The quest for biomarkers: Computational exploration of the cancer genome

de Ronde, J.J.

Citation for published version (APA):
BREAST CANCER SUBTYPE SPECIFIC CLASSIFIERS OF RESPONSE TO NEOADJUVANT CHEMOTHERAPY DO NOT OUTPERFORM CLASSIFIERS TRAINED ON ALL SUBTYPES

Jorma J. de Ronde¹,²,‡, Marc Jan Bonder¹,‡, Esther H. Lips²,³, Sjoerd Rodenhuis⁴, Lodewyk F.A. Wessels¹,⁵

Department of Molecular Carcinogenesis¹, Department of Molecular Pathology², Department of Pathology³, and Department of Medical Oncology⁴, The Netherlands Cancer Institute, The Netherlands Faculty of EEMCS, Delft University of Technology, Delft, The Netherlands⁵

‡ contributed equally

submitted
ABSTRACT

Introduction: Despite continuous efforts, not a single predictor of breast cancer chemotherapy resistance has made it into the clinic yet. However, it has become clear in recent years that breast cancer is a collection of molecularly distinct diseases. With ever increasing amounts of breast cancer data becoming available, we set out to study if gene expression based predictors of chemotherapy resistance that are specific for breast cancer subtypes can improve upon the performance of generic predictors.

Methods: We trained predictors of resistance that were specific for a subtype and generic predictors that were not specific for a particular subtype, i.e. trained on all subtypes simultaneously. Through a rigorous double-loop cross-validation we compared the performance of these two types of predictors on the different subtypes on a large set of tumors all profiled on the same expression platform (n = 394). We evaluated predictors based on either mRNA gene expression or clinical features.

Results: In the context of HER2+, ER- breast cancer, the clinical feature based subtype specific predictor outperformed the generic, non-specific predictor. This can be explained by the fact that the generic predictor included features -HER2 and ER status- that were predictive over the whole set, but not within this subtype. In all other scenarios the generic predictors outperformed the subtype specific predictors or showed equal performance.

Conclusions: Since it depends on the specific context which type of predictor - subtype specific or generic- performed better, it is highly recommended to evaluate both specific and generic predictors when attempting to predict treatment response in breast cancer.
INTRODUCTION

Breast cancer is a heterogeneous disease, which can be subdivided into distinct subtypes. Based on clinical features, including age, tumor grade, and TNM stage, patients can be stratified into more homogeneous subgroups. By measuring estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) expression, we can stratify tumors into a HER2 positive set (HER2+, either ER+ or ER-), a Luminal set (HER2-, ER+) and a Triple Negative set (HER2-, ER-, PR-; TN). These three different subtypes of breast cancer are considered distinct diseases and are approached accordingly in the clinic. Breast cancer patients of all subtypes receive chemotherapy as part of the treatment process.

Neoadjuvant therapy, the administration of therapeutic agents before the main treatment (typically surgery), reduces the mortality of breast cancer patients [1]. However, some patients only experience the downside of the therapy (i.e. toxicity) and not the benefit (i.e. increased survival). It has been shown that patients who achieve a pathological complete response after neoadjuvant treatment have a higher chance of relapse free survival [2, 3]. Because neoadjuvant treatments can be very toxic and not all patients benefit from the treatment, it would be desirable if the non-responders to therapy can be accurately separated from the responders. Multiple approaches to predict the response of patients to neoadjuvant treatment have been undertaken [4-10], a number of which were reviewed in [11]. Unfortunately, up to now none of the published predictors have been applied in the clinical setting. This lack of clinical implementation can be attributed to the fact that the results that were reported in the original papers are either not reproducible in external data sets or the fact that the reported accuracy is not high enough to be used in clinical decision making.

The best performing predictor of chemotherapy response published to date has an AUC of 0.805 [8]. In a subsequent independent validation study, AUCs of 0.711 (T-FAC treatment) and 0.584 (FAC treatment) were achieved for this same predictor [12]. While this is an improvement over previous attempts, the accuracy of this predictor is insufficient for clinical decision making. This classifier was trained on all available samples in the training dataset, irrespective of breast cancer subtype.

