2008

Identifying genomic signatures for predicting breast cancer outcomes

Shruti Rathnagiriswaran
West Virginia University

IDENTIFYING GENOMIC SIGNATURES FOR PREDICTING BREAST CANCER OUTCOMES

Shruti Rathnagiriswaran

Thesis Submitted to the
College of Engineering and Mineral Resources
At West Virginia University
In partial fulfillment of the requirements
For the degree of

Master of Science
In

Electrical Engineering

Committee:

Nancy Lan Guo, Ph.D., Chair
Bojan Cukic, Ph.D., Chair
Tim Menzies, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, WV

2008

Keywords: Machine Learning, Feature Selection, Breast Cancer prognostic prediction, Nearest Centroid Classification, Cox proportional hazard function, Kaplan-Meier Analysis

ABSTRACT

Identifying Genomic Signatures for predicting Breast Cancer outcomes

Shruti Rathnagiriswaran

Predicting the risk for recurrence in breast cancer patients is a critical task in clinics. Recent developments in DNA microarrays have fostered tremendous advances in molecular diagnosis and prognosis of breast cancer.

The first part of our study was based on a novel approach of considering the level of genomic instability as one of the most powerful predictors of clinical outcome. A systematic technique was presented to explore whether there is a linkage between the degree of genomic instability, gene expression patterns, and clinical outcomes by considering the following hypotheses; first, the degree of genomic instability is reflected by an aneuploidy-specific gene signature; second, this signature is robust and allows breast cancer prediction of clinical outcomes. The first hypothesis was tested by gene expression profiling of 48 breast tumors with varying degrees of genomic instability. A supervised machine learning approach of employing a combination of feature selection algorithms was used to identify a 12-gene genomic instability signature from a set of 7657 genes. The second hypothesis was tested by performing patient stratification on published breast cancer datasets using the genomic instability signature. The results concluded that patients with genomically stable breast carcinomas had considerably longer disease-free survival times compared to those with genomically unstable tumors. The gene signature generated significant patient stratification with distinct relapse-free and overall survival (log-rank tests; $p < 0.05$; $n = 469$). It was independent of clinical-pathological parameters and provided additional prognostic information within sub-groups defined by each of them.

The importance of selecting patients at high risk for recurrence for more aggressive therapy was realized in the second part of the study, considering the fact that breast cancer patients with advanced stages receive chemotherapy, but only half of them benefit from it. The FDA recently approved the first gene test for cancer; MammaPrint, for node-negative primary breast cancer. Oncotype DX is a commercially available gene test for tamoxifen-treated, node-negative, and estrogen receptor-positive breast cancer. These signatures are specific for early stage breast cancers. A population-based approach to the molecular prognosis of breast cancer is needed for more rational therapy for breast cancer patients. A 28-gene expression signature was identified in our previous study using a population-based approach. Using this signature, a patient-stratification scheme was developed by employing the nearest centroid classification algorithm. It generated a significant stratification with distinct relapse-free survival (log-rank tests; $p < 0.05$; $n =1337$) and overall survival (log-rank tests; $p < 0.05$; $n = 806$), based on the transcriptional profiles that were produced on a diverse range of microarray platforms. This molecular classification scheme could enable physicians to make treatment decisions based on specific characteristics of patients and their tumor, rather than population statistics. It could further refine subgroups defined by traditional clinical-pathological parameters into prognostic risk groups. It was unclear, whether a common gene set could predict a poor outcome in breast and ovarian cancer, the most common malignancies in women. The 28-gene signature generated significant prognostic categorization in ovarian cancers (log-rank tests; $p < 0.0001$; $n = 124$), thus, confirming the clinical applicability of the gene signature to predict breast and ovarian cancer recurrence.

Acknowledgements

I would like to express my gratitude to my advisor, Dr. Nancy Lan Guo for giving me an opportunity to work on this project for two years. Dr. Guo has been my mentor, and provided moral and financial support during my course work at West Virginia University. I thank Dr.Guo for being very patient and supportive in guiding me throughout this project. I owe my success and recognition in this project to Dr.Guo. Apart from the academic knowledge that I acquired in this lab, I also tried to imbibe some positive traits from Dr.Guo. This project is supported by the NIH/NCRR P20 RR16440-0.

I convey my sincere thanks to my committee members, Dr. Bojan Cukic and Dr. Tim Menzies for their guidance and honorable presence in my committee. I thank Dr. Cukic for his acceptance to co-chair my committee and providing me financial support in my final semester.

