Scientific Research Publishing

Advances in Breast Cancer Research, 2021, 10, 141-155
https://www.scirp.org/journal/abcr
ISSN Online: 2168-1597
ISSN Print: 2168-1589

A Predictive Model for Pathologic Complete Response in Breast Cancer Patients Treated with Neoadjuvant Chemotherapy Using Machine Learning

Isaac Kim[1], Kwanbum Lee[1], Seung Ah Lee[1], Yeon Hee Park[2], Seung Ki Kim[1*]

[1]Department of Surgery, Bundang CHA Medical Center, CHA University School of Medicine, Yatap-ro, Gyeonggi-do, Republic of Korea
[2]Division of Hematology-Oncology, Department of Internal Medicine, Samsung Medical Center, Sungkyunkwan University, School of Medicine, Irwon-ro, Gangnam-gu, Seoul, Republic of Korea
Email: *mdsky@cha.ac.kr

How to cite this paper: Kim, I., Lee, K., Lee, S.A., Park, Y.H. and Kim, S.K. (2021) A Predictive Model for Pathologic Complete Response in Breast Cancer Patients Treated with Neoadjuvant Chemotherapy Using Machine Learning. *Advances in Breast Cancer Research*, **10**, 141-155.
https://doi.org/10.4236/abcr.2021.104012

Received: July 20, 2021
Accepted: August 23, 2021
Published: August 26, 2021

Abstract

Background: In patients with breast cancer after Neoadjuvant Chemotherapy (NAC), pathological Complete Response (pCR) was associated with better long-term outcomes. We here attempted to predict pCR using machine learning. **Patients and Methods:** From 2008 to 2017, 1308 breast cancer patients underwent NAC before surgery, of whom 377 patients underwent Cancer SCAN™ for gene data. Of 377, 238 were analyzed here, with 139 excluded due to incomplete medical data. **Results:** The pCR (−) vs. (+) group had 200 vs. 38 patients. In our predictive model with gene data, the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve was 0.909 and accuracy was 0.875. In another model without gene data, the AUC of ROC curve was 0.743 and accuracy was 0.800. We also conducted internal validation with 72 patients undergoing NAC and Cancer SCAN™ during July 2017 and April 2018. When we applied a 0.4 threshold value, accuracy was 0.806 and 0.778 in the predictive model with vs. without gene profiles, respectively. **Conclusion:** The present predictive model may be a useful and easy-to-access tool for pCR-prediction in breast cancer patients treated with NAC.

Keywords

Breast Neoplasm, Neoadjuvant Therapy, Chemotherapy, Response, Prediction

1. Introduction

Neoadjuvant Chemotherapy (NAC) has long been used for decreasing the tumor

size to either increase operability [1] [2]. In patients with NAC, pathological Complete Response (pCR) has been proposed as a surrogate endpoint for the prediction of long-term clinical benefits, such as Disease-Free Survival (DFS) and Overall Survival (OS) [3] [4]. Especially, there was the strongest association between pCR and long-term outcome in patients with aggressive breast cancer subtypes (triple negative, HER2-positive and hormone-receptor-negative) [5].

In previous reports, various ways were used for the prediction of pCR in breast cancer patients treated with NAC. Magnetic resonance imaging had a predictive value [6] [7] and high Tissue Infiltrating Lymphocyte (TIL) status was an independent factor for prediction [8]. In addition, pathologic factors such as Ki 67 proliferation index [9] or transcripts such as long non-coding RNAs were associated with pCR [10]. Meanwhile, the machine learning method has recently emerged as a new way of a prediction tool for effective and accurate decisions [11].

In this study, we present an easy-to-use prediction tool for pCR using machine learning. We used data from clinical characteristics and gene expression profiles. Gene profiles came from Cancer SCAN™, a targeted sequencing platform designed at Samsung Medical Center [12].

2. Methods

2.1. Study Population

We performed a retrospective chart review of 1308 breast cancer patients who underwent NAC and surgery between August 2008 and June 2017 at Samsung Medical Center in Seoul, Korea. Among them, 377 patients who underwent Cancer SCAN™ were included. Cancer SCAN™ test was conducted only on patients who agreed to provide genetic information. As part of this study, DNA sequencing results and electronic medical records including pathology reports were reviewed. 139 cases were excluded from analysis due to incomplete medical data and 238 cases were included for analysis. We used additional retrospective data from 72 patients who underwent NAC, surgery and Cancer SCAN™ between July 2017 and April 2018 for internal validation. This study adhered to the tenets of the Declaration of Helsinki and was approved by the institutional review board (IRB) of Samsung Medical Center (IRB No. 2018-05-035).

The available data for the cohorts included age at diagnosis, subtype (e.g., Hormone Receptor [HR] positive/Human Epidermal growth factor 2 [HER2] receptor negative, HR positive/HER2 positive, HR negative/HER2 positive, HR negative/HER2 negative), histopathology (e.g., Invasive Ductal Carcinoma [IDC], Invasive Lobular Carcinoma [ILC], mixed), menopausal status, family history for breast cancer, regimen for NAC (e.g., AC [adriamycin, cyclophosphamide], AC + D/T [docetaxel/taxol], AC + D/T + Herceptin [H], AC + Paclitaxel + Carboplatin, TCHP [docetaxel, carboplatin, trastuzumab, pertuzumab], others), Multiplicity, pathological T-stage, axillary nodal evaluation (clinical N0, axillary fine needle aspiration [FNA] result), results of supraclavicular and internal mammary lymph node (IMLN) FNA, Ki67 status, tumor marker level (carcinoembryonic antigen [CEA],

carcinoma antigen15-3 [CA15-3]) and gene profile. We defined pCR as breast and also axillary pCR simultaneously. Breast pCR was defined as no invasive disease (ypT0 or ypTis) on final pathologic results. Axillary pCR was defined as no metastasis (ypN0) or isolated tumor cell on final pathologic results.