In this study, we set out to determine if we could increase the accuracy of the chemotherapy response predictors by creating predictors that are specific for each breast cancer subtype, instead of a predictor that was trained on all subtypes combined. In order to make sure that our results would not be biased towards a certain type of methodology we included six feature selection approaches and six classifiers. These methods reflect commonly used methods in the context of class prediction. Additionally, to answer the question of how generally applicable our results are, we included predictors based on gene expression data and predictors based on clinical features.
METHODS

Samples
For this study we selected three gene expression data sets, all hybridized on either the Affymetrix HG U133A platform or the Affymetrix HG U133 plus 2.0 platform. The three sets consisted of the microarray quality control II (MAQC2) breast cancer dataset[13]; a breast cancer dataset from the Department of Breast and Endocrine Surgery of the Osaka University (BESOU) [14] and a breast cancer dataset from the MD Anderson Cancer Center (MDACC) [12]. For all three datasets, informed consent from the included patients and approval from an ethics committee were obtained (the University of Texas M.D. Anderson Cancer Center ethical committee for the MAQC2 study, the Ethics Review Committee at Osaka University Hospital for the BESOU dataset, and the institutional review boards of each participating institution for the MDACC study). These datasets are publicly available from the Gene Expression Omnibus website (MAQC2 GEO ID is GSE16716, BESOU GEO ID is GSE32646, and MDACC GEO ID is GSE20271). From these sets we extracted the samples that were treated with taxol followed by 5-fluorouracil, (Adriamycin or Epirubicin), cyclophosphamide (T-FAC or T-FEC). Patients that were treated with a regimen including trastuzumab or FAC alone were removed from our dataset in order to get a more homogeneous treatment group. After this selection the dataset consisted of 394 samples in total. Unfortunately not all of these had all clinical data available, so for analyses based on clinical features the total dataset consisted of 374 samples. Supplementary table 1 shows an overview of the characteristics of this set of patients.

In these datasets, a pathological complete response (pCR) was defined as no residual invasive cancer in the breast and axillary lymph nodes. Samples were

<table>
<thead>
<tr>
<th>Stratification</th>
<th>pCR (%)</th>
<th>No pCR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical subtypes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HER2 positive</td>
<td>31 (38)</td>
<td>51 (62)</td>
</tr>
<tr>
<td>Luminal</td>
<td>14 (9)</td>
<td>185 (91)</td>
</tr>
<tr>
<td>Triple Negative</td>
<td>42 (37)</td>
<td>71 (63)</td>
</tr>
<tr>
<td>HER2 positive, ER negative</td>
<td>25 (56)</td>
<td>20 (44)</td>
</tr>
<tr>
<td>HER2 positive, ER positive*</td>
<td>6 (16)</td>
<td>31 (84)</td>
</tr>
<tr>
<td>No stratification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>87 (22)</td>
<td>307 (78)</td>
</tr>
</tbody>
</table>

*The HER2 positive, ER positive group was not included in the analysis due to the small sample size.
defined to be ER-positive or PR-positive when 10% or more of the tumor cells showed positive staining of ER or PR respectively, based on immunohistochemistry. Samples were marked as positive for HER2 when there was strong membrane staining (3+) or the sample had a gene copy number equal to or greater than 2.0 as measured by FISH (FISH ratio greater than 2.0 for the BESOU dataset).

Data preparation
All datasets (raw data) were downloaded from GEO [15]. The samples were background corrected and normalized using GeneChip-RMA [16] and subsequently log2 transformed. Since we combined datasets which originated from different experiments and platforms we applied a normalization strategy to enable reliable combination of the different datasets. First we verified the individual probe quality by employing ProbeMapper [17]. We selected the probes that were mapped to the same transcript by the vendor, the bioconductor annotation packages and a BLAST mapping of the probes to the set of transcripts of the latest human genome build (Hg19). In addition, we discarded the probes that were mapped to multiple genes. The probes that passed these tests were median centered, i.e. we set the median expression of each probe per dataset to match the median expression of that probe in the MAQC2 dataset. The median centering was performed by taking into consideration the ratio of clinical subtypes and response rates in each dataset, i.e. the median expression of a gene within a dataset was determined on a subset that was chosen such that it matched the other datasets with respect to the percentages of clinical subtypes and response rates. We chose the largest possible subsets that satisfied these constraints. For the subsequent analysis we selected the most informative probe per gene, i.e. the probe which showed the highest standard deviation across all data sets.

