-4422. <http://doi.org/10.1200/JCO.2007.10.6823>
7. Schneeweiss A, Chia S, Hickish T, Harvey V, Eniu A, Hegg R, Tausch C, Seo JH, Tsai YF, Ratnayake J, McNally V, Ross G, Cortes J (2013) Pertuzumab plus trastuzumab in combination with standard neoadjuvant anthracycline-containing and anthracycline-free chemotherapy regimens in patients with HER2-positive early breast cancer: a randomized phase II cardiac safety study (TRYPHAENA). *Ann Oncol* 24:2278-2284. <http://doi.org/10.1093/annonc/mdt182>
8. Swain SM, Ewer MS, Viale G, Delaloge S, Ferrero JM, Verrill M, Colomer R, Vieira C, Werner TL, Douthwaite H, Bradley D, Waldron-Lynch M, Kiermaier A, Eng-Wong J, Dang C, Group BS (2018) Pertuzumab, trastuzumab, and standard anthracycline- and taxane-based chemotherapy for the

- neoadjuvant treatment of patients with HER2-positive localized breast cancer (BERENICE): a phase II, open-label, multicenter, multinational cardiac safety study. *Ann Oncol* 29:646-653.  
<http://doi.org/10.1093/annonc/mdx773>
9. Wu J, Li S, Jia W, Su F (2011) Response and prognosis of taxanes and anthracyclines neoadjuvant chemotherapy in patients with triple-negative breast cancer. *J Cancer Res Clin Oncol* 137:1505-1510.  
<http://doi.org/10.1007/s00432-011-1029-6>
  10. Hyo-Eun Kim HHK, Boo-Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, Eun-Kyung Kim (2020) Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health* 2:e138-e148.  
[http://doi.org/https://doi.org/10.1016/S2589-7500\(20\)30003-0](http://doi.org/https://doi.org/10.1016/S2589-7500(20)30003-0)
  11. Wenbin Yue ZW, Hongwei Chen, Annette Payne, Xiaohui Liu (2018) Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. *Designs* 2
  12. Akay MF (2009) Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications* 36:3240-3247.  
<http://doi.org/10.1016/j.eswa.2008.01.009>
  - 13.
  14. Edge SB BD, Compton CC, Fritz AG, Greene FL, Trotti A, (editors). (2010) *AJCC cancer staging manual*, 7th edition. Springer, France
  15. Onitilo AA, Engel JM, Greenlee RT, Mukesh BN (2009) Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival. *Clin Med Res* 7:4-13.  
<http://doi.org/10.3121/cm.2009.825>
  16. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Morgan Kaufmann Publishers Inc., 340 Pine Street, Sixth Floor San Francisco CA United States, p 7
  17. Snoek J, Hugo Larochelle, and Ryan P. Adams (2012) Practical bayesian optimization of machine learning algorithms. Curran Associates Inc, 57 Morehouse Lane Red Hook, NY, United States, p 9
  18. Terence Parr KT, Christopher Csiszar, Jeremy Howard (2018) Beware Default Random Forest Importances
  19. Bottini A, Berruti A, Tampellini M, Morrica B, Brunelli A, Gnocchi E, Brizzi MP, Aguggini S, Fara E, Alquati P, Dogliotti L (1997) Influence of neoadjuvant chemotherapy on serum tumor markers CA 15-3, MCA, CEA, TPS and TPA in breast cancer patients with operable disease. *Tumour Biol* 18:301-310.  
<http://doi.org/10.1159/000218043>
  20. Wu SG, He ZY, Zhou J, Sun JY, Li FY, Lin Q, Guo L, Lin HX (2014) Serum levels of CEA and CA15-3 in different molecular subtypes and prognostic value in Chinese breast cancer. *Breast* 23:88-93.  
<http://doi.org/10.1016/j.breast.2013.11.003>
  21. Lips EH, Mukhtar RA, Yau C, de Ronde JJ, Livasy C, Carey LA, Loo CE, Vrancken-Peeters MJ, Sonke GS, Berry DA, Van't Veer LJ, Esserman LJ, Wesseling J, Rodenhuis S, Shelley Hwang E, Investigators