2.2. DNA Extraction and Sequencing

Genomic DNA (250 ng) from each tissue was sheared in a Covaris S220 Ultrasonicator (Covaris, Woburn, MA) and used with CancerSCAN™ probes and a Sure Select XT reagent kit HSQ (AgilentTechnologies) for construction of a library according to the manufacturer's protocol [12].

This panel is designed to enrich exons of 81 genes, covering 366.2 kb of the human genome. After enriched exome libraries were multiplexed, the libraries were sequenced on a HiSeq 2500 sequencing platform (Illumina). Briefly, a paired-end DNA sequencing library was prepared through gDNA shearing, end-repair, A-tailing, paired-end adaptor ligation, and amplification. After hybridization of the library with bait sequences for 27 hours, the captured library was purified and amplified with an index barcode tag, and library quality and quantity were assessed [12]. We defined mutation as single nucleotide variants or copy number variation or translocation.

2.3. Statistical Analysis

Variables were compared between pCR (−) and pCR (+) groups using chi-squared test or Fisher's exact test, while mean age was compared between the two groups via Mann-Whitney *U* tests with SAS version 9.4 (SAS Institute, Cary, NC, USA). Receiver Operating Characteristic (ROC) curves and Areas Under the ROC Curve (AUC) were calculated. All tests were two-sided and a p-value of <0.05 was considered statistically significant.

2.4. Machine Learning

Azure Machine Learning (Azure ML; Microsoft, Redmond, WA, USA) is a cloud service that enables the execution of machine learning processes. The Azure Machine Learning Studio (Microsoft, Redmond, WA, USA) is also available as a workspace to help users build and test predictive models [13]. We built a supervised machine learning classification model using the Azure ML platform. This was accomplished using the steps of: 1) edit the data; 2) split the data; 3) train the model; 4) score the model; and 5) evaluate the model (**Figure 1**). We split the modeling data (238 cases) into training and testing sets using a randomized 60 - 40 split. We then trained our training set using a Two-class Bayes point machine method [14] for the prediction of pCR.

3. Results

3.1. Patient Characteristics

The clinicopathologic characteristics of included patients are summarized in **Table**

1. The pCR (−) group had 200 patients and the pCR (+) group had 38 patients. The median age was older in pCR (+) group (p-value = 0.038) and pCR (−) group had more premenopausal patients than pCR (+) group (p-value = 0.045). IDC, AC/AC + Taxane regimen and triple negative breast cancer (HR-/HER2-) subtype were majority in both groups. There was no difference in both groups according to family history, subtype, multiplicity, T stage, axillary nodal status, Ki-67 and tumor marker status. In gene profile results, only BRCA2 mutation was associated with pCR (+) status statistically (p-value = 0.014). Patients with BRCA2 mutation were more in pCR (−) group (36.5%) than pCR (+) group (15.8%). We developed a predictive model with 238 cases using the Azure ML platform (Figure 1) using various classification algorithms, such as Two-class Decision Forest, Two-class Decision Jungle, Two-class Decision Forest, Two-class Support Vector Machine, and Two-class Neural Network. Among them, Two-class Bayes Point Machine was the most suitable method for prediction of pCR. We assessed Area Under the Curve (AUC). The AUC of the Receiver Operating Characteristic (ROC) curve was 0.909 and accuracy was 0.875 (Figure 2(a)). In addition, we developed a predictive model without gene profiles. We used only clinical data but patients pool (n = 238) and process were same with previous model. Through additional model, the AUC of ROC curve was 0.743 and accuracy was 0.800 (Figure 2(b)).

3.2. Predictive Model

We developed a predictive model with 238 cases using the Azure ML platform (Figure 1) using various classification algorithms, such as Two-class Decision Forest, Two-class Decision Jungle, Two-class Decision Forest, Two-class Support Vector Machine, and Two-class Neural Network. Among them, Two-class Bayes Point Machine was the most suitable method for prediction of pCR. We assessed Area Under the Curve (AUC). The AUC of the receiver operating characteristic (ROC) curve was 0.909 and accuracy was 0.875 (Figure 2(a)). In addition, we developed a predictive model without gene profiles. We used only clinical data but patients pool (n = 238) and process were same with previous model. Through additional model, the AUC of ROC curve was 0.743 and accuracy was 0.800 (Figure 2(b)).

3.3. Validation

We also conducted internal validation using 72 patients who underwent NAC and Cancer SCAN™ during July 2017 and April 2018. When we applied a 0.4 threshold value, accuracy was 0.806 in predictive model with gene profiles and 0.778 in model without gene profiles respectively (Table 2). As threshold value was decreased, sensitivity was increased but specificity was decreased.

3.4. Clinical Application

The Azure ML platform provides a function for the set-up of web services: (http://docs.microsoft.com/en-us/azure/machine-learning/studio/consume-web-services).

Table 1. The baseline characteristics of enrolled patients.

