Supplemental Data

Study design and participants

In the PHERGain trial, eligible participants were aged 18 years or older with centrally confirmed, HER2+ disease according to the 2018 American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP) criteria, stage I–IIIA, invasive, operable breast cancer (≥1.5 cm tumour size) with at least one breast lesion evaluable by [^{18}F]FDG-PET/CT (SUVmax≥1.5 x SUVmean liver + 2 standard deviation [SD]), an Eastern Cooperative Oncology Group (ECOG) performance status of 0 or 1, and a baseline left ventricular ejection fraction of at least 55%. Randomization was stratified by hormone receptor (HR) status. Between June 26, 2017, and April 24, 2019, 356 patients were randomly assigned (1:4) to arms A and B in the PHERGain trial. Patients in arm A (n=71) were treated with HP plus docetaxel and carboplatin regardless of their metabolic response after two cycles of chemotherapy + HP. All patients assigned to arm B initially received two cycles of HP (± endocrine therapy). Consecutively, metabolic responders in arm B (n=227) continued HP (± endocrine therapy) treatment for six further cycles; metabolic non-responders in this arm (n=58) were switched to six cycles of HP plus docetaxel and carboplatin. Adjuvant treatment was selected according to the neoadjuvant treatment administered, pathologic response, HR status, and clinical stage at diagnosis (*1*).

[^{18}F]FDG-PET/CT

[^{18}F]FDG-PET/CT scans were obtained with a maximum of seven days between baseline imaging and treatment initiation, according to study protocol. Each ^{18}F-FDG-PET/CT scan reading from the two independent reviewers underwent a thorough consensus process, in which it was extensively reviewed and approved. Discrepancies were resolved through input from a third reviewer.

[^{18}F]FDG-PET/CT procedures were based on the European Association of Nuclear Medicine (EANM) through their EANM Research Limited (EARL) subsidiary (*2*). All sites were EARL accredited. All [^{18}F]FDG-PET/CT scans had to be done on the same approved device of the imaging centre with identical acquisition and reconstruction settings. SUV was calculated as the ratio of tissue radioactivity concentration to the administered dose, divided by bodyweight. SUVmax was defined as the highest pixel SUV within a tumour. For each target lesion, metabolic response was calculated as the percentage decrease from baseline using the formula: SUVmax at baseline – SUVmax at follow-up/SUVmax at baseline.

PHERGain study procedures

At baseline, the primary tumour was evaluated locally at each investigator site for feasibility of surgery. In accordance with the National Comprehensive Cancer Network Clinical Practice Guidelines in Oncology 2017, axillary ultrasound with or without fine-needle aspiration at baseline and surgical staging of the axilla during breast surgery were mandatory. Surgery was done 2–6 weeks after the last dose of study treatment (*1*).

Baseline core biopsies of the primary tumour were taken for histological diagnosis. Local assessment of oestrogen receptor (ER) and progesterone receptor (PR), and central assessment of HER2 status were performed. Tumours

with ER and/or PR expression of 1% or more were defined as HR-positive. HER2 status was regarded as positive if the immunohistochemistry (IHC) result was 3+ or 2+ and confirmed by positive in-situ hybridization.

Statistical considerations

The prespecified analysis aimed to select the best cut-off value of ΔSUV_{max} to predict pCR in HP arm. So, the analysis was conducted in patients who have been allocated in arm B and avoid chemotherapy because they fulfilled the predefined metabolic response definition (cut-off 40%) of the PHERGain study (primary analysis set, n=227).

The ΔSUVmax capability of predicting pCR was evaluated in terms of area under the receiver operating characteristic curve (AUC). The 95% confidence interval (95% CI) was calculated using the De Long method. Based on Rice and Harris, we have considered that the predictive capacity of ΔSUV_{max} would not be acceptable with AUC values less than or equal to 63.9%. Alternatively, AUC values greater or equal than 71.4% were considered appropriate (*3*). The study will meet this target predictive capability if the lower boundary of the 95% CI for AUC is greater than 63.9%. The 227 metabolic responders in arm B will attain a 5% precision (half with of the confidence interval) to detect an AUC of 71.4% (*4*).