- IST (2012) Lobular histology and response to neoadjuvant chemotherapy in invasive breast cancer. *Breast Cancer Res Treat* 136:35-43. <http://doi.org/10.1007/s10549-012-2233-z>
22. Mamounas EP, Bryant J, Lembersky B, Fehrenbacher L, Sedlacek SM, Fisher B, Wickerham DL, Yothers G, Soran A, Wolmark N (2005) Paclitaxel after doxorubicin plus cyclophosphamide as adjuvant chemotherapy for node-positive breast cancer: results from NSABP B-28. *J Clin Oncol* 23:3686-3696. <http://doi.org/10.1200/JCO.2005.10.517>
23. Rastogi P, Anderson SJ, Bear HD, Geyer CE, Kahlenberg MS, Robidoux A, Margolese RG, Hoehn JL, Vogel VG, Dakhil SR, Tamkus D, King KM, Pajon ER, Wright MJ, Robert J, Paik S, Mamounas EP, Wolmark N (2008) Preoperative chemotherapy: updates of National Surgical Adjuvant Breast and Bowel Project Protocols B-18 and B-27. *J Clin Oncol* 26:778-785. <http://doi.org/10.1200/JCO.2007.15.0235>
24. Petrelli F, Coiu A, Borgonovo K, Cabiddu M, Ghilardi M, Lonati V, Barni S (2014) The value of platinum agents as neoadjuvant chemotherapy in triple-negative breast cancers: a systematic review and meta-analysis. *Breast Cancer Res Treat* 144:223-232. <http://doi.org/10.1007/s10549-014-2876-z>
25. Hamy AS, Belin L, Bonsang-Kitzis H, Paquet C, Pierga JY, Lerebours F, Cottu P, Rouzier R, Savignoni A, Lae M, Reyat F (2016) Pathological complete response and prognosis after neoadjuvant chemotherapy for HER2-positive breast cancers before and after trastuzumab era: results from a real-life cohort. *Br J Cancer* 114:44-52. <http://doi.org/10.1038/bjc.2015.426>
26. Houssami N, Macaskill P, von Minckwitz G, Marinovich ML, Mamounas E (2012) Meta-analysis of the association of breast cancer subtype and pathologic complete response to neoadjuvant chemotherapy. *Eur J Cancer* 48:3342-3354. <http://doi.org/10.1016/j.ejca.2012.05.023>
27. Mougalian SS, Soulos PR, Killelea BK, Lannin DR, Abu-Khalaf MM, DiGiovanna MP, Sanft TB, Puzstai L, Gross CP, Chagpar AB (2015) Use of neoadjuvant chemotherapy for patients with stage I to III breast cancer in the United States. *Cancer* 121:2544-2552. <http://doi.org/10.1002/cncr.29348>
28. Mayer IA, Abramson VG, Lehmann BD, Pietenpol JA (2014) New strategies for triple-negative breast cancer—deciphering the heterogeneity. *Clin Cancer Res* 20:782-790. <http://doi.org/10.1158/1078-0432.CCR-13-0583>
29. Imamura M, Morimoto T, Nomura T, Michishita S, Nishimukai A, Higuchi T, Fujimoto Y, Miyagawa Y, Kira A, Murase K, Araki K, Takatsuka Y, Oh K, Masai Y, Akazawa K, Miyoshi Y (2018) Independent prognostic impact of preoperative serum carcinoembryonic antigen and cancer antigen 15-3 levels for early breast cancer subtypes. *World J Surg Oncol* 16:26. <http://doi.org/10.1186/s12957-018-1325-6>
30. Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, Taube S, Somerfield MR, Hayes DF, Bast RC, Jr., American Society of Clinical O (2007) American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J Clin Oncol* 25:5287-5312. <http://doi.org/10.1200/JCO.2007.14.2364>
31. Landmann A, Farrugia DJ, Zhu L, Diego EJ, Johnson RR, Soran A, Dabbs DJ, Clark BZ, Puhalla SL, Jankowitz RC, Brufsky AM, Ahrendt GM, McAuliffe PF, Bhargava R (2018) Low Estrogen Receptor