		Total	pCR (−) (n = 200)	pCR (+) (n = 38)	p-value
Age	≤35	62	53 (26.5%)	9 (23.7%)	0.253
	35< age ≤45	92	79 (39.5%)	13 (34.2%)	
	45< age ≤55	57	48 (24.0%)	9 (23.7%)	
	55< age	27	20 (10.0%)	7 (18.4%)	
	Mean age (range)		41.9 (25 - 66)	45.3 (31-68)	0.038
Menopause	Post	46	34 (17.0%)	12 (31.6%)	0.045
	Pre	192	166 (83.0%)	26 (68.4%)	
Family History	None	214	178 (89.0%)	36 (94.7%)	0.386
	Breast cancer FHx	24	22 (11.0%)	2 (5.3%)	
Regimen	AC/AC + Taxane	144	128 (64%)	16 (42.1%)	0.076
	AC + Taxane + Herceptin	57	43 (21.5%)	14 (36.8%)	
	AC + Paclitaxel + Carboplatin	13	12 (6.0%)	1 (2.6%)	
	TCHP	8	3 (1.5%)	5 (13.2%)	
	Others	16	14 (7.0%)	2 (5.3%)	
Pathology	IDC	229	191 (95.5%)	38 (100%)	0.334
	ILC	3	3 (1.5%)	0	
	Mixed	6	6 (3.0%)	0	
Subtype	HR+/HER2−*	57	54 (27.0%)	3 (7.9%)	0.570
	HR+/HER2+#	31	20 (10.0%)	11 (28.9%)	
	HR−/HER2+	39	31 (15.5%)	8 (21.1%)	
	HR−/HER2−	111	95 (47.5%)	16 (42.1%)	
Multiplicity	None	121	98 (49.0%)	23 (60.5%)	0.218
	Multiplicity	117	102 (51.0%)	15 (39.5%)	
T stage	1	41	31 (15.5%)	10 (26.3%)	0.107
	2	108	92 (46.0%)	16 (42.1%)	
	3	87	76 (38.0%)	11 (28.9%)	
	4	2	1 (0.5%)	1 (2.6%)	
Axillary Nodal evaluation	Clinical N0	35	30 (15.0%)	5 (13.2%)	0.431
	Axillary FNA negative	48	42 (21.0%)	6 (15.8%)	
	Axillary FNA positive	155	128 (64.0%)	27 (71.1%)	
SCN FNA	Undone	225	169 (94.5%)	36 (94.7%)	1.000
	Positive	13	11 (5.5%)	2 (5.3%)	
	Negative	0	0 (0.0%)	0 (0.0%)	
IMLN FNA	Undone	226	189 (94.5%)	37 (97.4%)	1.000
	Positive	11	11 (5.5%)	0 (0.0%)	
	Negative	1	0 (0.0%)	1 (2.6%)	

Continued

Ki-67	1	37	34 (17.0%)	3 (7.9%)	1.000
	2	72	54 (27.0%)	18 (47.4%)	
	3	48	44 (22.0%)	4 (10.5%)	
	4	81	68 (34.0%)	13 (34.2%)	
Tumor marker	CEA elevation	10	8 (4.0%)	2 (5.3%)	1.000
	None	228	192 (96.0%)	36 (94.7%)	
	CA15-3 elevation	19	18 (9.0%)	1 (2.6%)	0.219
	None	219	182 (91.0%)	37 (97.4%)	
Gene profile	ESR mutation	10	10 (5.0%)	0 (0.0%)	0.226
	Wild type	228	190 (95.0%)	38 (100.0%)	
	EGFR mutation	17	16 (8.0%)	1 (2.6%)	0.323
	Wild type	221	184 (92.0%)	37 (97.4%)	
	ERRB2 mutation	72	57 (28.5%)	15 (39.5%)	0.183
	Wild type	166	143 (71.5%)	23 (60.5%)	
	PIK3CA mutation	50	43 (21.5%)	7 (18.4%)	0.829
	Wild type	188	157 (78.5%)	31 (81.6%)	
	BRCA1 mutation	70	58 (29.0%)	12 (31.6%)	0.846
	Wild type	168	142 (71.0%)	26 (68.4%)	
	BRCA2 mutation	79	73 (36.5%)	6 (15.8%)	0.014
	Wild type	159	127 (63.5%)	32 (84.2%)	
	TP53 mutation	189	161 (80.5%)	28 (73.7%)	0.382
	Wild type	49	39 (19.5%)	10 (26.3%)	
	PTEN mutation	25	24 (12.0%)	1 (2.6%)	0.143
	Wild type	213	176 (88.0%)	37 (97.4%)	

Abbreviation: FHx, family history; AC, adriamycin + cyclophosphamide; TCHP, docetaxel + carboplatin + trastuzumab + pertuzumab; IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; HR, hormone receptor; HER2, human epidermal growth factor receptor 2; FNA, fine needle aspiration; SCN, supraclavicular lymph node; IMLN, internal mammary lymph node; CEA, carcino embryonic antigen; CA15-3, carcinoma antigen 15-3. *HER2− means HER2 1+ or 2+ with SISH negative, #HER2+ means HER2 3+ or 2+ with SISH positive.

Table 2. The predictive results of validation with 72 patients.

Threshold value	Predictive model	Sensitivity	Specificity	Precision	Accuracy
0.2	With gene profiles	0.82	0.56	0.45	0.639
	Without gene profiles	0.64	0.68	0.47	0.667
0.3	With gene profiles	0.64	0.88	0.70	0.806
	Without gene profiles	0.55	0.84	0.60	0.750
0.4	With gene profiles	0.55	0.92	0.75	0.806
	Without gene profiles	0.45	0.92	0.71	0.778

Sensitivity = true positive/true positive + false negative; specificity = true negative/ false positive + true negative; Precision, positive predictive value = true positive/ true positive + false positive; accuracy = true positive+ true negative/total.