The selection of best cut-off of ΔSUV_{max} was based on sensitivity and specificity. Sensitivity was the number of patients achieving a pCR with a ΔSUV_{max} value higher than the cut-off (true positives for ΔSUV_{max}) among all patients with pCR. The specificity was the number of patients not achieving pCR with a ΔSUV_{max} value lower than the cut-off (true negatives for ΔSUV_{max}) among all patients without pCR). Both measures were combined to calculate the balanced accuracy: (Sensitivity+Specificity)/2. We selected the ΔSUV_{max} cut-off that achieved the highest values of balanced accuracy.

Best cut-off for the PHERGain sample

We randomly split the dataset for arm B and metabolic responders in training dataset (with 80% of patients) and test data set (with 20% of patients). The best cut-off of the ΔSUV_{max} was calculated in the training dataset with the 10 repeated 10-fold cross-validation method. The best cut-off selected in the training data set was evaluated in the test dataset.

Best cut-off for the target population

The best cut-off selected for the PHERGain sample was corrected based on the correlation between two independent reviewers (r) to account inter evaluator differences (*5*). The 95% CI for the optimal cut-off of ΔSUV_{max} was calculated as follows:

Lower boundary = Sample cut-off – 1.96 * $SD_{\Delta SUV_{max}}$ * $\sqrt{(1-r)}$

Upper boundary = Sample cut-off + 1.96 * $SD_{\Delta SUV,max}$ * $\sqrt{(1-r)}$

Were $SD_{\Delta SUV,max}$ is the standard deviation of ΔSUV_{max} and 1.96 is z-score for a 0.025 one-sided alpha level.

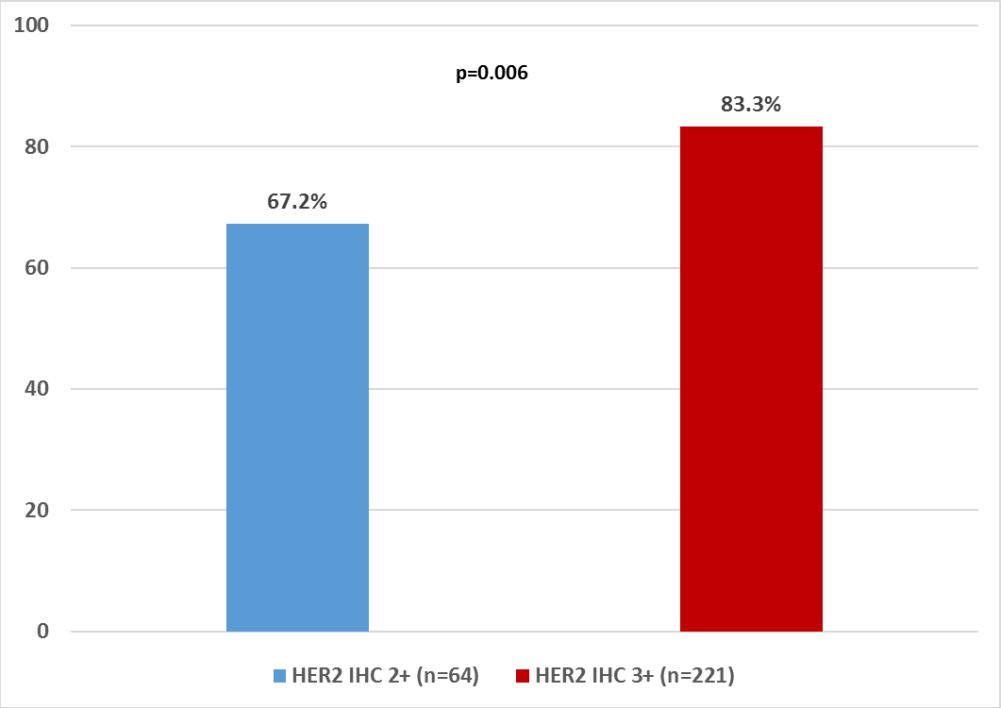
The upper boundary for the confidence interval was defined as the final cut-off for the target population. It maximizes the probability of achieving a pCR in the patients selected to avoid chemotherapy in comparison with the lower boundary for confidence interval.

The clinical outcomes have been compared between the PHERGain cut-off (40%), the best cut-off for the PHERGain sample, and the cut-offs selected for target population (lower and upper boundaries). The number of patients achieving pCR has been reported for -5%, 0, and 5% to 90% ΔSUV_{max} values in patients avoiding chemotherapy (primary analysis set, n=227), in patients with chemotherapy initiated after two treatment cycles (metabolic non-responders in arm B, n=58), and in patients receiving chemotherapy from the beginning of the study (arm A, n=71). Only metabolic non-responders have missing ΔSUV_{max} values. They have been imputed with the mean value for metabolic non-responders in their study arm to preserve the Intention to treat principle. Results for imputed and raw analyses were presented.