- (ER)-Positive Breast Cancer and Neoadjuvant Systemic Chemotherapy: Is Response Similar to Typical ER-Positive or ER-Negative Disease? *Am J Clin Pathol* 150:34-42.  
<http://doi.org/10.1093/ajcp/aqy028>
32. Sheri A, Smith IE, Hills M, Jones RL, Johnston SR, Dowsett M (2017) Relationship between IHC4 score and response to neo-adjuvant chemotherapy in estrogen receptor-positive breast cancer. *Breast Cancer Res Treat* 164:395-400. <http://doi.org/10.1007/s10549-017-4266-9>
33. van Mackelenbergh MT, Denkert C, Nekljudova V, Karn T, Schem C, Marme F, Stickeler E, Jackisch C, Hanusch C, Huober J, Fasching PA, Blohmer JU, Kummel S, Muller V, Schneeweiss A, Untch M, von Minckwitz G, Weber KE, Loibl S (2018) Outcome after neoadjuvant chemotherapy in estrogen receptor-positive and progesterone receptor-negative breast cancer patients: a pooled analysis of individual patient data from ten prospectively randomized controlled neoadjuvant trials. *Breast Cancer Res Treat* 167:59-71. <http://doi.org/10.1007/s10549-017-4480-5>
34. Konecny G, Pauletti G, Pegram M, Untch M, Dandekar S, Aguilar Z, Wilson C, Rong HM, Bauerfeind I, Felber M, Wang HJ, Beryt M, Seshadri R, Hepp H, Slamon DJ (2003) Quantitative association between HER-2/neu and steroid hormone receptors in hormone receptor-positive primary breast cancer. *J Natl Cancer Inst* 95:142-153. <http://doi.org/10.1093/jnci/95.2.142>
35. Dowsett M, Harper-Wynne C, Boeddinghaus I, Salter J, Hills M, Dixon M, Ebbs S, Gui G, Sacks N, Smith I (2001) HER-2 amplification impedes the antiproliferative effects of hormone therapy in estrogen receptor-positive primary breast cancer. *Cancer Res* 61:8452-8458.
36. (CDER) USDoHaHSFaDACfDEaR (2014) Guidance for Industry Pathological Complete Response in Neoadjuvant Treatment of High-Risk Early-Stage Breast Cancer: Use as an Endpoint to Support Accelerated Approval
37. Kim JY, Park D, Son DS, Nam SJ, Kim SW, Jung HH, Kim YJ, Park G, Park WY, Lee JE, Park YH (2017) Circulating tumor DNA shows variable clonal response of breast cancer during neoadjuvant chemotherapy. *Oncotarget* 8:86423-86434. <http://doi.org/10.18632/oncotarget.21198>
38. Moss J, Zick A, Grinshpun A, Carmon E, Maoz M, Ochana BL, Abraham O, Arieli O, Germansky L, Meir K, Glaser B, Shemer R, Uziely B, Dor Y (2020) Circulating breast-derived DNA allows universal detection and monitoring of localized breast cancer. *Ann Oncol* 31:395-403.  
<http://doi.org/10.1016/j.annonc.2019.11.014>
39. Bownes RJ, Turnbull AK, Martinez-Perez C, Cameron DA, Sims AH, Oikonomidou O (2019) On-treatment biomarkers can improve prediction of response to neoadjuvant chemotherapy in breast cancer. *Breast Cancer Res* 21:73. <http://doi.org/10.1186/s13058-019-1159-3>

## Tables

**Table1. Baseline characteristics according to pathologic complete response (N=2065)**

| Characteristics                 | Non-pCR<br>(N=1433) | Pathologic CR<br>(N=632) | P-value            |
|---------------------------------|---------------------|--------------------------|--------------------|
| Age, Median ( SD)               | 46.1 (±9.8)         | 47.7 (±9.7)              | .009 <sup>a</sup>  |
| BMI <sup>1</sup> , Median ( SD) | 23.0 (±3.6)         | 22.8 (±3.3)              | .305 <sup>a</sup>  |
| Menopausal status               |                     |                          | .011 <sup>b</sup>  |
| Premenopause                    | 469 (65.8)          | 244 (34.2)               |                    |
| Postmenopause                   | 964 (71.3)          | 388 (28.7)               |                    |
| Smoking                         |                     |                          | .921 <sup>b</sup>  |
| Never smoker                    | 1395 (69.1)         | 615 (30.9)               |                    |
| Ex/current smoker               | 38 (69.4)           | 17 (30.6)                |                    |
| CEA, Mean ( SD)                 | 2.5 (±15.1)         | 2.8 (±28.9)              | .787 <sup>a</sup>  |
| CA-15-3, Mean ( SD)             | 17.0 (±21.9)        | 12.3 (±9.7)              | <.001 <sup>a</sup> |
| Histology                       |                     |                          | .239 <sup>b</sup>  |
| Invasive ductal carcinoma       | 1285 (69.0)         | 578 (31.0)               |                    |
| Non-IDC <sup>2</sup>            | 148 (73.3)          | 54 (34.2)                |                    |
| Estrogen receptor (Allred)      | (N=2064)            |                          | <.001 <sup>b</sup> |
| 0                               | 616 (60.9)          | 395 (39.1)               |                    |
| 1                               | 0 (0)               | 0 (0)                    |                    |
| 2                               | 48 (57.8)           | 35 (42.2)                |                    |
| 3                               | 43 (54.4)           | 36 (45.6)                |                    |
| 4                               | 46 (64.8)           | 25 (35.2)                |                    |
| 5                               | 28 (58.3)           | 20 (41.7)                |                    |
| 6                               | 41 (63.1)           | 24 (36.9)                |                    |
| 7                               | 152 (84.4)          | 28 (15.6)                |                    |
| 8                               | 459 (87.1)          | 68 (12.9)                |                    |
| Progesterone receptor           |                     |                          | <.001 <sup>b</sup> |
| 0                               | 736 (59.9)          | 493 (40.1)               |                    |
| 1                               | 1 (100)             | 0 (0)                    |                    |