Feature selection and classification
After probe matching and gene median centering, 12010 genes were entered into the subsequent analysis. In addition to expression based predictors we also investigated the predictive power of clinical features. For these predictors we selected ER status, PR status, HER2 status, TNM-stage, age and grade as features to be used the predictors. All of these are routinely determined in the clinic and were directly available in the datasets we selected.

We employed three feature selection approaches for the expression based predictors and three for the clinical predictors. We chose different feature selection approaches for the expression and the clinically based predictors, since the expression features are all continuous valued and most of the clinical features are categorical, implying different requirements for feature selection. The feature selection approaches employed by the expression predictors were: Wilcoxon-Mann-Whitney test (WMW), Wilcoxon-Mann-Whitney test after removal of features with a Pearson correlation exceeding 0.75 (WMW-uncor.) and the ratio of between- to within group sum of
squares (BWR). These methods are either available in R [18] or were implemented in R. The ranking approaches employed for the clinical features were: information gain (Inf.gain), correlation feature selection (CFS) and RELIEF. These feature selection methods are available in the "FSelector" R package [19]. The features were ranked by their respective scores when comparing responder samples to non-respond samples (i.e. their discriminative power) in order to select the best features for classification.

The classifiers that we employed were the same for the expression data and the clinical data. We tested the following classifiers: the J48 decision tree (J48), the 3-nearest neighbor classifier (3NN) based on Euclidean space, the nearest mean classifier (in Euclidean space), logistic regression (LREG), naive Bayes (NB) and a support vector machine (SVM) (WeKa SMO with default settings). This represents a wide range of classifiers including linear and non-linear classifiers as well as classifiers designed for discrete and continuous features. All classifiers were implemented in the “RWeka” R package [20, 21].

We implemented each possible combination of classifier and feature ranking approach, except for the J48 classifier since this classifier includes its own feature selection approach. This resulted in 16 clinical and 16 expression based predictors. We will refer to a specific combination of a set of selected features and the associated classifier as a predictor.

Estimation of predictive performance
To estimate the accuracy of the predictors we employed the area under the receiver operator curve (AUC). The AUC is estimated through a double loop cross-validation strategy, illustrated in Figure 1.

Full details of the procedure are outlined in the supplementary materials and methods, in summary the following steps are taken: 1) the input dataset is divided into 2/3 for the training set and 1/3 for the validation set; 2) the training set (i.e. 2/3 of all data) is employed to determine the optimal number of features to be used in each feature selection method and classifier combination (i.e. for each of the 16 possible combinations of feature selection method and classifier, a single, optimal number of features is determined); 3) the training set is then used to train each of the 16 predictors with the number of features previously determined to be optimal for each combination; 4) finally the performance of each feature selection and classifier combination is assessed by applying these 16 predictors to the validation set (i.e. 1/3 of all data). The split into training and validation sets was performed such that the ratios of subtypes and response rates within the two sets were equal. This whole procedure was repeated 15 times in order to get a more accurate average performance per predictor and subtype. This double loop cross-validation procedure is similar to the approach that was employed by Popovici et al.[13] and Wessels et al.[22].