Continuous variables are expressed as median and interquartile range (IQR) represented by the 25–75 percentile and categorical variables as, absolute value and percentage. The 95% CI for categorical variables were calculated with Clopper-Pearson method. Patient characteristics for metabolic responders in arm B were compared with pCR and without pCR. Patient characteristics for arm B were compared between metabolic responders and non-responders. Clinical outcomes as pCR and rate of patients with chemotherapy were compared between study arms. These comparisons were conducted with a logistic regression model based on Wald test. The correlation between ΔSUV_{max} after two cycles from the reviewer 1 and 2 was analysed with Pearson method and the 95% CI.

The 95% CI for the AUC and optimal cut-off estimation of ΔSUV_{max} were reported to align with the one-sided primary hypothesis test for the neoadjuvant setting in the PHERGain trial. This primary endpoint was met at a nominal α level of 2.5% (*1*). All other statistical tests were two-sided and considered statistically significant at $p<0.05$. The exploratory findings were set at a two-sided 0.1 level and were considered as trends toward significance. All statistical analyses were done with R version 4.02 (2020-06-22).

Supplemental Figure 1: The rate of [^{18}F]FDG-PET/CT responders between patients with HER2 IHC 3+ and 2+ status.



IHC=immunohistochemistry. The analysis has been conducted with logistic regression model based on Wald test.

Supplemental Table 1. Demographic and baseline clinical characteristics in patients avoiding neoadjuvant chemotherapy ([18F]FDG-PET/CT responders in arm B) by pCR status, according to the original cut-off.

	All patients (N=227)	With pCR (N=86)	Without pCR (N=141)	p-value
Median age (years, IQR)	51 (45.0 to 59.0)	52 (45.2 to 60.8)	50 (45.0 to 58.0)	0.372
<=50 years	110 (48.5%)	38 (44.2%)	72 (51.1%)	0.315
>50 years	117 (51.5%)	48 (55.8%)	69 (48.9%)	
Postmenopausal				
No	117 (51.5%)	41 (47.7%)	76 (53.9%)	0.363
Yes	110 (48.5%)	45 (52.3%)	65 (46.1%)	
ECOG				
0	209 (92.1%)	78 (90.7%)	131 (92.9%)	0.551
1	18 (7.9%)	8 (9.3%)	10 (7.1%)	
Stage				
I	21 (9.3%)	10 (11.6%)	11 (7.8%)	0.34
II	173 (76.2%)	64 (74.4%)	109 (77.3%)	
IIIA	33 (14.5%)	12 (14.0%)	21 (14.9%)	
Nodal status				
Negative	117 (51.5%)	49 (57.0%)	68 (48.2%)	0.201
Positive	110 (48.5%)	37 (43.0%)	73 (51.8%)	
Hormone receptor status				
ER -/PR -	70 (30.8%)	31 (36.0%)	39 (27.7%)	0.185
ER+ and/or PR+	157 (69.2%)	55 (64.0%)	102 (72.3%)	
HER2 IHC score and FISH analysis				
2+ and FISH+	43 (18.9%)	11 (12.8%)	32 (22.7%)	*0.068*
3+	184 (81.1%)	75 (87.2%)	109 (77.3%)	
Ki67%				
≤20%	39 (17.2%)	13 (15.1%)	26 (18.4%)	0.597
>20%	177 (78.0%)	67 (77.9%)	110 (78.0%)	
Not evaluated	11 (4.8%)	6 (7.0%)	5 (3.5%)	
Tumour Grade				
GI (well differ.)	4 (1.8%)	3 (3.5%)	1 (0.7%)	0.163
GII (moderately differ.)	86 (37.9%)	36 (41.9%)	50 (35.5%)	
GIII (poorly differ.)	105 (46.3%)	38 (44.2%)	67 (47.5%)	
Gx	32 (14.1%)	9 (10.5%)	23 (16.3%)	
Median SUV_{max} at baseline (IQR)	10.4 (6.2–15.5)	8.8 (5.6–15.5)	10.8 (6.8–15.4)	0.248
Median SUV_{max} at 6 weeks (IQR)	2.2 (1.3–3.6)	1.6 (0.7–2.6)	2.7 (1.7–4.7)	**<0.001**
ΔSUV_{max} Median (IQR)	69.6% (57.5 to 79.9)	77.8% (67.0 to 85.4)	63.3% (54.8 to 74.8)	**<0.001**
Mean ± SD **	68.5% ± 14.6	75.20% ± 14.1	64.4% ± 13.6	