|                         |             |            |                    |
|-------------------------|-------------|------------|--------------------|
| 2                       | 22 (62.9)   | 13 (37.1)  |                    |
| 3                       | 53 (70.7)   | 22 (29.3)  |                    |
| 4                       | 76 (79.2)   | 20 (20.8)  |                    |
| 5                       | 65 (83.3)   | 13 (16.7)  |                    |
| 6                       | 90 (92.8)   | 7 (7.2)    |                    |
| 7                       | 121 (90.3)  | 13 (9.7)   |                    |
| 8                       | 269 (84.0)  | 51 (16.0)  |                    |
| <b>HER2<sup>3</sup></b> |             |            | <.001 <sup>b</sup> |
| <b>Positive</b>         | 393 (52.1)  | 362 (47.9) |                    |
| <b>Negative</b>         | 1040 (79.4) | 270 (20.6) |                    |
| <b>Ki-67</b>            | (N=2022)    |            | <.001 <sup>b</sup> |
| <b>1+</b>               | 368 (82.0)  | 81 (18.0)  |                    |
| <b>2+</b>               | 444 (66.5)  | 224 (33.5) |                    |
| <b>3+</b>               | 333 (68.2)  | 155 (31.8) |                    |
| <b>4+</b>               | 257 (61.6)  | 160 (38.4) |                    |
| <b>Clinical stage</b>   |             |            | <.001 <sup>b</sup> |
| <b>IA</b>               | 10 (52.6)   | 9 (47.6)   |                    |
| <b>IIA</b>              | 151 (57.6)  | 111 (42.4) |                    |
| <b>IIB</b>              | 247 (59.4)  | 169 (40.6) |                    |
| <b>IIIA</b>             | 592 (72.5)  | 225 (27.5) |                    |
| <b>IIIB</b>             | 68 (81.9)   | 15 (18.1)  |                    |
| <b>IIIC</b>             | 365 (78.0)  | 103 (22.0) |                    |
| <b>Clinical T stage</b> |             |            | <.001 <sup>b</sup> |
| <b>0</b>                | 4 (50.0)    | 4 (50.0)   |                    |
| <b>1</b>                | 75 (54.0)   | 64 (46.0)  |                    |
| <b>2</b>                | 775 (65.7)  | 405 (34.3) |                    |
| <b>3</b>                | 465 (77.4)  | 136 (22.6) |                    |
| <b>4</b>                | 114 (83.2)  | 23 (16.8)  |                    |

|  |            |            |                    |
|--|------------|------------|--------------------|
| <b>Clinical N stage</b>                  |            |            | <.001 <sup>b</sup> |
| <b>0</b>                                 | 173 (60.5) | 113 (39.5) |                    |
| <b>1</b>                                 | 352 (62.1) | 215 (37.9) |                    |
| <b>2</b>                                 | 543 (73.0) | 201 (27.0) |                    |
| <b>3</b>                                 | 365 (78.0) | 103 (22.0) |                    |
| <b>Chemotherapy regimen</b>              |            |            | <.001 <sup>b</sup> |
| <b>AC<sup>4</sup>-T<sup>5</sup></b>      | 963 (79.3) | 251 (20.7) |                    |
| <b>CTx<sup>6</sup> with trastuzumab</b>  | 227 (60.5) | 148 (39.5) |                    |
| <b>CTx with trastuzumab + pertuzumab</b> | 90 (33.2)  | 181 (66.8) |                    |
| <b>AC-T plus platinum</b>                | 65 (92.9)  | 5 (7.1)    |                    |
| <b>AC alone</b>                          | 54 (84.4)  | 10 (15.6)  |                    |
| <b>Others</b>                            | 34 (47.9)  | 37 (52.1)  |                    |

