Using Multi-Task Learning to Predict New Tumor Events in Breast and Prostate Cancer

Anand Ahuja and Karl Tayeb
EN.600.438 Computational Genomics

1 Introduction

Whole genome sequencing techniques such as RNAseq can elucidate expression patterns that characterize biological pathways. Our project seeks to predict new tumor events in cancer patients using genetic profiles at first contact. We found that numerous other studies analyzed machine learning methods for cancer classification and traditional prediction questions for different types of cancer. Ultimately, we decided to develop a model to predict new tumor events, which has more confounding factors than determining the presence or absence of cancer; to answer a more complicated question we thought we could improve on traditional prediction methods by using multitask prediction. The general intuition behind this decision is that new tumor events occur in every form of cancer, so in theory there also should exist common genetic mechanisms associated with new tumor events in the different types of cancer. Thus our question is significant because we seek to apply multitask learning techniques to predict new tumor events; this not only would hopefully create a more accurate model but also explore the limitations of multitask learning. Multitask learning is an approach to transfer learning, so our question also explores the relatedness of new tumor events in breast and prostate cancer. The concept of using two related prediction tasks in parallel to learn from each other and create more accurate models motivated our question – this type of machine learning seemed directly applicable to creating the most biologically significant answers by utilizing more relevant information to account for limitations in the data.

2 Related Works

With the improvements in DNA sequencing and falling costs, computational genomics research has sought to use this explosion in data to understand the nuanced genetic signatures in cancer for prediction, prevention, and care. Applications of machine learning to cancer prediction are certainly not novel, but many limitations and biases develop with analyses using traditional prediction models that implement single feature sets and specific data as it does in joint predictions. A now relatively old review of machine learning in cancer prediction from 2007 highlights that many of the early limitations in developing models include data dimensionality, vague multiple predictor models, and lack of reproducibility. Part of this stems from early computational genomics' lack of power to make specific predictions and only apply to determine "outcome for breast cancer patients at this particular hospital…they may become irrelevant over time."[1] Furthermore a 2014 review of machine learning applications in cancer prognosis and prediction highlighted the sheer volume of computational genomics projects geared towards developing a better understanding of cancer. One of the most common limitations noted in these studies was dataset size and quality; similarly the author notes that many learning algorithms could have been able to extract more accurate and reliable predictions using a better developed feature set. Aside from model validation, the review notes that now with the rise of high throughput technologies, computational models frequently use multiple data types to make more generalized predictions.[2] Our project sought to exploit the common signatures of new tumor events between different types of cancer to create a better model of prediction of tumor recurrence. Prediction of tumor recurrence in breast cancer patients has been explored several times – one of the earliest studies developed an artificial neural network using 2441 breast cancer patients to predict relapse over 5 years.[3] The study was heralded for its breadth of data and attempts to create general models using data from multiple institutions; furthermore

[1] Cruz, J. A., & Wishart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics*, *2*, 59–77.

[2] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction.*Computational and Structural Biotechnology Journal*, *13*, 8–17.

[3] Laurentiis, Michelino et al. "A Prognostic Model That Makes Quantitative Estimates Of Probability Of Relapse For Breast Cancer Patients". *Clinical Cancer Research* 5.12 (1999): 4133-4139. Web. 8 May 2016.

their machine learning approach yielded higher prediction accuracy than traditional staging methods of cancer recurrence prediction.

Our aim was to create a model that would exceed such prediction capabilities by implementing multi-task learning to generalize our model, incorporating information from two types of cancer. Learning tasks in parallel using a shared feature set should create a more generalized and hopefully informative model because what is learned during one task may help learn other tasks.[4] This type of machine learning has been applied to cancer, but only thus far in cancer classification analyses. Our project was partially inspired by the successful application of transfer learning to cancer prediction in both leukemia and prostate cancer using neural networking with multiple back propagation.[5] While this study demonstrated the successful application of multitask learning within a given cancer, we sought to apply multitask learning to improve prediction across cancer types. Vladimir Cherkassky at the University of Minnesota released an open source package for his multitask support vector machine SVM+MTL.[6] We planned to use this package to take a multitask learning approach to predicting new tumor event occurrence in breast cancer and prostate cancer.