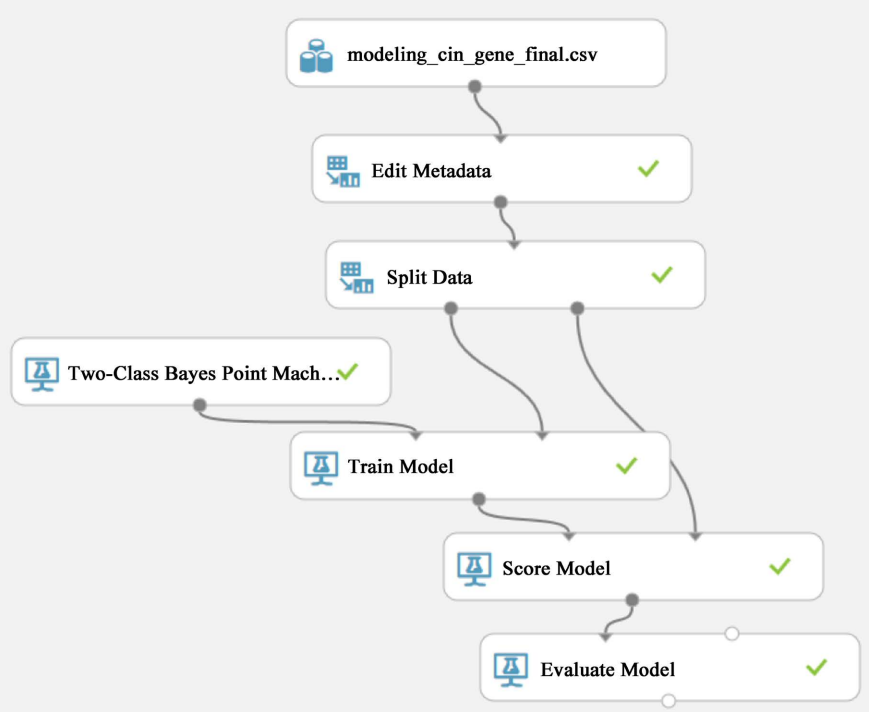

Figure 1. The workflow of modeling using Azure ML (Microsoft, Redmond, WA, USA). It consists of establishing a dataset, editing the metadata, employing an algorithm (e.g., Two-Class Bayes Point Machine), splitting the data, training the model, scoring the model, and evaluating the model.

After using the Azure ML predictive model as a web service, we used a Representational State Transfer application programming interface to send data and obtain predictions in real-time. For example, when we input data according to each variable excluding the final value "pCR," an external application communicated with a machine learning workflow scoring model in real-time, enabling the predicted value to be calculated in only a few seconds (**Figure 3**).

4. Discussion

The prediction of pCR in breast cancer patients treated with NAC is important in terms of management. The scope of surgery could vary depending on whether pCR or not and it is possible to consider novel NAC in the case of non-responder.

In previous study, various methods were used for prediction of pCR in patients treated with NAC. One of them was prediction using breast MRI. Weber *et al.* studied predictive value of MRI before and after NAC in128 patients [7]. MRI had a positive predictive value of 63.4% and negative predictive value of 84.1% for in-breast pCR. Moreover, Positive predictive value of axillary pCR was 65.6% and negative predictive value was 66.7% [7]. Lee *et al.* conducted retrospective study in 74 patients treated with NAC and underwent breast MRI before NAC [6]. They showed that perfusion parameters of tumor, background parenchyma of contralateral breast and their combination in pretreatment breast MRI allow early prediction for pCR of breast cancer. The highest predictive power