Note. <sup>a)</sup> t-test for continuous variables, <sup>b)</sup> Chi-square test for categorical variables

<sup>1</sup>Body mass index; <sup>2</sup>invasive ductal carcinoma; <sup>3</sup>human epidermal growth factor receptor 2; <sup>4</sup>adriamycin/cyclophosphamide; <sup>5</sup> taxane; <sup>6</sup> chemotherapy

**Table 2. Hyper-parameters of Light GBM**

| Parameters           | Optimal value |
|----------------------|---------------|
| Bagging fraction     | 0.40          |
| Feature fraction     | 0.37          |
| Lambda L1            | 0.05          |
| Lambda L2            | 0.91          |
| Max depth            | 5.04          |
| Min child weight     | 4.59          |
| Min split gain       | 0.09          |
| Number of leaves     | 197.02        |
| Learning rate        | 0.005         |
| Number of estimators | 8000          |
| Class weight         | Balanced      |
| Importance type      | Gain          |

**Table 3. Performance of pCR prediction model in different pathology-based subtypes**

| Subtypes                  | All   | HR <sup>3+</sup> /HER2 <sup>4-</sup> | HR+/HER2+ | HR-HER2+ | TNBC <sup>5</sup> |
|---------------------------|-------|--------------------------------------|-----------|----------|-------------------|
| pCR <sup>1</sup> rate (%) | 30.6  | 10.6                                 | 39.9      | 45.4     | 30.3              |
| AUC <sup>2</sup>          | 0.810 | 0.840                                | 0.716     | 0.753    | 0.653             |
| Sensitivity               | 0.775 | 0.721                                | 0.679     | 0.617    | 0.668             |
| Specificity               | 0.685 | 0.791                                | 0.657     | 0.782    | 0.538             |

<sup>1</sup> Pathologic complete response; <sup>2</sup> area under the curve; <sup>3</sup> hormone receptor; <sup>4</sup> human epidermal growth factor receptor; <sup>5</sup> triple-negative breast cancer