Data are n (%), unless otherwise specified. ΔSUV_{max}= SUV_{max} reduction at 6 weeks (or 2 cycles). ECOG=Eastern Cooperative Oncology Group. ER=Oestrogen receptor. FISH=Fluorescence In Situ Hybridization. G= Tumour grade. Gx=Tumour grade cannot be assessed. IHC=Immunohistochemistry. IQR=Interquartile range. pCR=Pathological complete response. PR=Progesterone receptor. SD=Standard deviation. SUV_{max}=Maximum standardized uptake value.

Patient characteristics were compared between patients with pCR and without pCR. We have used a logistic regression model based on Wald test.

Standard deviation in all patients is used to calculate the 95% CI for the optimal cut-off of ΔSUV_{max}.

Supplemental Table 2: Patient characteristics in arm B by [^{18}F]FDG-PET/CT responders and non-responders.

	All patients (n=285)	[^{18}F]FDG-PET/CT Responder (n=227)	[^{18}F]FDG-PET/CT Non-Responder (n=58)	p-value**
Median age (years, IQR)	50 (45 to 59)	51 (45 to 59)	48 (45 to 60)	0.524
<=50 years	146 (51.2%)	110 (48.5%)	36 (62.1%)	*0.066*
>50 years	139 (48.8%)	117 (51.5%)	22 (37.9%)	
Postmenopausal				
No	146 (51.2%)	117 (51.5%)	29 (50%)	0.834
Yes	139 (48.8%)	110 (48.5%)	29 (50%)	
ECOG				
0	264 (92.6%)	209 (92.1%)	55 (94.8%)	0.477
1	21 (7.4%)	18 (7.9%)	3 (5.2%)	
Stage				
I	24 (8.4%)	21 (9.3%)	3 (5.2%)	0.293
II	219 (76.8%)	173 (76.2%)	46 (79.3%)	
IIIA	42 (14.7%)	33 (14.5%)	9 (15.5%)	
Nodal status				
Negative	145 (50.9%)	117 (51.5%)	28 (48.3%)	0.657
Positive	140 (49.1%)	110 (48.5%)	30 (51.7%)	
Hormone receptor status				
ER- / PR-	93 (32.6%)	70 (30.8%)	23 (39.7%)	0.203
ER+ or PR+ or both	192 (67.4%)	157 (69.2%)	35 (60.3%)	
HER2 IHC score and FISH analysis				
2+ and FISH positive	64 (22.5%)	43 (18.9%)	21 (36.2%)	**0.006**
3+	221 (77.5%)	184 (81.1%)	37 (63.8%)	
Ki67%				
≤20%	51 (17.9%)	39 (17.2%)	12 (20.7%)	0.525
>20%	220 (77.2%)	177 (78%)	43 (74.1%)	
Not evaluated	14 (14.9%)	11 (4.8%)	3 (5.2%)	
Tumour Grade				
GI (well differentiated)	6 (2.1%)	4 (1.8%)	2 (3.4%)	0.433
GII (moderately differentiated)	109 (38.2%)	86 (37.9%)	23 (39.7%)	
GIII (poorly differentiated)	127 (44.6%)	105 (46.3%)	22 (37.9%)	
Gx*	43 (15.1%)	32 (14.1%)	11 (19%)	
Median SUV_{max} at baseline (IQR)	10.4 (6.4–15.9)	10.4 (6.2–15.5)	11.2 (7.4–18.2)	0.238
Median SUV_{max} at 6 weeks (IQR)	2.8(1.6-5)	2.2 (1.3–3.6)	8 (4.8-14.7)	**<0.001**
Median ΔSUV_{max} (IQR) ***	63.7%(44.3 to 77.6)	69.6% (57.5 to 79.9)	18.2% (2.7 to 30.8)	**<0.001**

Data are n (%), unless otherwise specified. ΔSUV_{max}=SUV_{max} reduction at 6 weeks (or two cycles). ER=oestrogen receptor. Gx=Tumour grade cannot be assessed. HP=trastuzumab and pertuzumab. IHC=immunohistochemistry. PR=progesterone receptor. SD=standard deviation. SUV_{max}=the maximum Standardized Uptake Value

Patient characteristics were compared between patients with FDG-PET responders and non-responders. We have used a logistic regression model based on Wald test.