3 Data

We analyzed UNC IlluminaHiSeq RNASeq Data Level 3 Data from The Cancer Genome Atlas. In particular, we used the gene expression data from Breast Invasive Carcinoma (BRCA) and Prostate Adenocarcinoma (PRAD) patients. For our analyses we focused on data from tumor, matched normal individuals – this notation indicates that the data is from tumor but for which matched normal tissue exists. The BRCA data set had 514 tumor matched individuals and ~10% had new tumor events. The PRAD data set had 398 tumor matched patient samples and ~20% had new tumor events. After ensuring patients had relevant follow up tumor information, and omitting patients without follow-up, we had 359 BRCA samples and 278 PRAD samples. For each sample expression signals of particular composite exons of 20531 genes were obtained from the RNAseq analysis.

4 Methods

First, we split our data 60/40 for a training set and a test set.

A histogram of our data indicated that we needed to process our data prior to analyses because as indicated in the left panel of Figure 1, our data was very heavily skewed right. We implemented the Shapiro-Wilk normality test to evaluate how well each gene fits a normal distribution after log and various power transformations. The log transformation yielded the highest mean and median results of likelihood that data was from a normal distribution. After normalization, we sought to scale our data to have mean of 0 and variance of 1.

Our approach to feature selection was borrowed from Chen and Huang in their paper on applying multitask learning to identify cancer.[7] Gene by gene we use the signal-to-noise statistic which is simply the difference in means between the two groups (new-tumor and non-new-tumor) divided by the sum of the variances in each group. On data that follows a normal distribution this is a good measure of which genes have significantly different expression differences between the two conditions. We rank genes by signal-to-noise statistic to

[4] Caruna, Rich. "Multitask Learning". *Machine Learning* 28 (1997): 41-75. Print.

[5] Huang, Zone-Wei, and Austin Chen. "A New Multi-Task Learning Technique to Predict Classification of Leukemia and Prostate Cancer." Ed. David D. Zhang. *Medical Biometrics*. Berlin: Springer, 2010. 11-20. Print.

[6] Cherkassky, Vladamir. Generalized SMO for SVM+MTL. Computer software. Predictive Learning and Advanced Learning Technologies. University of Minnesota, 2008. Web.

[7] Chen, A. H., & Huang, Z. W. (2010). *A new multi-task learning technique to predict classification of leukemia and prostate cancer*. In *Medical Biometrics*(pp. 11-20). Springer Berlin Heidelberg.

select the genes for our models. Multitask learning requires that the two tasks being learned have the same feature space. To achieve this, we thought of ways of selecting genes significant to both prediction tasks. The two methods we attempt are selecting the top k genes for each cancer to form a joint feature-space of 2k genes. The other is to iteratively take the top ranked genes from BRCA, PRAD, and the overlap between both cancers. This overlap was selected by ranking the minimum signal to noise statistic of each gene between each cancer.