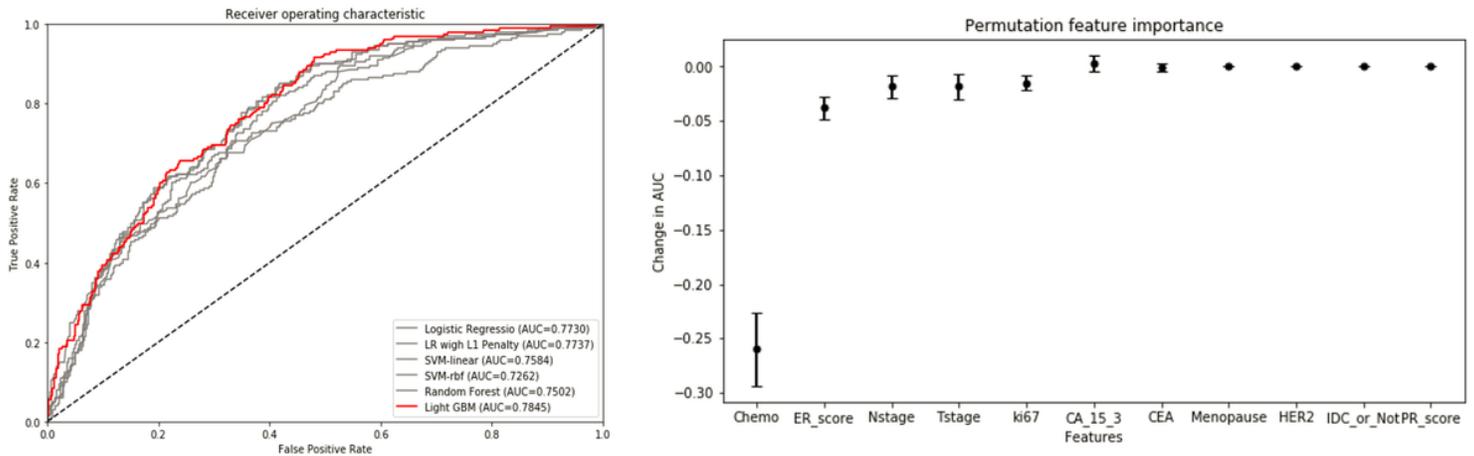
## Supplementary Data

The Supplementary Figure and Tables are not available with this version

### Supplementary figure 1. Correlation between all features

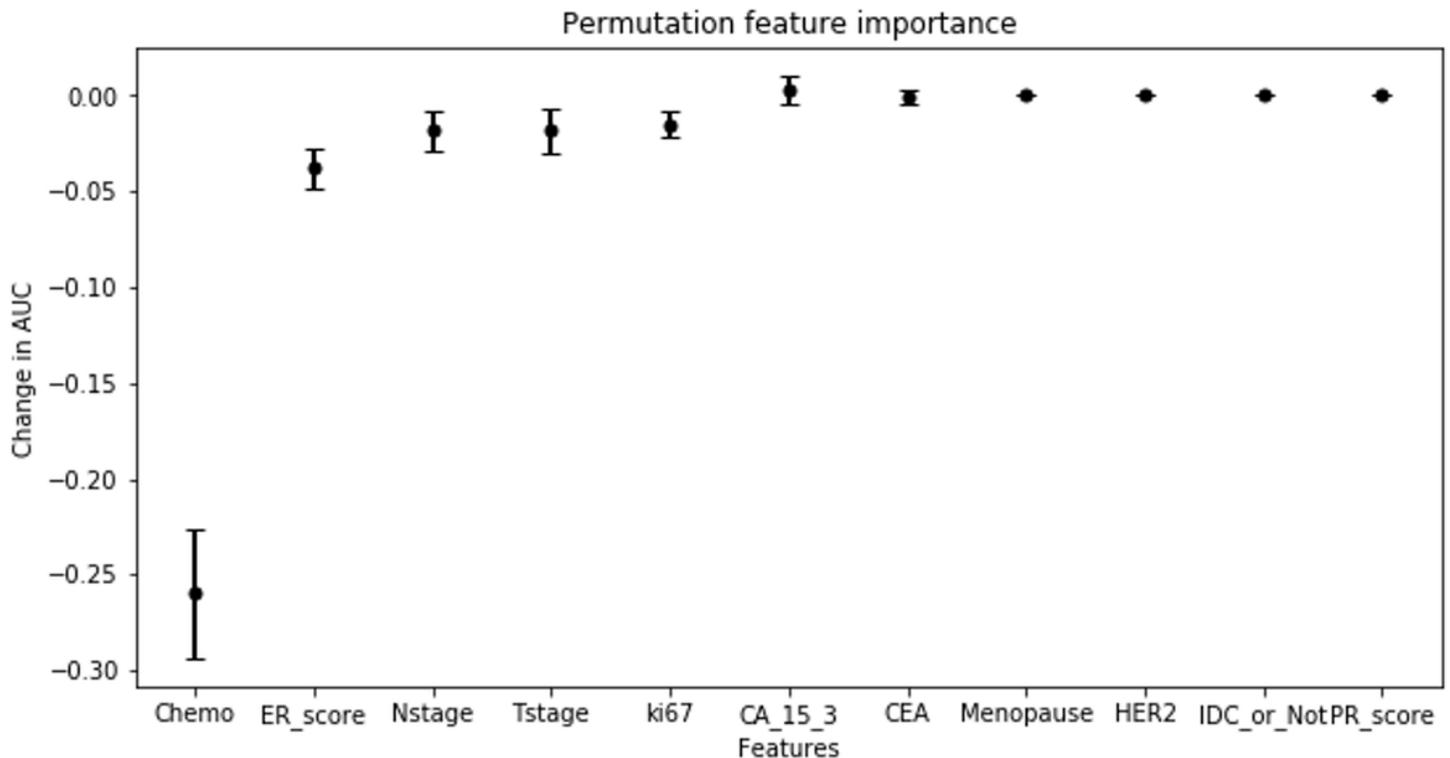
(Note. pCR, pathologic complete response; Menopause, menopausal status; HER2, human epidermal growth factor receptor 2; Chemo, chemotherapy regimen; IDC\_or\_Not, histology; Co\_Cancer, Comorbidity (Cancer) ; Co\_Hepatitis, Comorbidity (Hepatitis); Co\_Metabolic, Comorbidity (Metabolic) ;Co\_CVD, Comorbidity (cerebrovascular disease); Co\_others, Comorbidity (others) ; ER\_score, estrogen receptor; PR\_score, progesterone receptor; cstageN, clinical stage)