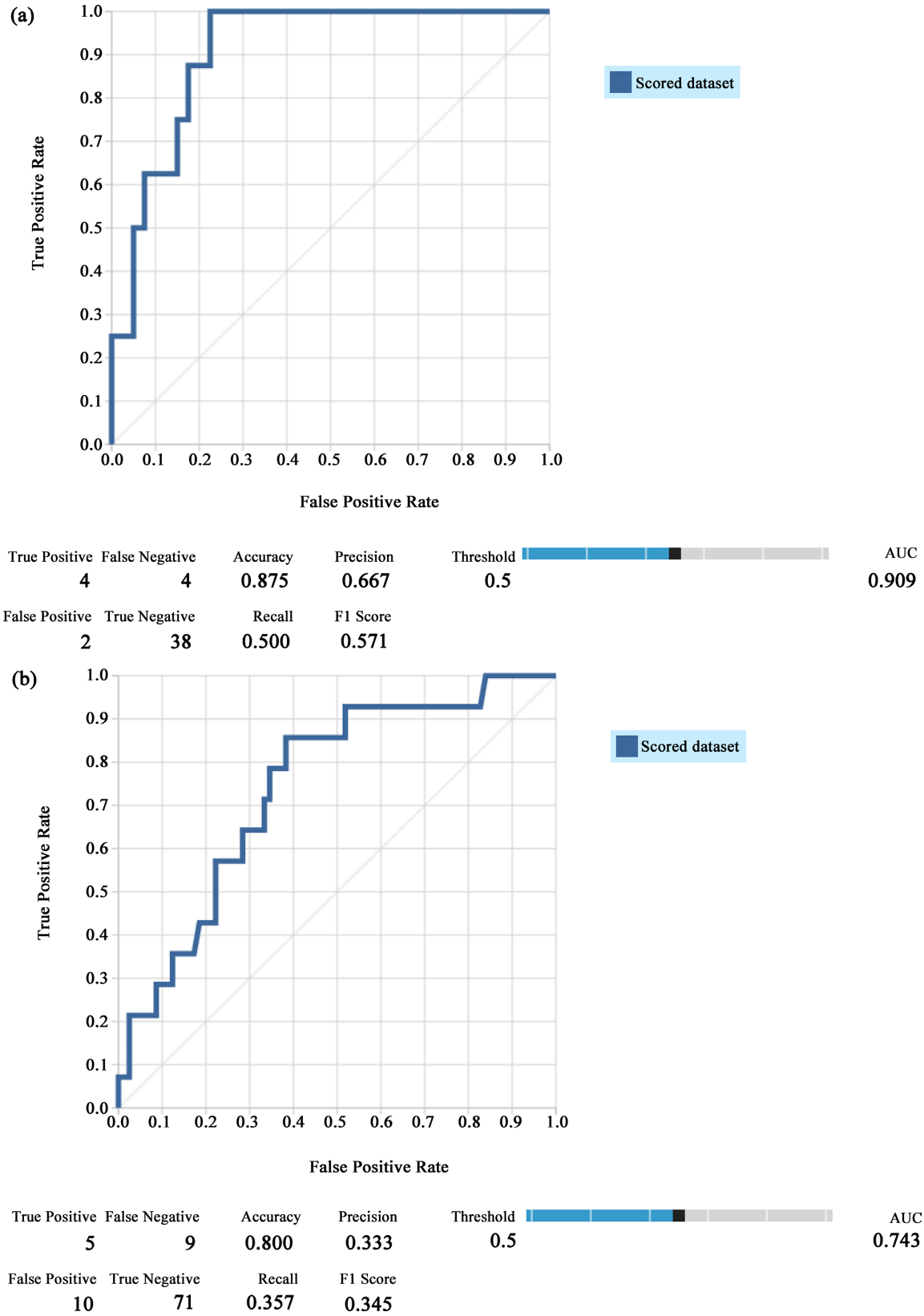

Figure 2. (a) The Receiver Operating Characteristic (ROC) curve of our predicted model with gene data. The Area Under the Curve (AUC) of ROC curve was 0.909; (b) The receiver operating characteristic (ROC) curve of our predicted model without gene data. The area under the curve (AUC) of ROC curve was 0.743.

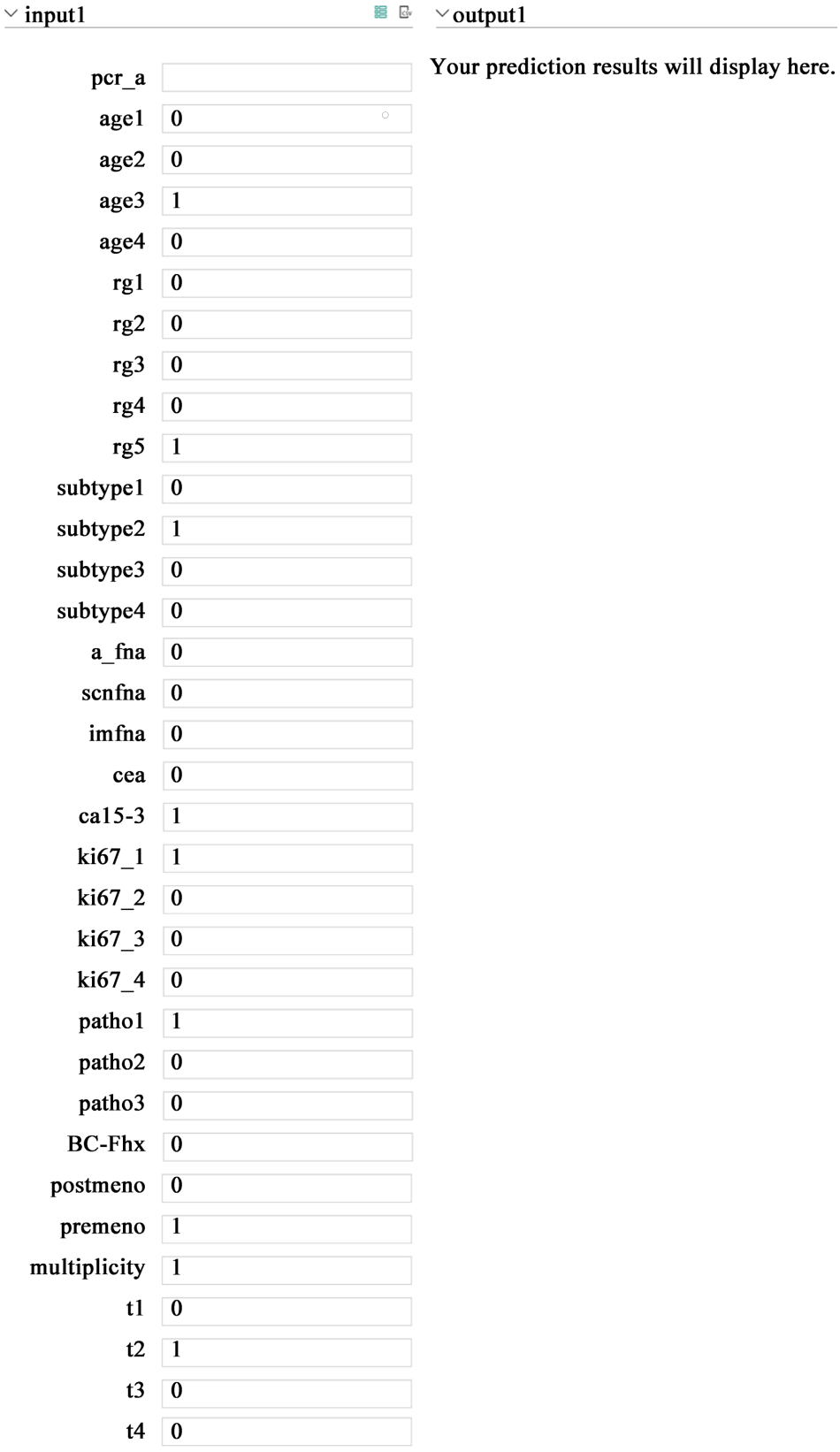

Figure 3. An illustration of web service usage for our predictive tool. For example, when we input data according to each variable excluding the final value "pCR", the predicted value to be calculated in only a few seconds. The meaning of each variable was shown in **Supplementary** Table A1.

for pCR was 0.807 of AUC (p-value = 0.002) [6].

Ki67 was also predictive value for pCR in previous studies [9] [15] [16]. Cabrera *et al.* showed that no reduction of Ki67 significantly increased the hazard ratio of recurrence and death by 3.39 (95% confidence interval [CI] 1.8 - 6.37) in OS and 7.03 in DFS (95%CI 2.6 - 18.7) [9]. Brown *et al.* conducted scoring of Ki67 expression for prediction of response to NAC and showed that both the average and maximum score was directly correlated to pCR (average p-value = 0.0002; maximum p-value = 0.0011) [16].