I would like to give a token of appreciation to all my lab-mates; Rama Kanth Mettu, Ying-Wooi Wan, Kursad Tosun, and Swetha Bose Nutakki for their timely help and friendly attitude. It was great fun working with them. I thank Dr. Yan Ma for her initial coordination and support in my project. I thank my friends for their encouragement

I thank Dr. Thomas Ried for confiding on us and providing us the data.

Last but not the least; I thank my parents, my grandparents, my aunt, my brother, Sharan, and all my relatives for their dedicated prayers and moral support. I am indebted to all of them. I thank the Almighty for showering his grace upon me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Breast cancer, which originates from the breast epithelial cells, is the most commonly diagnosed cancer in women, with an incidence rate more than twice that of colorectal cancer and cervical cancer and about three times that of lung cancer. It is the second leading cause of deaths related to cancer in women in the United States (1). The National Cancer Institute estimates 182,460 new cases and 40,480 deaths related to breast cancer among women in the United States in 2008[1]. Breast cancer is usually treated by removing the tumor and the involved lymph nodes. There are different kinds of therapies that frequently follow surgery namely; radiation therapy for women who have large tumors or many involved lymph nodes or in case of breast conservation, endocrine therapy for women with tumors that express the estrogen receptor (ER+), chemotherapy for women who develop a high risk for a poor outcome due to cases such as large tumors, involved lymph nodes, advanced disease or inflammatory breast cancer (1).

Predicting the recurrence of breast cancer is one of the most quintessential tasks for physicians. It helps them to determine the most appropriate care based on the individual patient's risk and treatment preferences, especially to avoid harsh therapies that may not be effective[2]**.** There are different predictive tests that evaluate the risk of recurrence and help the doctors to make more personalized decisions related to patient management. The traditional methods

[1] http://www.cancer.gov/cancertopics/types/breast

[2] http://www.earthtimes.org/articles/show/mammaprintr-breast-cancer-test-provides-valuable-insight-for-personalized-treatment decisions,414055.shtml

include decisions made on the basis of the location and size of the tumor, the grade and the type of the tumor. Numerous novel strategies have been introduced in the last two decades in the diagnosis and treatment of breast cancer. An innovative technology namely, genomic-based microarray profiling has fostered tremendous breakthroughs in diagnosis and prognosis of breast cancer. Using this technology, molecular signatures are identified and serve as a diagnostic tool for treating breast cancer.

Genomic instability, as reflected in heterogeneous nuclear DNA content, serves as one of the most powerful predictors of clinical outcome. It is defined as the loss of stability due to abnormal genetic changes occurring serially in cell-populations at a high rate, as they descend from the same ancestral cell[3]. The importance of genomic instability in the prognosis of breast cancer can be reflected by the fact that it causes metastases in breast cancer. Studies have shown that breast cancer patients with genomically stable tumors have considerably longer disease-free survival times compared to those with genomically unstable tumors (2;3). In predicting breast cancer outcomes, the observations resulting from nuclear DNA content and genomic instability were found to be similar to those resulting from gene expression signatures (4-7).

Considering these significances of genomic instability, a linkage was established between the degree of genomic instability, gene expression patterns, and clinical outcomes in the first part of our study. Gene expression profiling of 48 breast cancer patients with varying degrees of genomic instability was used and the differences between the groups of genomic instability were explored. This led to the development of a 12-gene genomic instability signature. A combination of different machine learning algorithms was employed to identify the 12-gene genomic

[3] http://www.ratical.org/radiation/CNR/GenomicInst.html

instability signature from an initial set of 7657 genes. The extent to which the genomic instability-associated gene expression patterns could allow the prediction of breast cancer outcomes was determined using a nearest centroid classification algorithm on 469 patient samples. The association between the genomic instability-defined risk groups and traditional prognostic factors of breast cancer was studied.

Two gene expression-based tests have been identified in the recent past to predict the outcomes of breast cancer, namely, Oncotype DX, and Mammaprint. Oncotype DX (8) of Genomic Health (Redwood City, CA) is the first multi-gene expression test that is available commercially. It has the ability to predict the consequences of chemotherapy and recurrence in early-stage breast cancer. It has been recommended for patients with lymph node-negative, estrogen receptor-positive breast cancer treated with tamoxifen (1), by both the American Society of Clinical Oncology (ASCO) and the National Comprehensive Cancer Network[4]. MammaPrint (9;10) is an another prognostic test for breast cancer recurrence that consists of a set of 70 genes. It was commercialized by Agendia (Amsterdam, The Netherlands) and was recently approved by the Food and Drug Administration (FDA). It is recommended for use in node-negative women under age 61 and with a tumor size less than 5 cm. However, these tests have been identified on particular subgroups of patients. A population-based prognostic gene signature is needed for deciding the kind of breast cancer treatment.