For the non-subtype specific predictors, we divided all data of all subtypes into three equal parts, i.e. the dataset was subdivided into 2/3 of the samples
Figure 1. Cartoon of the double loop cross-validation scheme. Our analysis employed a double loop cross-validation. The inner loop determines the optimal number of features to be used by a specific combination of feature selection and classifier, here depicted by the green block. This inner loop uses 2/3 of all data (i.e. the training data), the remaining 1/3 is employed to measure the performance of the trained classifier (i.e. a 3 fold cross-validation setup). The outer loop is repeated 15 times in order to get an average AUC for each predictor.

in the training set and the remaining 1/3 in the validation set. For the subtype specific predictors the input dataset consisted only of samples of the specific subtype being analyzed (i.e. the 2/3 training set and 1/3 validation set consisted only of samples belonging to the relevant subtype). In order to compare the performance of the non-subtype specific predictor with the subtype specific predictor on a given subtype, say TN, the non-subtype specific predictor was trained on a training set consisting of all subtypes, but only validated on a the validation set consisting of TN samples.

Finally we selected the best performing predictor per subtype for both the subtype specific and the non-subtype specific predictors in order to compare their
performances. To compare the AUC values of non-subtype specific predictors to the AUC values of subtype specific predictors we employed the two-sided t-test.

RESULTS

Subtype specific versus non-subtype specific predictors
We employed a stratification based on ER and HER2 status. We stratified patients into a HER2 positive group (HER2+), a Luminal group (HER2-, ER+) and a Triple Negative group (HER2-, ER-)(TN). We further subdivided the HER2-positive group based on ER status, resulting in a HER2-positive and ER-negative group (HER2+, ER-) and a HER2-positive and ER-positive group (HER2+, ER+). Unfortunately, the HER2 positive and ER positive group had too few samples to analyze this group separately with adequate power (n = 37).

Figure 2 shows the comparison of the best performing classifiers for the subtype specific and non-subtype specific predictors.

Predictors based on clinical features
The AUCs for the clinical subtype specific and non-subtype specific predictors were highly similar for most sets (HER2+, TN and ER+)(Red boxplots in Figure 2). Only in the HER2+, ER- subgroup there was a significant difference, where the subtype specific predictor had a significantly higher AUC than the non-subtype specific (multiple testing corrected p-value: $4.30 \times 10^{-5}$). None of the other groups showed a significant difference in AUC between the subtype specific and non-subtype specific clinical predictors.

Predictors based on gene expression data
The AUC comparisons for the expression based predictors (blue boxplots in Figure 2) mostly mirrored the results for the clinical features based predictors. All groups showed highly similar results for the subtype specific and the non-subtype specific predictors. Only in the Luminal group, a borderline significant difference in AUC between the subtype specific and non-subtype specific predictor was observed (corrected p-value: 0.0709).

Table 2 shows the best performing feature selection method-classifier combination for each dataset (these correspond to predictors whose AUCs were compared in Figure 2). The exact average AUC scores and details about the different data sets can be found in Table S-2.

DISCUSSION

From our analysis we can conclude that there is very little difference in performance between building subtype specific predictors compared to building non-subtype specific predictors (i.e. combining data of all subtypes and training on that dataset). This is true for the predictors employing clinical features as well as the predictors
employing gene expression predictors. We did observe that in the luminal subgroup all the non-subtype specific models (gene expression or clinical features based) tended to achieve better AUC scores compared to the subtype specific predictors, although this difference was not significant. This could be explained by the fact that positive events (i.e. pathological complete responses) are rare in this set, only 7% (n=14) of the ER-positive patients achieved a pCR in our set. Having few samples in the minority class will make it difficult to train a predictor that correctly classifies samples from this minority class, leading to lower AUC values. The higher performance of the non-subtype specific predictor (compared to the
For the TN and HER2+ subtypes, there was no clear difference between the subtype specific and the non-subtype specific predictors. Only for the HER2-positive and ER-negative subgroup, we found a significant difference in clinical feature based predictor performance. The subtype specific predictor, based on clinical features, was superior to the non-subtype specific model. While the AUC of the subtype specific model was 0.59, the AUC of the non-subtype specific model was exactly 0.5. This could be explained by the fact that in most iterations of the predictor training, the non-subtype specific predictor included only a single feature. This single feature was ER status (in some cases combined with either PR or HER2). Taken over the whole dataset, ER status (or PR and HER2) is highly predictive of response to chemotherapy. However, the HER2+, ER-subtype is, by definition, completely ER negative and HER2 positive, so the ER and HER2+ status feature will not yield any predictive power in this subgroup. The subtype specific predictor did not include ER or HER2 status, but included age, stage and grade. Given that the subtype specific predictor included more informative features, it was no surprise that it could achieve higher AUC values. This scenario forms a clear example of when a subtype specific predictor is preferred, i.e. a feature is predictive over the whole dataset, but is non-informative within a specific subgroup of that dataset. When we look at the same subtype, but now employing gene expression data, the difference between the subtype specific and non-subtype