In addition, Tumor infiltrated lymphocytes (TIL) was associated with pCR in patients treated with NAC [8] [17] [18]. For example, Denkert *et al.* investigated intratumoral lymphocytes in a total 1058 pretherapeutic breast cancer biopsy from two NAC study [17]. Results showed that the percentage of intratumoral lymphocytes was a significant independent parameter for pCR (training cohort; p-value = 0.012; validation cohort p-value = 0.001) [17].

Our data showed that menopause and BRCA2 mutation were associated with pCR. The pCR (+) group had less premenopausal patients (p-value = 0.045) and also less BRCA2 mutation case (p-value = 0.014) than pCR (−) group. There were some study about association between BRCA mutation and pCR [19] [20]. Minckwitz *et al.* revealed that BRCA mutation was predictor for higher pCR rates after NAC (anthracycline/taxane based) in TNBC [19]. According to Arun *et al.*, BRCA1 status was independently associated with higher pCR rates [20]. Among 317 patients who underwent BRCA testing and NAC, 26 of 57 (46%) BRCA1 carriers achieved pCR, compared with 3 of 23 (13%) BRCA2 carrier and 53 of 237 (22%) BRCA non-carriers (p-value < 0.001) [20]. However, the association between menopause and pCR was not confirmed in the previous studies and our study alone cannot sufficiently explain the relationship between menopause and pCR.

We included DCIS in pCR definition. In Mazouni's study, residual DCIS in patients treated NAC does not adversely affect survival or local recurrence rate therefore inclusion of patients with residual DCIS in the definition of pCR is justified [21]. And also, the definition for pCR has been not standardized in clinical trials [22] [23].

Among previous articles, there were some studies revealed that it is difficult to reflect pCR with only clinical variables. Bear *et al.* insisted that there is no clinically useful molecular predictor of response to any cytotoxic drug used in the treatment of breast cancer [24]. Hortobagyi *et al.* also reveal that clinical parameters such as tumor size, estrogen or HER-2 receptor status, histologic or nuclear grade, or the expression of single molecular markers (*i.e.*, Bcl-2, p53, MDR-1, and so on) show weak association with response and are not regimen-specific, which limits their utility in selecting chemotherapy treatment [25]. Our study contains data from gene profiles and it was our one of advantage. There were other studies about prediction of pCR with gene data in patients treated with NAC. Wang *et al.* used lncRNA signature to predict pCR rate [10] and Ayer *et al.* selected a 74-gene k-NN model for predictors of pCR to T/FAC neoadjuvant

therapy [26]. Overall, a 78% (14 of 18) predictive accuracy was observed, with a 100% (three of three) positive predictive value for pCR, a 73% (11 of 15) negative predictive value, a sensitivity of 43% (three of seven), and a specificity of 100% (11 of 11) [26].

Our study is valuable as the analysis contained not only clinical findings, but also gene profiles, and was developed with machine learning. After deploying the Azure ML predictive model as a web service, we used a Representational State Transfer application programming interface to send data and obtain predictions in real-time. Meanwhile, variable factors were measured according to the official international standard but there could be minimal differences among centers. Our predictive model can incorporate data from other centers and still provide proper results for each center, so any disparity among centers or hospitals could be diminished. In addition, our predictive model showed reliable result. The accuracy was 0.875 in modeling and 0.810 in validation group. Through additional predictive model without gene profiles, the accuracy was 0.806 in modeling and 0.778 in validation group. If the center is not able to use gene data, you had better use the second model. Moreover, we made a prediction model using only the variables that can be obtained before NAC. Our model did not require data during NAC therefore it is more consistent with the meaning of prediction before NAC.

Our study has several limitations. First, the number of patients enrolled in this study was relatively small. Because we used data from patients underwent NAC, surgery but also Cancer SCAN™. If we did not use gene data, we would have enrolled more than one thousand patients treated with NAC followed by surgery. Second, only internal validation was performed. To increase study reliability, an analysis with a larger number of patients and external validation regardless of race is needed.

5. Conclusion

Our predictive model presented a useful and easy-to-access tool for the prediction of pCR in breast cancer patients treated with NAC. After additional evaluation with a larger patient group and external validation, our model could be more widely used.

Acknowledgements

This study was presented for oral presentation session in Global Breast Cancer Conference 2019.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Loibl, S., Denkert, C. and von Minckwitz, G. (2015) Neoadjuvant Treatment of Breast

Cancer—Clinical and Research Perspective. *Breast*, **24**, S73-S77. https://doi.org/10.1016/j.breast.2015.07.018

[2] Kummel, S., Holtschmidt, J. and Loibl, S. (2014) Surgical Treatment of Primary Breast Cancer in the Neoadjuvant Setting. *British Journal of Surgery*, **101**, 912-924. https://doi.org/10.1002/bjs.9545

[3] Cameron, D.A., Anderson, E.D., Levack, P., Hawkins, R.A., Anderson, T.J., Leonard, R.C., *et al.* (1997) Primary Systemic Therapy for Operable Breast Cancer—10-Year Survival Data after Chemotherapy and Hormone Therapy. *British Journal of Cancer*, **76**, 1099-1105. https://doi.org/10.1038/bjc.1997.514

[4] Liedtke, C., Mazouni, C., Hess, K.R., Andre, F., Tordai, A., Mejia, J.A., *et al.* (2008) Response to Neoadjuvant Therapy and Long-Term Survival in Patients with Triple-Negative Breast Cancer. *Journal of Clinical Oncology*, **26**, 1275-1281. https://doi.org/10.1200/JCO.2007.14.4147

[5] Cortazar, P., Zhang, L., Untch, M., Mehta, K., Costantino, J.P., Wolmark, N., *et al.* (2014) Pathological Complete Response and Long-Term Clinical Benefit in Breast Cancer: The CTNeoBC Pooled Analysis. *Lancet*, **384**, 164-172. https://doi.org/10.1016/S0140-6736(13)62422-8