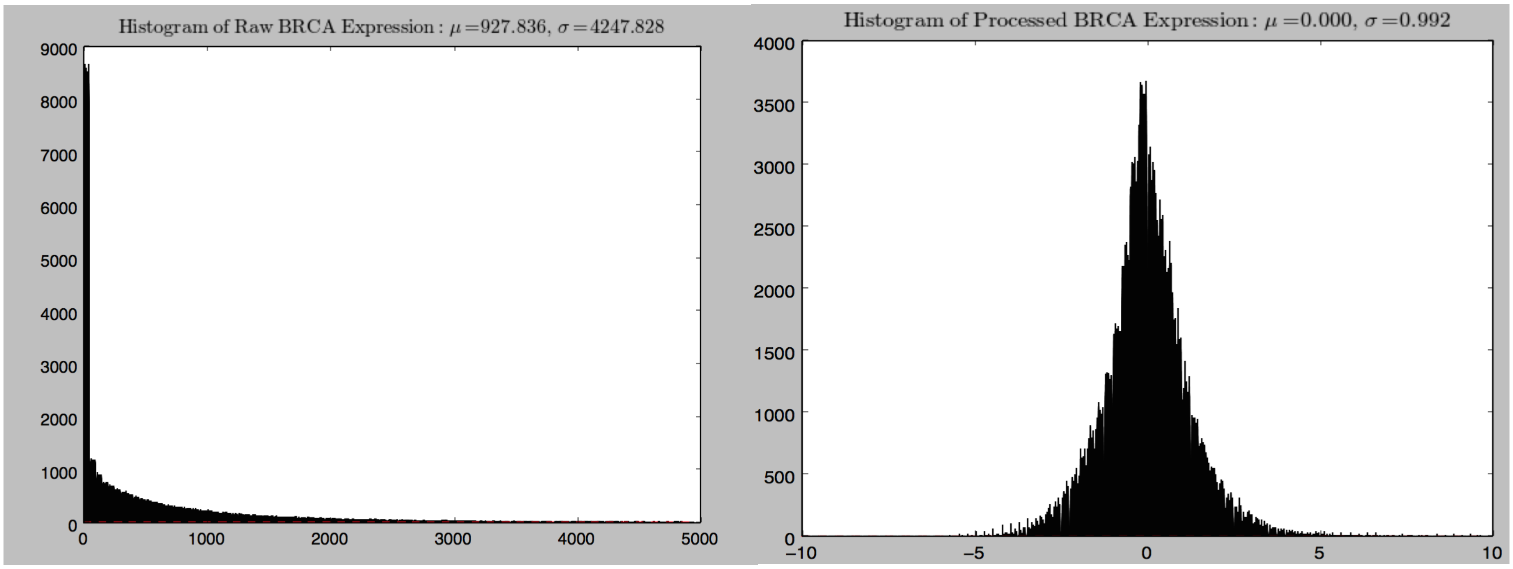

Figure 1. The left panel is a distribution of the data prior to normalization and the right panel illustrates the data after pre-processing appears to fit normal distribution. BRCA representative of PRAD data preprocessing. Generated by test-dist.py in /src

Our baseline models consisted of a number of linear kernel support vector machines. We tested a number of feature selection approaches (BRCA-only, PRAD-only, topmixed, and mixed) to select a varying number of genes (50, 100, 200, 300, 500). For each model we tested a range of the tuning parameter C (0.01, 0.1, 1, 10, 100), selecting the parameter that performed best on average on the validation set during 5-fold cross validation. Note, features were selected for each training subset during cross-validation. In our initial presentation of preliminary results we found that selecting features on the entire training data, including the validation set, caused the performance on the validation set to be drastically better than that on the test set which is held out completely.

After preliminary results yielded low prediction accuracies, we explored confounding errors resulting from batch effects. We used an MD Anderson Cancer Center web statistical package to analyze batch effects from TGCA datasets. The TGCA Batch Effect's tool quantitatively measures batch effects using the Dispersion Separability Statistic represents the ratio of intra-sample homogeneity to the homogeneity of the inter-batch homogeneity. Based on these metrics the BRCA and PRAD data from TCGA was not found the have batch effects.[8]

[8] *TGCA Batch Effects Tool*. Computer software. *Quantitative Measures of Batch Effects*. Vers. 2. MD Anderson - Department of Bioinformatics and Computational Biology, n.d. Web.

To build our multitask model we use a multitask SVM package for MATLAB by University of Minnesota's Vladimir Cherkassky. The idea behind this model is to learn an 'average' predictive model referred to as a decision space for both predictive tasks, as well as corrective models that add back specificity for each predictive task.

To gauge the performance of our models we used ROC curves which is created by plotting the true positive rate against the false positive rate at varying classification thresholds. The area under the curve (AUC) is typically used for comparison of different machine learning models. To understand the multitask model's effectiveness we compared the AUC from the joint model with 2K features to the independent learning models with K features and 2K features using both feature selection methods. Additionally, we also recorded model accuracy, precision, and recall. In particular, we are interested in learning a model with high recall scores as this indicates the models ability to positively classify actual new tumor events. We are less concerned with regular model accuracy – since there are comparatively few examples of new tumor events in the dataset any model that prefers to classify samples as non-new-tumor events will have somewhat high accuracies. That said, we chose to compare our models first and foremost on ROC-AUC because it considers with what confidence the model positively or negatively classifies samples.