Standard deviation in all patients is used to calculate the 95% CI for the optimal cut-off of ΔSUV_{max}.

Supplemental Table 3: Patient characteristics in arm B between patients [^{18}F]FDG-PET responders and non-responders

- **Methods**

Lower boundary = Sample cut-off – 1.96 * $SD_{\Delta SUV_{max}}$ * √(1-r)

Upper boundary = Sample cut-off + 1.96 * $SD_{\Delta SUV,max}$ * √(1-r)

Where "r" is the correlation between two independent reviewers (0.974), the $SD_{\Delta SUV,max}$ is the standard deviation of ΔSUV_{max} (14.6), the sample cut-off is 72.6%, and 1.96 is z-score for a 0.025 one-sided alpha level.

- **Results**

Lower boundary = 72.6% – 1.96*(14.6 * √(1-0.974)) = 68.0%Point estimation = 72.6%

Upper boundary = 72.6 + 1.96*(14.6 * √(1-0.974)) = 77.2% ≈ 77%

Supplemental Table 4: Patient characteristics in arm B between patients [^{18}F]FDG-PET/CT responders and non-responders based on the ΔSUV_{max} cut-off of ≥ 77%.

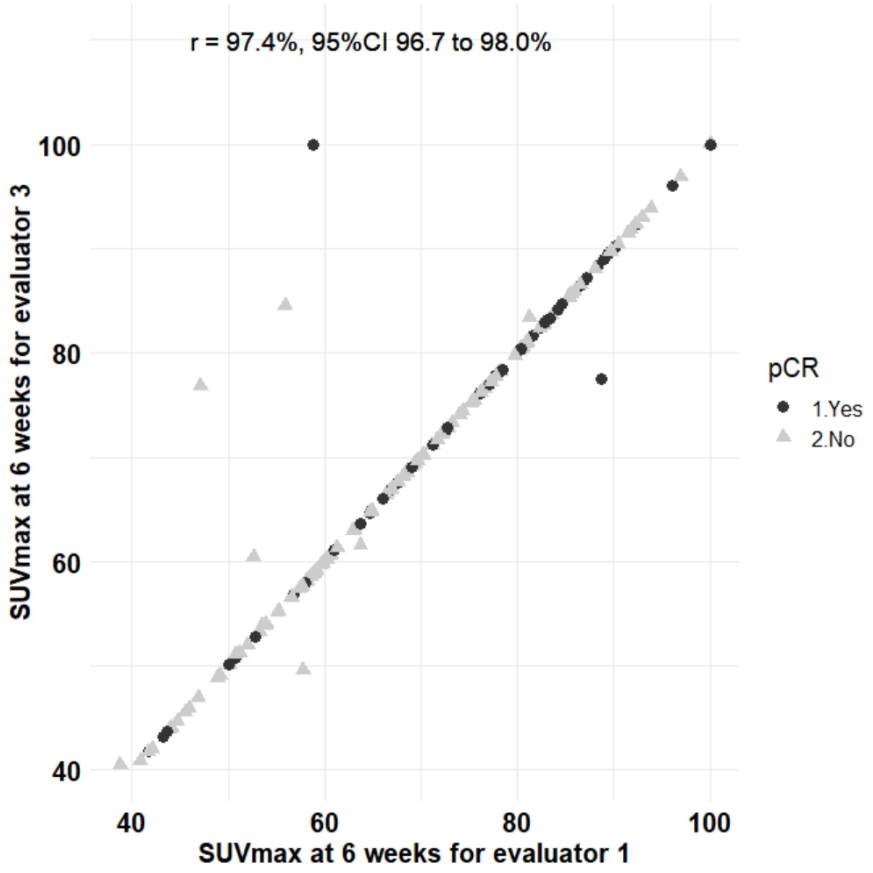
	All patients (n=285)	[^{18}F]FDG-PET/CT Responder (ΔSUV_{max} ≥ 77) (n=74)	[^{18}F]FDG-PET/CT (ΔSUV_{max} < 77) (n=211)	p-value**
Median age (years, IQR)	50 (45 to 59)	53 (47.2 to 61.8)	50 (44 to 58)	**0.015**
<=50 years	146 (51.2%)	27 (36.5%)	119 (56.4%)	**0.004**
>50 years	139 (48.8%)	47 (63.5%)	92 (43.6%)	
Postmenopausal				
No	146 (51.2%)	33 (44.6%)	113 (53.6%)	0.186
Yes	139 (48.8%)	41 (55.4%)	98 (46.4%)	
ECOG				
0	264 (92.6%)	71 (95.9%)	193 (91.5%)	0.215
1	21 (7.4%)	3 (4.1%)	18 (8.5%)	
Stage				
I	24 (8.4%)	8 (10.8%)	16 (7.6%)	0.401
II	219 (76.8%)	56 (75.7%)	163 (77.3%)	
IIIA	42 (14.7%)	10 (13.5%)	32 (15.2%)	
Nodal status				
Negative	145 (50.9%)	39 (52.7%)	106 (50.2%)	0.715
Positive	140 (49.1%)	35 (47.3%)	105 (49.8%)	
Hormone receptor status				
ER- / PR-	93 (32.6%)	31 (41.9%)	62 (29.4%)	**0.05**
ER+ or PR+ or both	192 (67.4%)	43 (58.1%)	149 (70.6%)	
HER2 IHC score and FISH analysis				
2+ and FISH positive	64 (22.5%)	8 (10.8%)	56 (26.5%)	**0.007**
3+	221 (77.5%)	66 (89.2%)	155 (73.5%)	
Ki67%				
≤20%	51 (17.9%)	11 (14.9%)	40 (19%)	0.441
>20%	220 (77.2%)	59 (79.7%)	161 (76.3%)	
Not evaluated	14 (14.9%)	4 (5.4%)	10 (4.7%)	
Tumour Grade				
GI (well differentiated)	6 (2.1%)	1 (1.4%)	5 (2.4%)	0.604
GII (moderately differentiated)	109 (38.2%)	29 (39.2%)	80 (37.9%)	
GIII (poorly differentiated)	127 (44.6%)	33 (44.6%)	94 (44.5%)	
Gx*	43 (15.1%)	11 (14.9%)	32 (15.2%)	
Median SUV_{max} at baseline (IQR)	10.4 (6.4–15.9)	11.4 (7.9–20.1)	10 (5.9–15.5)	**0.027**
Median SUV_{max} at 6 weeks (IQR)	2.8 (1.6-5)	1.1 (0.7–2.1)	8 (4.8-14.7)	**<0.001**
Median ΔSUV_{max} (IQR) ***	63.7% (44.3 to 77.6)	83.8% (80.4 to 86.9)	56% (38.4 to 66.7)	**<0.001**