[6] Lee, J., Kim, S.H. and Kang, B.J. (2018) Pretreatment Prediction of Pathologic Complete Response to Neoadjuvant Chemotherapy in Breast Cancer: Perfusion Metrics of Dynamic Contrast Enhanced MRI. *Scientific Reports*, **8**, Article No. 9490. https://doi.org/10.1038/s41598-018-27764-9

[7] Weber, J.J., Jochelson, M.S., Eaton, A., Zabor, E.C., Barrio, A.V., Gemignani, M.L., *et al.* (2017) MRI and Prediction of Pathologic Complete Response in the Breast and Axilla after Neoadjuvant Chemotherapy for Breast Cancer. *Journal of the American College of Surgeons*, **225**, 740-746. https://doi.org/10.1016/j.jamcollsurg.2017.08.027

[8] Asano, Y., Kashiwagi, S., Goto, W., Takada, K., Takahashi, K., Hatano, T., *et al.* (2018) Prediction of Treatment Response to Neoadjuvant Chemotherapy in Breast Cancer by Subtype Using Tumor-Infiltrating Lymphocytes. *Anticancer Research*, **38**, 2311-2321. https://doi.org/10.21873/anticanres.12476

[9] Cabrera-Galeana, P., Munoz-Montano, W., Lara-Medina, F., Alvarado-Miranda, A., Perez-Sanchez, V., Villarreal-Garza, C., *et al.* (2018) Ki67 Changes Identify Worse Outcomes in Residual Breast Cancer Tumors after Neoadjuvant Chemotherapy. *Oncologist*, **23**, 670-678. https://doi.org/10.1634/theoncologist.2017-0396

[10] Wang, G., Chen, X., Liang, Y., Wang, W. and Shen, K. (2017) A Long Noncoding RNA Signature That Predicts Pathological Complete Remission Rate Sensitively in Neoadjuvant Treatment of Breast Cancer. *Translational Oncology*, **10**, 988-997. https://doi.org/10.1016/j.tranon.2017.09.005

[11] Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I. (2015) Machine Learning Applications in Cancer Prognosis and Prediction. *Computational and Structural Biotechnology Journal*, **13**, 8-17. https://doi.org/10.1016/j.csbj.2014.11.005

[12] Shin, H.T., Choi, Y.L., Yun, J.W., Kim, N.K.D., Kim, S.Y., Jeon, H.J., *et al.* (2017) Prevalence and Detection of Low-Allele-Fraction Variants in Clinical Cancer Samples. *Nature Communications*, **8**, Article No. 1377. https://doi.org/10.1038/s41467-017-01470-y

[13] Miller, S., Curran, K. and Lunney, T. (2016) Cloud-Based Machine Learning for the Detection of Anonymous Web Proxies. 2016 27*th Irish Signals and Systems Conference*, Londonderry, 21-22 June 2016, 1-6. https://doi.org/10.1109/ISSC.2016.7528443

[14] Chang, E., Goh, K., Sychay, G. and Wu, G. (2003) CBSA: Content-Based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines. *IEEE Transactions on Circuits and Systems for Video Technology*, **13**, 26-38. https://doi.org/10.1109/TCSVT.2002.808079

[15] Yerushalmi, R., Woods, R., Ravdin, P.M., Hayes, M.M. and Gelmon, K.A. (2010) Ki67 in Breast Cancer: Prognostic and Predictive Potential. *The Lancet Oncology*, **11**, 174-183. https://doi.org/10.1016/S1470-2045(09)70262-1

[16] Brown, J.R., DiGiovanna, M.P., Killelea, B., Lannin, D.R. and Rimm, D.L. (2014) Quantitative Assessment Ki-67 Score for Prediction of Response to Neoadjuvant Chemotherapy in Breast Cancer. *Laboratory Investigation*, **94**, 98-106. https://doi.org/10.1038/labinvest.2013.128

[17] Denkert, C., Loibl, S., Noske, A., Roller, M., Muller, B., Komor, M., *et al.* (2010) Tumor-Associated Lymphocytes as an Independent Predictor of Response to Neoadjuvant Chemotherapy in Breast Cancer. *Journal of Clinical Oncology*, **28**, 105-113. https://doi.org/10.1200/JCO.2009.23.7370

[18] Yamaguchi, R., Tanaka, M., Yano, A., Tse, G.M., Yamaguchi, M., Koura, K., *et al.* (2012) Tumor-Infiltrating Lymphocytes Are Important Pathologic Predictors for Neoadjuvant Chemotherapy in Patients with Breast Cancer. *Human Pathology*, **43**, 1688-1694. https://doi.org/10.1016/j.humpath.2011.12.013

[19] Von Minckwitz, G., Hahnen, E., Fasching, P.A., Hauke, J., Schneeweiss, A., Salat, C., *et al.* (2014) Pathological Complete Response (pCR) Rates after Carboplatin-Containing Neoadjuvant Chemotherapy in Patients with Germline BRCA (g BRCA) Mutation and Triple-Negative Breast Cancer (TNBC): Results from GeparSixto. *Journal of Clinical Oncology*, **32**, 1005. https://doi.org/10.1200/jco.2014.32.15_suppl.1005

[20] Arun, B., Bayraktar, S., Liu, D.D., Gutierrez Barrera, A.M., Atchley, D., Pusztai, L., *et al.* (2011) Response to Neoadjuvant Systemic Therapy for Breast Cancer in BRCA Mutation Carriers and Noncarriers: A Single-Institution Experience. *Journal of Clinical Oncology*, **29**, 3739-3746. https://doi.org/10.1200/JCO.2011.35.2682

[21] Mazouni, C., Peintinger, F., Wan-Kau, S., Andre, F., Gonzalez-Angulo, A.M., Symmans, W.F., *et al.* (2007) Residual Ductal Carcinoma *in Situ* in Patients with Complete Eradication of Invasive Breast Cancer after Neoadjuvant Chemotherapy Does Not Adversely Affect Patient Outcome. *Journal of Clinical Oncology*, **25**, 2650-2655. https://doi.org/10.1200/JCO.2006.08.2271