In the second part of our study, a challenging task was sought to accurately classify breast cancer patients into subgroups of good prognosis and poor prognosis, in an attempt to improve the breast cancer survival rate. The driving force behind this study is the fact that most of the

[4] http://www.biospace.com/news_print.aspx?NewsEntityId=98423

breast cancer patients receive chemotherapy but only half of them benefit from it (11). It is critical to select patients at high risk for recurrence for additional chemotherapy. Moreover, this part of the study also sought to investigate whether a common gene set could predict poor outcomes in both breast and ovarian cancer, considering the fact that women with breast cancer are easily susceptible to ovarian cancer according to epidemiological studies (12).

This part of the study sought to validate the predictive power of a 28-gene expression signature that was identified in our previous study (13) using a population-based approach to the molecular prognosis of breast cancer. The training dataset contained 99 patients having different histologies, while the validation datasets contained a total of 1734 breast cancer patients with heterogeneous disease stages, having gene expressions generated from different microarray platforms, and 124 ovarian cancer patients with advanced stage (III or IV). A scheme was developed for applying the prognostic gene signature in patient stratification, based on transcriptional profiles generated on a diverse range of microarray platforms, using a nearest centroid classification (NCC) algorithm. The association between the gene signature and traditional clinical-pathological factors was accessed in quantifying breast cancer disease-free survival and overall survival.

The thesis is organized as follows. Chapter 2 discusses the background of our study, Chapter 3 elucidates the development and validation of the 12-gene genomic instability signature, Chapter 4 describes the validation of the 28-gene expression signature that was identified earlier, and finally, Chapter 5 provides conclusion to the research work.

Chapter 2

Background

2.1 Introduction

The three primary concerns in cancer prediction/prognosis are as follows: 1) the prediction of cancer susceptibility where, one tries to predict the likelihood of developing a type of cancer prior to the occurrence of the disease, 2) the prediction of cancer recurrence where, one tries to predict the likelihood of redeveloping cancer after the apparent resolution of the disease, 3) the prediction of cancer survivability where, one tries to predict an outcome after the diagnosis of the disease (14). In the latter two cases, the success of the disease prognosis is dependent on the quality of diagnosis.

Machine Learning is a part of artificial intelligence that uses statistical, probabilistic and optimization tools to study existing examples and then uses the "prior" information to classify new data or identify new patterns (14). Machine learning methods have been used extensively in the past as an aid for cancer diagnosis, but recently, researchers have started applying machine learning techniques for cancer prediction and prognosis too (14).

The remainder of the chapter is organized as follows. Section 2.2 describes gene expression profiling, Section 2.3 discusses the need for feature selection, Section 2.4 reviews the machine learning algorithms that were used in our research, Section 2.5 and Section 2.6 explain the statistical methods that were adopted in our research, Section 2.7 reviews the related work done in previous studies, and finally, Section 2.8 provides a summary of the chapter.

2.2 Gene Expression Profiling

Gene Expression Profiling or Microarray analysis is an emerging technology for identifying genes. It has been successfully used in the prognosis and therapy of breast cancer and other diseases in the recent years. This has been accomplished by a variety of microarray platforms. In this analysis, the composition of cellular messenger ribonucleic acid (mRNA) is identified. It provides the measure of the number of mRNA transcripts derived from a gene (1). This technology involves several thousands of genes and thus causes the dimensionality to be very high. It becomes a major limitation in many pattern recognition problems when the sample size is small. Moreover, the large number of features leads to the degradation in the performance of the classifiers, if the number of samples is relatively very small. This causes the problem of 'Curse of dimensionality' which is a term coined by Richard Bellman that describes the problem caused by the exponential increase in volume as a result of adding extra dimensions to a space[5]. In the case of cancer classification, the number of genes is as large as thousands of genes but the number of samples is relatively small because of the limitations in the availability of samples, acquisition, time and cost (15).