## Figures



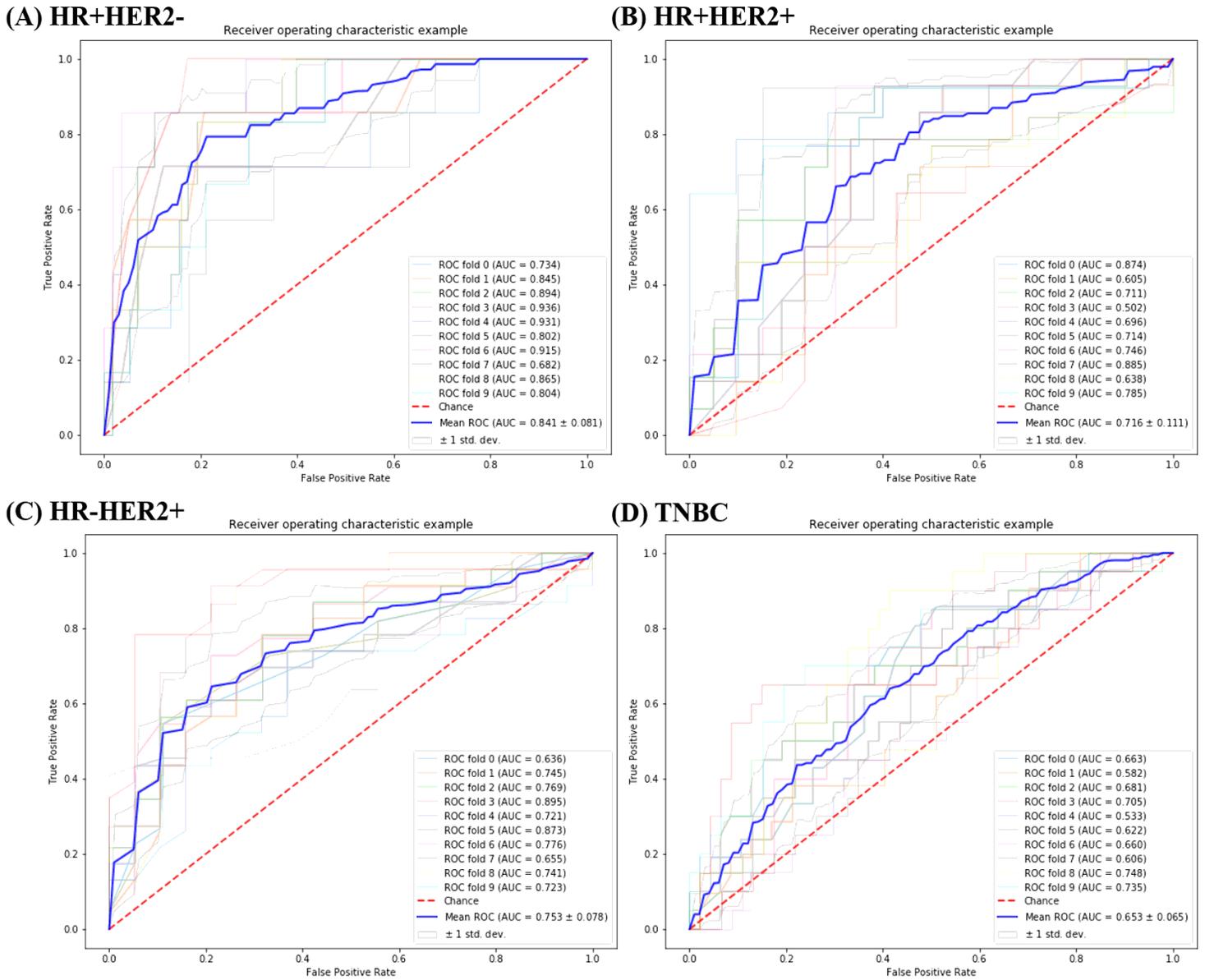
**Figure 1**

(A) Comparison of six machine learning models, (B) Area under the curve of LightGBM model with 10-fold expansion