5 Results

Baseline results:

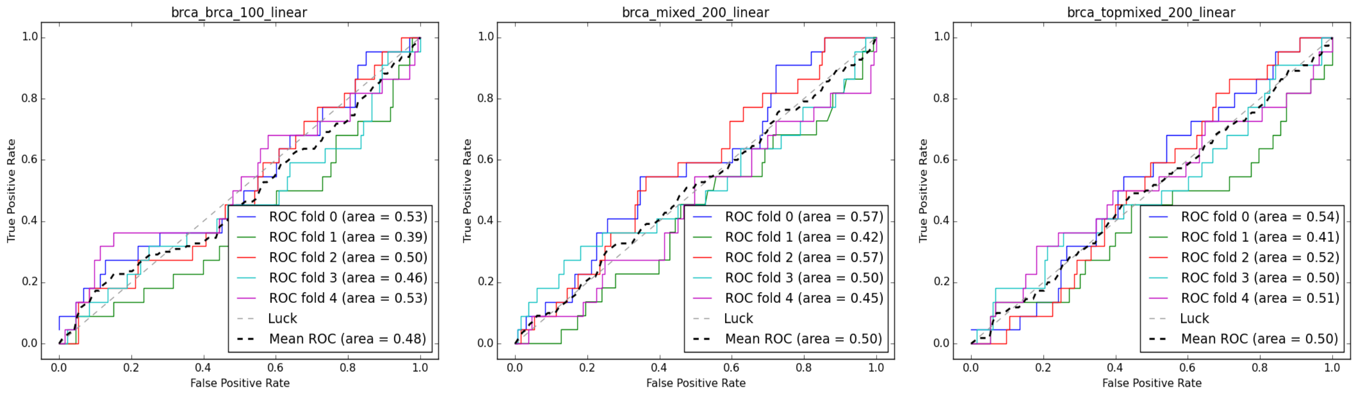

Figure 2. ROC on BRCA Test Set Data The baseline models in BRCA patients predicted new tumor events at no worse than randomly predicting outcomes. Title indicates <predictive task>_<feature selection>_<# features>_<SVM kernel> BRCAThe mean AUC was .48 in the independent model, .50 in the mixed model, and .50 in the topmixed model.

Model	brca_brca_100_linear	brca_topmixed_200_linear	brca_mixed_200_linear
Val Acc	0.941666667	0.821713615	0.810719875
Val Prec	0.755479798	0.254996799	0.296748521
Val Rec	0.662222222	0.241666667	0.306666667
Val ROC-AUC	0.111203937	0.634947257	0.653209218
Test Acc	0.735483871	0.763870968	0.762580645
Test Prec	0.153448935	0.222578692	0.277656046
Test Rec	0.090909091	0.163636364	0.236363636
Test ROC-AUC	0.525632262	0.550786056	0.568967874

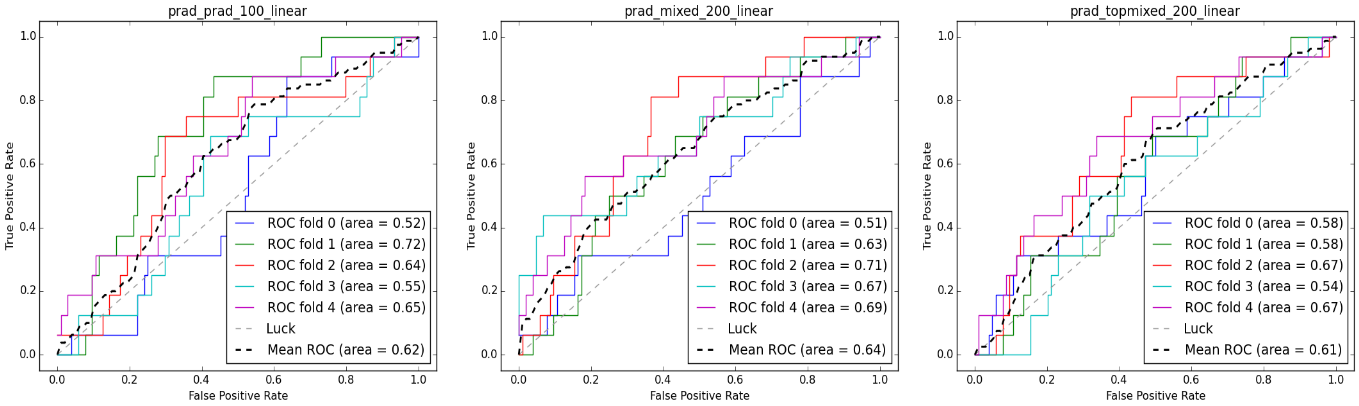

Figure 3. ROC on PRAD Test Set Data The baseline models in PRAD patients had a higher prediction accuracy than the BRCA models. Title indicates <predictive task>_<feature selection>_<# features>_<SVM kernel> The mean AUC was .62 in the independent model, .64 in the mixed model, and .61 in the topmixed model.