2.3 Need for Feature Selection

Large number of features in a high dimensional dataset causes noise and introduces an error. Moreover, not all the features are important for performing an analysis on the dataset. This can be explained mathematically by considering a p-dimensional random variable *X* such that

$$X = (x_1, x_2, x_3.., x_p)^T \qquad (2.1)$$

[5] http://en.wikipedia.org/wiki/Curse_of_dimensionality

S is a *k*-dimensional subset of *X* having lower dimensions and is represented by

$$S = (s_1, s_2, s_3 \ldots s_k)^T \tag{2.2}$$

where $k \leq p$ and *S* contains the important features extracted by using some algorithm (criterion) (16). Another drawback with high dimensional datasets is that they may require more samples or observations to extract the important features.

The problems associated with the high dimensional datasets can be minimized by having a priori information about the features. The process of extracting important and relevant features is called Feature Extraction. This is usually done by employing machine learning algorithms. Basically, in this process, a subset of input variables is selected by eliminating features with little or no predictive information.

Feature selection can significantly improve the comprehensibility of the resulting classifier models and often builds a model that generalizes better to unseen points. Feature selection is thus defined as a process in which a data space is transformed into a feature space that, in theory, has exactly the same dimension as the original data space. However practically, the transformation involves a reduction in the number of the effective features but retains most of the intrinsic information of the data. Thus, this technique aims to minimize information loss while maximizing reduction in dimensionality.

2.4 Classification Algorithms

This section talks about the various classification algorithms that we used in our research to obtain the gene signatures.

2.4.1 Naïve Bayes Algorithm

The Naive Bayes algorithm is a machine learning algorithm used for classification. It is based on Bayes rule that makes two assumptions; 1) the attributes *X1…. X_n* are all conditionally independent of one another, given the class *Y* , 2) the predictive process is not influenced by any hidden or latent attributes (17;18). It is usually employed in supervised induction tasks, in which the ultimate goal is to accurately predict the class of test instances when the training instances include class information (18).

From the definition of conditional independence, given random variables *X, Y, Z; X* is conditionally independent of *Y* given *Z*, if and only if the probability distribution of *X* is independent of the value of *Y* given *Z*; mathematically, (17)

$$(\forall i, j, k) P(X=x_i | Y=y_j, Z=z_k) = P(X=x_i | Z=z_k) \tag{2.3}$$

If $\boldsymbol{X}$ denotes a vector of attributes $[x_1, x_2]$ and *Y* denotes a class,

$$P(\boldsymbol{X}|Y) = P(x_1, x_2|Y) \tag{2.4}$$

from general property of probabilities

$$= P(x_1|x_2, Y)\, P(x_2|Y) \tag{2.5}$$

from (2.3)

$$= P(x_1|Y)\, P(x_2|Y) \tag{2.6}$$

Thus it can be summarized that, if $\boldsymbol{X}$ contains *n* attributes which are conditionally independent of one another for a given class *Y*, then (17)

$$P\ (\boldsymbol{X}/Y) = P(x_1, x_2, \ldots, x_n\ /Y) = \prod_{i=1}^{n} P(x_i/Y) \qquad (2.7)$$

From Bayes Rule, we have

$$P(Y = y_k\ /\ x_1..x_n) = \frac{P(Y = yk)\ \prod_i P(xi\ |Y = yk)}{\sum_j P(Y = y\ j) \prod_i P(Xi|Y = yj)} \qquad (2.8)$$

Where Y is a discrete-valued variable denoting the class and $x_1 \ldots x_n$ are the discrete-valued attributes. Equation (2.8) represents the fundamental equation for Naïve Bayes Classifier (17). The Naïve Bayes Classification rule is used to calculate the most probable class to which an unknown sample belongs to (17). It is given by the equation.

$$Y \leftarrow arg\ max_{yk} \quad \frac{P(Y = yk)\ \prod_i P(xi\ |Y = yk)}{\sum_j P(Y = y\ j) \prod_i P(Xi|Y = yj)} \qquad (2.9)$$

Since the denominator is independent of yk, the equation (2.9) can further be reduced to

$$Y \leftarrow arg\ max_{yk}\ P(Y = yk) \prod_i P(xi|Y = yk) \qquad (2.10)$$

2.4.2 Random Forests

Random Forests is one such classification algorithms that directly provides measures of variable importance related to the relevance of the variable in the classification. Developed by Leo Breiman, this classification algorithm employs an ensemble of classification trees. The tree is built using bagging and random variable selection that results in the low correlation of the individual trees. In order to obtain low-bias trees, they are left unpruned. The fundamental principle governing random forests is that for each tree, a random vector is generated such that it is independent of previous random vectors but has the same distribution, and the tree is grown using the training set and the random vector together as inputs (19). This algorithm can also be used for feature selection.