Table 2. Characteristics of the optimal predictors for the different subtypes. In each cell the optimal combination of classifier, and feature selection method, as well as the optimal number of features selected, is shown.

<table>
<thead>
<tr>
<th>Stratification</th>
<th>Clinical</th>
<th>Gene Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Subtype specific</td>
<td>Non specific</td>
</tr>
<tr>
<td>Luminal</td>
<td>LREG-Relief 2</td>
<td>NM-Relief 5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triple Negative</td>
<td>NM-CFS 2</td>
<td>NB-Relief 6</td>
</tr>
<tr>
<td>HER2-positive</td>
<td>NB-Relief 5</td>
<td>NB-CFS 3</td>
</tr>
<tr>
<td>HER2-positive, ER-negative</td>
<td>3NN-CFS 1</td>
<td>NM-CFS 1</td>
</tr>
</tbody>
</table>

Legend: classifiers: NB=Naive Bayes, NM=Nearest Mean, LREG: Logistic regression, SVM=Support vector machine, 3NN=3-Nearest Neighbor; Feature selection methods: CFS=Correlated feature selection, WMW=Wilcoxon-Mann-Whitney, BWR = Ratio between to within class sum of squares, WMW-uncor. = Wilcoxon-Mann-Whitney where correlated features are removed, Inf.gain = information gain.
specific predictors disappears however. The selected features in that classifier are predictive over both the entire set and this specific subgroup of samples.

In our analysis we did not take into account that there is a large difference in the size of the training set between subtype specific and non-subtype specific predictors. That is, we compared the subtype specific predictors, the predictors based only on data from one subtype, to a predictor that was trained on data from all subtypes. This means that the training set sizes of the non-subtype specific predictors were two to four times as large as the subtype specific training sets (depending on the size of the subtype relative to the whole dataset). These training set size differences have an influence on predictor performance and performance estimation. Since we focused on comparing subtype specific to non-subtype specific predictors in real world situations, we opted to compare performance in this manner. When a dataset is to be analyzed, one would always want to use as much data as available so it would not make sense to only take part of the combined dataset purely for comparison reasons. Our results suggest that the larger training set of the generic predictor outweighs the benefits of a subtype specific training set. This in turn implies the important notion that there are features in the data that can predict response to therapy that are shared amongst different subtypes. That is, if there was no commonality between predictive factors in the different subtypes, the increased training set size would not increase the performance of the generic predictor.

It has been shown that patients benefit from receiving these subtype specific treatments instead of a general treatment [23, 24]. When patients receive treatment that is specific for a particular subtype, a predictor that combines all data will not only feature heterogeneity in subtypes, but also in treatment regimens received. It remains to be seen if a non-subtype specific predictor will, in such a scenario, still perform as well as a predictor that was trained for a specific treatment and subtype.