Model	prad_prad_100_linear	prad_topmixed_200_linear	prad_mixed_200_linear
Val Accuracy	0.65461039	0.741168831	0.72974026
Val Precession	0.354324426	0.445052864	0.436550459
Val Recall	0.312837163	0.384532135	0.388844489
Val ROC-AUC	0.540601326	0.655091201	0.6537381
Test Accuracy	0.661666667	0.673333333	0.698333333
Test Precession	0.266124576	0.238172896	0.280141863
Test Recall	0.3	0.25	0.3
Test ROC-AUC	0.496754808	0.509855769	0.521875

For each model we performed 5-fold cross validation on the training set, performing feature selection on the training portion of each fold, then evaluating accuracy, precision, recall, and ROC-AUC on the held out validation set and the training set. For each model we checked a range of tuning parameters C (0.01, 0.1, 1, 10, 100) for the linear SVM, selecting C that returned the highest average ROC-AUC. As pictured above we are unable to produce a meaningful predictive model for BRCA data across feature selection choices. For PRAD we see marginally better results, but ultimately models built on both cancers individually, regardless of the range of feature selection choices we make, fail to accurately predict positive new tumor events. We attributed the shortcoming of this model to the relatively small sample of positive new tumor events in the BRCA and PRAD datasets individually. It was our hope, that through multitask learning we would be able to build a better predictive model by identifying common underlying genetic factors that influenced new tumor events in both cancers.

Multitask Model Results:

For the multitask model we chose to use the top-mixed feature selection method described above to identify 200 genes to build a predictive model for both BRCA and PRAD. We trained a model with a linear kernel decision space and linear kernel corrective space. For this model there are two tuning parameters C, as in typical single-task SVM, and gamma which is a regularization parameter for the margin of the corrective spaces (a large gamma enforces that the corrective functions are small). We tested the model across a range of C (0.01, 0.1, 1, 10, 100) and gamma (10, 1, 0.1, 0.01, 0.001) values.

To test the models, we did 5-fold cross validation on each combination of parameters, selecting features on the training set alone, and recording accuracy, precision, recall, and ROC-AUC on the held out validation set and training data. Unfortunately, we found that across the board the multitask predictive model failed to perform better than the models learned for each cancer separately. Here, the ROC curves are telling, in that the AUC of ROC is always around (or even less than 0.5) indicating that the model is no more likely to assign a higher probability of new tumor events to actual new tumor event samples in the test set versus non-new tumor event samples in the test set.

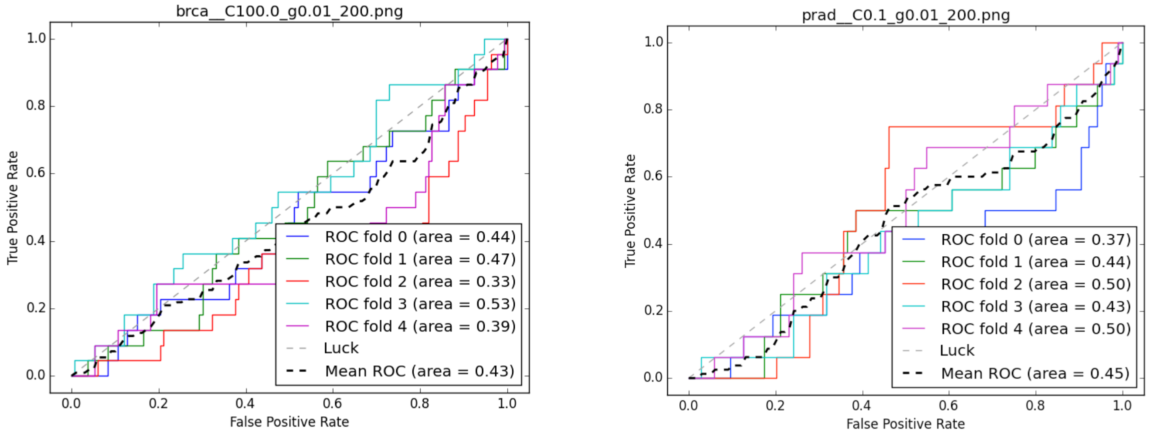

Figure 4. Typical results of the multitask SVM. The multitask SVM with linear decision and corrective space kernel fails to produce meaningful predictions of new tumor events.