Data are n (%), unless otherwise specified. ΔSUV_{max}=SUV_{max} reduction at 6 weeks (or two cycles). ER=oestrogen receptor. Gx=Tumour grade cannot be assessed. HP=trastuzumab and pertuzumab. IHC=immunohistochemistry. PR=progesterone receptor. SD=standard deviation. SUV_{max}=the maximum Standardized Uptake Value

Patient characteristics were compared between patients with FDG-PET responders and non-responders. We have used a logistic regression model based on Wald test.

Standard deviation in all patients is used to calculate the 95% CI for the optimal cut-off of ΔSUV_{max}.

Supplemental Figure 2: Correlation between ΔSUV$_{max}$ at 6 weeks between reviewer 1 and 2 by pCR in [^{18}F]FDG-PET/CT responders in arm B.



pCR= pathological complete response. r= Pearson correlation coefficient.

References

1. Pérez-García JM, Gebhart G, Ruiz Borrego M, et al. Chemotherapy de-escalation using an 18F-FDG-PET-based pathological response-adapted strategy in patients with HER2-positive early breast cancer (PHERGain): a multicentre, randomised, open-label, non-comparative, phase 2 trial. *The Lancet Oncology*. 2021;22:858-871.
2. Bossuyt V, Provenzano E, Symmans WF, et al. Recommendations for standardized pathological characterization of residual disease for neoadjuvant clinical trials of breast cancer by the BIG-NABCG collaboration. *Ann Oncol*. 2015;26:1280-1291.
3. Rice ME, Harris GT. Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law Hum Behav*. 2005;29:615-620.
4. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. *J Biomed Inform*. 2014;48:193-204.
5. Charter RA, Feldt LS. Confidence Intervals for True Scores: Is There a Correct Approach? *Journal of Psychoeducational Assessment*. 2001;19:350-364.