In recent papers studying outcome prediction after T-FAC or T-FEC and FAC or AC neoadjuvant therapy, no predictors specific for breast cancer subtypes were built [4, 8, 9, 13]. Therefore, we cannot directly assess if the subtype specific predictors that we built perform better or worse than previously published predictors. However, we could compare the non-subtype specific predictors (i.e. built and tested on all subtypes at once). More specifically, we compared the performance of our predictors to the predictors published alongside the three Affymetrix datasets which we included in our study. Unfortunately, the paper that was published alongside the BESOU data did not include AUC values so we could not compare their performance to ours. Instead, the authors of this paper[9], focused on finding the best model to achieve a high negative predictive value instead of a high AUC. In the paper accompanying the MAQC2 dataset[13], the authors presented an expression based predictor that combined LREG with BWR. The gene expression based predictor published with the MDACC dataset[12], was a predictor based on DLDA (Naive Bayes (NB)), which included 30 genes, selected based on the t-test p-value. The average AUC of the Popovici et al. model was: 0.805, Tabchy et al. reported an AUC 0.711 for the T-FAC
treated samples. Our highest average AUC value was 0.768 (NB with BWR using 30 features). Given the variance we observed around our AUC estimate, these AUC values are most likely not statistically significantly different.

Tabchy et al. also validated a nomogram based on clinical data, which achieved an AUC of 0.89 for the T-FAC treatment arms. Our best clinical feature predictor was the predictor employing NB and relief using 6 features, which achieved an average AUC of 0.796. Again, the variance of the double loop CV estimate is of such a magnitude that these values are most likely not significantly different.

A possible explanation for the slightly lower AUC values for our predictors could lie in the fact that we classified more samples and -potentially- our AUC estimates are more accurate due to our larger sample set. We included 2.9 times as many samples compared to Tabchy et al. and 1.7 times as many samples compared to Popovici et al. In addition, we combined data from multiple sets and even though the patient characteristics were not significantly different, this could have an influence on the AUC. Differences in estimated AUC values could also be attributed to the training and validation procedures employed. To test this we applied our double-loop cross validation on the Tabchy and Popovici datasets. Our best performing gene expression based predictors showed AUCs of 0.807 and 0.780 respectively, which is very close to the reported performances. This indicates that the published performances are robust, but dataset specific. As we trained and tested on a larger dataset, and as the training and testing procedure has been proven to be robust, our achieved AUC of 0.768 is probably closer to the real value. However, as indicated above, these values are within the error margins of the experiments and since we compared the relative performance of specific to generic predictors, this performance difference has no consequences for the results presented here.

CONCLUSIONS

In general, subtype specific predictors would be preferred over non-subtype specific predictors. However, when the phenotype to be predicted has a skewed distribution, like response in the Luminal subtype, a non-subtype specific predictor (i.e. based on all data combined) will outperform the subtype specific predictor. This implies that there are features that are predictive of response to chemotherapy that are shared amongst different subtypes and resistance mechanisms are not exclusive to particular subtypes. In specific cases where features are predictive over the whole dataset, but uninformative within a specific subgroup, like ER and HER status in the HER2-positive and ER-negative subgroup, a subtype specific predictor will offer a significant performance gain over non-subtype specific predictors.

For breast cancer specifically - with the exception of the HER2-positive, ER-negative group-, we can conclude that building a subtype specific predictor offers equal performance compared to a predictor based on all available data. However, when it is
unknown which of the scenarios mentioned above is present in the data, it would be advised to analyze both subtype specific and non-subtype specific predictors.

REFERENCES


LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNM stage</td>
<td>Tumor, Nodes, Metastases</td>
</tr>
<tr>
<td>ER</td>
<td>Estrogen Receptor</td>
</tr>
<tr>
<td>PR</td>
<td>Progesterone Receptor</td>
</tr>
<tr>
<td>HER2</td>
<td>Human Epidermal growth factor Receptor 2</td>
</tr>
<tr>
<td>TN</td>
<td>Triple Negative</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under the Curve</td>
</tr>
<tr>
<td>MAQC2</td>
<td>MicroArray Quality Control 2</td>
</tr>
<tr>
<td>BESOU</td>
<td>Breast and Endocrine Surgery of the Osaka University</td>
</tr>
<tr>
<td>MDACC</td>
<td>MD Anderson Cancer Center</td>
</tr>
<tr>
<td>GEO ID</td>
<td>Gene Expression Omnibus Identifier</td>
</tr>
<tr>
<td>T-FAC</td>
<td>Taxol Fluorouracil Adriamycin Cyclophosphamide</td>
</tr>
<tr>
<td>T-FEC</td>
<td>Taxol Fluorouracil Epirubicin Cyclophosphamide</td>
</tr>
</tbody>
</table>
FAC  Fluorouracil Adriamycin Cyclophosphamide
AC   Adriamycin Cyclophosphamide
pCR  pathological complete response
FISH Fluorescent In Situ Hybridization
WMW  Wilcoxon-Mann-Whitney
WMW-uncor. Wilcoxon-Mann-Whitney with correlated features removed
BWR  Between to Within Ratio of sum of squares
Inf.gain Information Gain
CFS  Correlated Feature Selection
3NN  3 Nearest Neighbour
LREG Logistic Regression
NB   Naive Bayes
SVM  Support Vector Machine
DLDA Diagonal Linear Discriminant Analysis