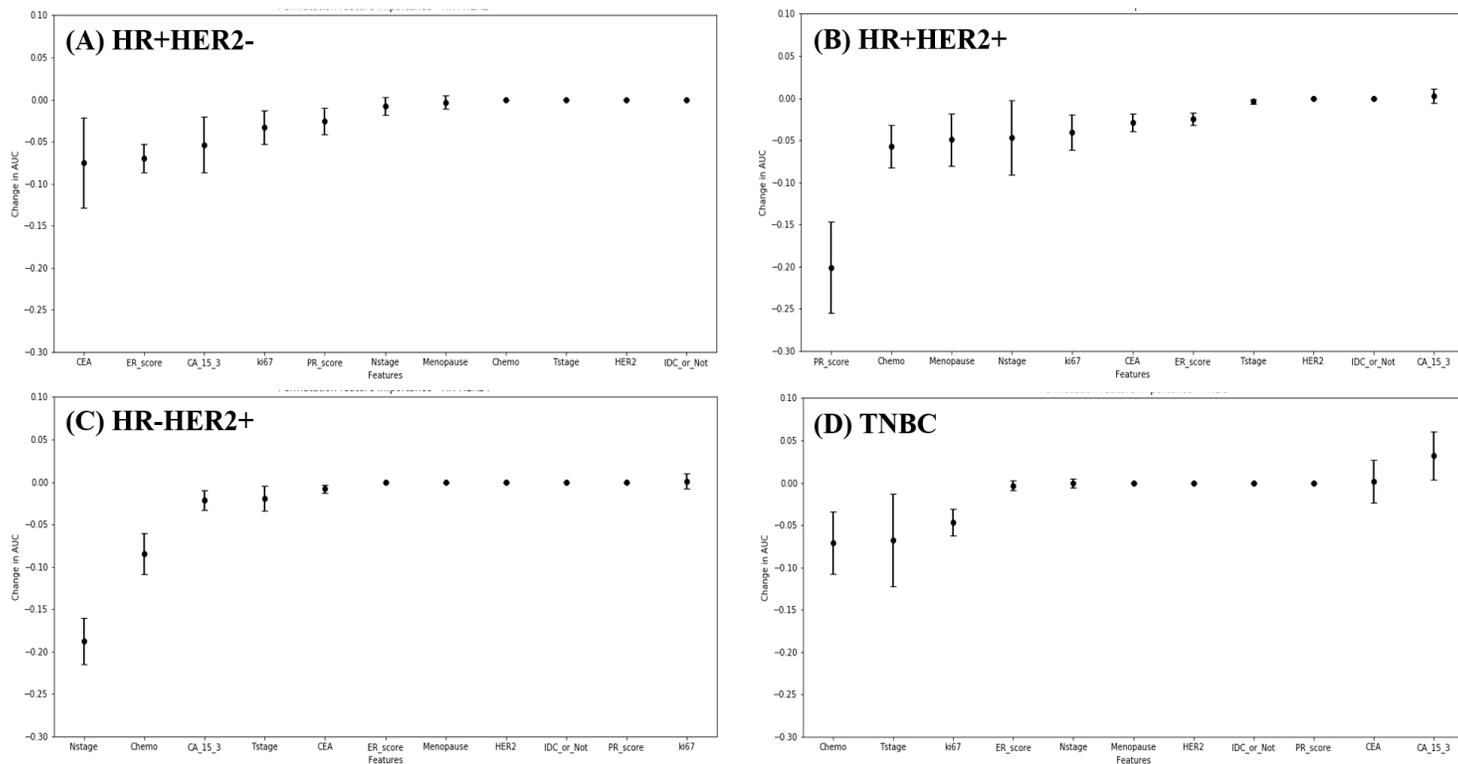
**Figure 2**

Permutation feature importance (Note. Chemo, Chemotherapy regimen; ER\_score, estrogen receptor; Menopause, menopausal status; HER2, human epidermal growth factor receptor 2; IDC\_or\_Not, histology; PR\_score, progesterone receptor)



**Figure 3**

Machine learning model for pathologic complete response prediction according to breast cancer subtype: (A) Hormone receptor (HR)+/Human epidermal growth factor receptor 2 (HER2)-, (B) HR+ /HER2+, (C) HR-HER2+, (D) Triple negative breast cancer



**Figure 4**

Permutation feature importance according to breast cancer subtype: (A) Hormone receptor (HR)+/Human epidermal growth factor receptor 2 (HER2)-, (B) HR+/HER2+, (C) HR-HER2+, (D) Triple negative breast cancer (Note. ER\_score, estrogen receptor; PR score, progesterone receptor; Menopause, menopausal status; Chemo, chemotherapy regimen; HER2, human epidermal growth factor receptor 2; IDC\_or\_Not, histology)