[22] Fumagalli, D., Bedard, P.L., Nahleh, Z., Michiels, S., Sotiriou, C., Loi, S., *et al.* (2012) A Common Language in Neoadjuvant Breast Cancer Clinical Trials: Proposals for Standard Definitions and Endpoints. *The Lancet Oncology*, **13**, e240-e248. https://doi.org/10.1016/S1470-2045(11)70378-3

[23] von Minckwitz, G., Rezai, M., Loibl, S., Fasching, P.A., Huober, J., Tesch, H., *et al.* (2010) Capecitabine in Addition to Anthracycline- and Taxane-Based Neoadjuvant Treatment in Patients with Primary Breast Cancer: Phase III GeparQuattro Study. *Journal of Clinical Oncology*, **28**, 2015-2023. https://doi.org/10.1200/JCO.2009.23.8303

[24] Bear, H.D., Anderson, S., Brown, A., Smith, R., Mamounas, E.P., Fisher, B., *et al.* (2003) The Effect on Tumor Response of Adding Sequential Preoperative Docetaxel to Preoperative Doxorubicin and Cyclophosphamide: Preliminary Results from National Surgical Adjuvant Breast and Bowel Project Protocol B-27. *Journal of Clinical Oncology*, **21**, 4165-4174. https://doi.org/10.1200/JCO.2003.12.005

[25] Hortobagyi, G., Hayes, D. and Pusztai, L. (2002) Integrating Newer Science into Breast Cancer Prognosis and Treatment: A Review of Current Molecular Predictors and Profiles. *American Society of Clinical Oncology* 2002 *Annual Meeting Summaries*, Cicago, 192-201.

[26] Ayers, M., Symmans, W.F., Stec, J., Damokosh, A.I., Clark, E., Hess, K., *et al.* (2004) Gene Expression Profiles Predict Complete Pathologic Response to Neoadjuvant Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide Chemotherapy in Breast Cancer. *Journal of Clinical Oncology*, **22**, 2284-2293. https://doi.org/10.1200/JCO.2004.05.166

Supplementary Table A1. Variables for Clinical Application

Variables	Character of variables	Meaning
Age 1 (yes = 1 or no = 0)	Categorical variable	Age ≤ 35
Age 2 (yes = 1 or no = 0)	Categorical variable	35 < Age ≤ 45
Age 3 (yes = 1 or no = 0)	Categorical variable	45 < Age ≤ 55
Age 4 (yes = 1 or no = 0)	Categorical variable	55 < Age
Rg 1 (yes = 1 or no = 0)	Categorical variable	AC or AC + Taxane regimen
Rg 2 (yes = 1 or no = 0)	Categorical variable	AC + Taxane + Herceptin regimen
Rg 3 (yes = 1 or no = 0)	Categorical variable	AC + Paclitaxel + Carboplatin regimen
Rg 4 (yes = 1 or no = 0)	Categorical variable	TCHP (docetaxel + carboplatin + tastuzumab + pertuzumab) regimen
Rg 5 (yes = 1 or no = 0)	Categorical variable	Other regimen
Subtype 1 (yes = 1 or no = 0)	Categorical variable	HR+/HER2−
Subtype 2 (yes = 1 or no = 0)	Categorical variable	HR+/HER2+
Subtype 3 (yes = 1 or no = 0)	Categorical variable	HR−/HER2+
Subtype 4 (yes = 1 or no = 0)	Categorical variable	HR−/HER2−
Afna (clinical N0 = 0, axillary FNA positive = 1, axillary FNA negative = 2)	Categorical variable	Axillary FNA
Scnfna (SCN FNA undone & negative = 0, SCN FNA positive = 1)	Categorical variable	Supra clavicula lymph node FNA
Imfna (IMLN FNA undone & negative = 0, IMLN FNA positive = 1)	Categorical variable	Internal mammary lymph node FNA
Cea (elevation = 1, normal range or lower range = 0)	Categorical variable	Carcinoembryonic antigen; 7 ng/ml <
Ca15-3 (elevation = 1, normal range or lower range = 0)	Categorical variable	Carcinoma antigen 15-3; 30 U/ml <
Ki-67 1 (yes = 1 or no = 0)	Categorical variable	Percentage of positive cell by IHC staining, Ki-67 < 25%
Ki-67 2 (yes = 1 or no = 0)	Categorical variable	Ki-67 25≤ <50%
Ki-67 3 (yes = 1 or no = 0)	Categorical variable	Ki-67 50≤ <75%
Ki-67 4 (yes = 1 or no = 0)	Categorical variable	Ki-67 75≤ <100%
Patho 1 (yes = 1 or no = 0)	Categorical variable	Invasive ductal carcinoma
Patho 2 (yes = 1 or no = 0)	Categorical variable	Invasive lobular carcinoma
Patho 3 (yes = 1 or no = 0)	Categorical variable	Mixed type
BC-Fhx (yes = 1 or no = 0)	Categorical variable	Family history of breast cancer
Postmeno (yes = 1 or no = 0)	Categorical variable	Post menopausal
Premeno (yes = 1 or no = 0)	Categorical variable	Pre menopausal
Multiplicity (yes = 1 or no = 0)	Categorical variable	
Clinical T1 (yes = 1 or no = 0)	Categorical variable	Tumor size ≤ 2 cm
Clinical T2 (yes = 1 or no = 0)	Categorical variable	2 cm < Tumor size ≤ 5 cm
Clinical T3 (yes = 1 or no = 0)	Categorical variable	5 cm < Tumor size
Clinical T4 (yes = 1 or no = 0)	Categorical variable	Invasion to skin or chest wall