AUTHORS’ CONTRIBUTIONS
JR participated in the design and coordination of the study, conceived of the study and drafted the manuscript, MB participated in the design and coordination of the study, conceived of the study, carried out the analyses and helped draft the manuscript, EL participated in the design and coordination of the study, SR participated in the design and coordination of the study, and LW participated in the design and coordination of the study, helped draft the manuscript, and conceived of the study. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS
JR and EL were funded by CTMM, the Center for Translational Molecular Medicine (www.ctmm.nl), project Breast CARE (Grant 03O-104).

COMPETING INTERESTS
The authors declare that they have no competing interests.
SUPPLEMENTAL MATERIALS
Supplemental methods

Estimation of predictive performance
To estimate the accuracy of the predictors we employed the area under the receiver operator curve (AUC). The AUC is estimated through a double loop cross-validation strategy.

Figure 1 illustrates the experimental design. In the inner-loop, marked in green (Figure 1), the optimal number of features is selected per classifier (C) and feature selection (FS) combination, i.e. per predictor. In the outer-loop, marked in blue (Figure 1), all training samples were used to train the predictor, employing the number of features per predictor that was determined in the previous steps. This classifier was then evaluated by measuring the predictive accuracy on the samples that were not used in either the training of the classifier or in selecting the optimal number of features. In both the inner loop and in the outer loop, a three-fold cross-validation (3FCV) was used. In the outer-loop the 3FCV is repeated 15 times and the 3FCV of the inner-loop was repeated 5 times per outer-loop iteration. We chose to use fewer iterations in the inner as compared to the outer loop as this had a large influence on the runtime and we reasoned that the confidence intervals of the inner loop did not require a high accuracy since we were only interested in the maximum average AUC to select the best number of features. The stratifications of the data into a training and an evaluation set were performed with respect to response rates.

We calculated the optimal number of features to be used for each predictor by averaging the best performing number of features selected in each inner loop iteration, per FS-C combination. To limit the number of tests to be performed, we fixed the number of features that could be used. For the clinical predictors this ranged from one to six features in steps of one and for the gene expression data the number of features ranged from 10 to 120 features in steps of 10.

To calculate the AUC, we evaluated the class posteriors returned by the predictors. The AUC calculation itself was performed using the “ROCR” R package[1]. By averaging the AUCs for the three folds repeated either 15 (outer) or 5 (inner) times, we could estimate the AUC performance of a new set including the corresponding confidence intervals.

The double loop cross-validation procedure is similar to the approach that was employed by Popovici et al.[2] and Wessels et al.[3].

REFERENCES
2. Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, Samuelson FW, Nikolsky Y,
BREAST CANCER SUBTYPE SPECIFIC CLASSIFIERS


Figure 1. Illustration of the double loop cross-validation scheme. Legend: 3FCV: 3 fold cross-validation, CFS: Correlated feature selection method, Inf.gain= information gain, WMW=Wilcoxon-Mann-Whitney test, WMW-uncor. = Wilcoxon-Mann-Whitney with correlated features removed, BWR = between to within group sum of squares. NMC = nearest mean classifier, LREG= logistic regression, NB= naïve Bayes, 3NN = k-nearest neighbor, SVM= support vector machine.