6 Conclusion and Discussion

The results of the multitask model we trained make it obvious that for the set-up we described, linear multitask SVM is not effective for modelling new tumor events. We consider some of the reason why this is the case.

One of the big challenges in multitask learning is identifying and quantifying task similarity. We hypothesized that there would be common genetic signatures between BRCA and PRAD that are predictive of new tumor events in both cancers, in which case a multitask learning approach could be used to share information between the cancer types and build a model with better generalization, but it is also possible that the similarity between these two tasks is not that great. We try to account for this when testing the model by testing the model at varying C and gamma values in the model, however, we found that across all parameter settings the multitask model failed to perform better than average.

The other issue we consider is that there may be too many clinical confounding factors, presenting a major challenge to prediction of new tumor events from RNA seq data in general. For example, our model does not consider what treatment methods were pursued for each cancer, which we could reasonably expect to have a significant impact on new tumor event occurrence.

Common Features Selected in Each Fold of Cross validation of Multitask Model

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	200	4	1	0	1
Fold 2	4	200	3	5	2
Fold 3	1	3	200	2	3
Fold 4	0	5	2	200	0
Fold 5	1	2	3	0	200

Figure 5. Common Features Selected in Folds of Multitask Model. Clearly, the features selected by signal to noise ranking are highly sensitive to the subset of samples on which features are selected.

Another concern we have is that there seems to be large variability in which features are selected by our feature selection procedure. Part of the motivation of taking a multitask learning approach is to exploit the similarities in genetic signatures that are most predictive of new tumor events in both cancer types. However, the features selected by our feature ranking procedure are sensitive to what subset of the training data they are selected on. As figure 6 shows, there is very little overlap between the 200 features selected on the training data of each fold of cross validation in the multitask model we trained. This suggest that noise in the data is dominating the feature selection process. Larger datasets, particularly having more examples of new tumor events would help resolve this problem.

7 Submission Notes

Please note that while we include the data and code to train the models this will take some amount of time to learn. For the multitask model it took over an hour and a half to train the model with 5-fold cross validation at all the parameters. To help expedite the process we have included in our submission saved pickle files and .mat files in the appropriate subdirectories so that you will be able to verify the model results with test_model.sh without necessarily training with train_model.sh. That said, train_model.sh will build the model given sufficient time. To minimize the submission size we did not submit the data from TCGA as in all it was several gigs of data. Instead we loaded the data and relevant clinical information into numpy array which we saved as pickles in the data subdirectory. process_data.sh essentially takes these arrays, performs the described transformations to the data and saves them to csv's.

The problem for this project was identified jointly by Karl and Anand. Karl found the multitask learning package and developed all computationally challenging aspects of the project. He was also responsible for the modular and script based development of the project. Anand processed the data and helped with portions of code throughout and focused on the analyses and writing portions of the project.

7 Works Referenced

Zhang, D., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage, 59(2), 895-907.

Kandaswamy, C., Silva, L. M., Alexandre, L. A., & Santos, J. M. (2016). High-Content Analysis of Breast Cancer Using Single-Cell Deep Transfer Learning. Journal of biomolecular screening, 1087057115623451.

Gönen, M., & Margolin, A. A. (2014, July). Kernelized Bayesian Transfer Learning. In AAAI (pp. 1831-1839).

De Laurentiis M, De Placido S, Bianco AR, Clark GM, Ravdin PM. (n.d.). A Prognostic Model That Makes Quantitative Estimates of Probability of Relapse for Breast Cancer Patients. Clinical Cancer Research, 1999 Dec;5(12):4133-9.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., ... & Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. Proceedings of the National Academy of Sciences, 98(20), 11462-11467.