The foremost step of random forests is to form diverse tree classifiers from a single training set. Each tree is built upon a random sample taken with replacement from the training set. This is called “bootstrap sample”. A random subset of the whole variables set is used for splitting the tree nodes. The classification decision of a new case is obtained by majority voting over all trees unless the cut-off value is user defined. In random forests, about one-third of the cases in the bootstrap sample are not used in growing the tree. These cases are called “out-of-bag” (OOB) cases and are used to evaluate the algorithm performance.

There are two measures of importance provided while implementing Random Forests in software package R; “mean decrease in accuracy” and “mean decrease in gini”. Mean decrease in accuracy considers the importance of an m^{th} variable as the difference between the “out-of-bag” error rate for the randomly permuted m^{th} variable (the error rate obtained by randomly rearranging the values of the m^{th} variable for the out-of-bag set, for each tree, and getting new classifications for the forest, by putting this permuted set down the tree.) and the original “out-of-bag” error rate. Mean decrease in gini considers the importance of an m^{th} variable as the sum of all decreases in impurity (measured by gini index) in the forest due to this variable, normalized by the number of trees (20).

The qualities that make Random Forest an ideal classifier are (19).

- It has good predictive performance even when there is noise in the predictive variables.
- It does not have over-fitting problems.

- It can be used both for problems involving two classes as well as multiple number of classes.
- It gives the measures of variable gene importance.
- It involves little need to fine-tune the parameters to achieve excellent performance.

2.4.3 Random Committee

It is one of the metalearning algorithms used in machine learning that takes classifiers and converts them into more powerful learners. Random Committee builds an ensemble of randomizable base classifiers by taking random samples of the same dataset and considering different random seed every time a classifier is built. The final prediction is the average of the predictions made by individual classifiers (17;21).

2.4.4 Relief Algorithm

Relief is an instance based attribute ranking scheme introduced by Kira and Rendell (22). Using this algorithm, the relevance of a feature can be found by estimating its ability to distinguish samples near to each other. The basis on which the algorithm works is that in dimensions of relevant features, the closest sample of the same class is expected to be closer than the closest sample of other classes (23).

Let us consider N training samples:

$\{\mathbf{x(1)},c(1)\},\{\mathbf{x(2)},c(2)\},\{\mathbf{x(3)},c(3)\},\ldots..,\{\mathbf{x(N)},c(N)\}$

Where $\mathbf{x}$ $\mathbf{(k)}$ $=[x_1(k), x_2(k), x_3(k),\ldots, x_n(k)]$ is the feature vector of sample k. $x_1, x_2 ,x_3,\ldots.x_n$ are the available features, and $c(k)$ is the class to which the sample k belongs .

The following criterion is used to estimate the relevance of a feature, say feature x_i

$$C_1 = \frac{1}{N}\sum_{k=1}^{N}[|x_i(k) - x_i^M(k)| - |x_i(k) - x_i^H(k)|] \tag{2.11}$$

Where $x_i^M(k)$ denotes the values of feature of the nearest-miss and $x_i^H(k)$ denotes the values of feature of the nearest-hit samples of sample k. The nearest-hit sample is a term referring to the nearest neighboring sample of the same class, while the nearest-miss sample refers to the nearest neighboring sample of the different class (23). Relief was defined for problems involving 2 classes and was later extended to Relief F algorithm that had the capability to handle noise and multiple datasets (24).

2.4.5 Nearest Centroid Method

Nearest centroid method is a fast and simple algorithm used for classification that works on the basis of classifying an unknown instance to the class whose centroid is closest to it. It considers the centroid of the cluster as a representative of the class. The learnt distance function is used to determine the closest centroid (25).

The arithmetic mean of a class C_j represents the prototype pattern for the class and is denoted by

$${}^{\mu}C_j = \frac{1}{|Cj|}\sum_{xi\in Cj} x_i \tag{2.12}$$

where x_i represents the training samples that have the class $C_{j.}$

Using this algorithm, a class label of an unknown instance x is predicted as:

$$C(x) = arg\ min_{\ Cj}\ d\ ({}^{\mu}C_j, x) \tag{2.13}$$

where $d(x,y)$ denotes the distance function (26).

The distance function measures the strictness of dependence between the two vectors (27). This method is usually preferred in biological applications because of its favorable invariance properties i.e. the correlation between the variables is not affected by an addition of a constant offset to the components of the data or by applying a multiplicative factor (27).

Pearson Correlation provides the degree of linear dependence of vectors *x* and *w* by

$$R\,(x, w) = \frac{\sum_{i=1}^{d}(xi-\ \mu x).(wi-\mu w)}{\sqrt{\sum_{i=1}^{d}(xi-\ \mu x)^{2}}.\sqrt{\sum_{i=1}^{d}(wi-\mu w)^{2}}} \qquad (2.14)$$

where μx and μw are the respective means of the vectors *x* and *w*. The equation is standardized by the multiplication of the standard deviations of the vectors after subtracting their respective means. This causes the Pearson correlation to be invariant (27).

The nearest centroid classification is an efficient method for classifying the new instances without any feature selection. It is one of the simplest and extremely fast classifier. For cases involving two classes, the nearest centroid algorithm is linear and implicitly encodes a thresholding hyperplane that separates the two classes (26).

2.5 Kaplan-Meier Analysis and Log-Rank Test

2.5.1 Kaplan- Meier Analysis

Kaplan-Meier analysis is a recommended statistical technique used in clinical trials for estimating the proportion of the population of people who would survive a given length of time under the same circumstances, given a set of observed survival times including censored times (times for which the period of observation was cut-off before the event of interest occurred) (28).

It is a non-parametric (actuarial) technique that estimates time-related events[6] by analyzing the distribution of patient survival times following their recruitment to a study (29).

[6] http://www.isixsigma.com/dictionary/Kaplan-Meier-780.htm

This analysis allows estimation of survival over time even when patients are censored (dropped out or are studied for different lengths of time)[7]. It is usually followed by plotting the cumulative survival function on a linear scale with the time on the x-axis and the cumulative survival on the y-axis. The curve generally slopes down with fewer surviving cases as the time increases. The plot is generally a step function, in which the estimated survival probabilities are constant between adjacent death times and decrease at each death (28). The steepness of the curve indicates the efficacy of the treatment being investigated. Kaplan-Meier curves can also be used to test the statistically significant differences between the survival curves associated with two different treatments (29).

2.5.2 Mathematical Expression

Let us assume that there are N individuals observed from time 0 to sometime T, the true survival time of each individual is X_i and the distribution function is

$$F^*(x) = P\,(X_i \leq x) \tag{2.15}$$

Let the survival function be

$$F(x) = 1 - F^*(x) \tag{2.16}$$

Let the censoring variables Y_i be independent of the survival variables X_i and have the distribution function

$$H^*(y) = P\,(Y_i \leq y) \tag{2.17}$$

$H^*(y)$ represents the probability that the individual is censored by time y.

[7] http://biostat.mc.vanderbilt.edu/twiki/pub/Main/ClinStat/km.lam.pdf

$$\text{Let } H(y) = 1\text{-}H^*(y) \tag{2.18}$$

Let the observed survival times be denoted as t_i where

$$t_i = X_i \text{ when } X_i \leq Y_i \text{ and } X_i < T. \tag{2.19}$$

The Kaplan-Meier estimate $F_N(t)$ is computed by ranking the values t_i in ascending order $t_1 \leq t_2 \leq t_3 ... \leq t_n$ where, where t_j is the j^{th} largest unique survival time (30).

The estimated conditional probability of surviving beyond time t_j is

$$P_j = 1\text{-} \frac{dj}{rj} \tag{2.20}$$

Where dj is the number of individuals who experience the event of interest at time t_j, and rj is the number of individuals at risk just before t_j, inclusive of those censored at t_j. Basically, the estimated conditional probability represents the ratio of the number of patients surviving beyond time t_j to the number of patients at risk (28;30).

Thus the Kaplan-Meier estimate is the product of these conditional probabilities (28).

$$F_N(t) = \prod_{j=1}^{i} P_j \quad \text{for } t_i \leq t \leq t_{i+1} \tag{2.21}$$

2.5.3 Log- rank test

Log-rank test is a hypothesis test and is the most popular test for comparing the survival of groups. It is a non-parametric test and is sometimes called as Mantel-Cox test. It considers the entire follow-up period of a patient. It is generally used when the data is right-censored i.e. the censoring is non-informative either because there was no event observed till the end of the study or the individual lost to follow-up or the event was not recorded properly [8].

This test was first proposed by Nathan Mantel and was named “log-rank test” by Richard and Julian Peto[9]. The advantage of using the log-rank test is that we need not have prior knowledge of the shape of the survival curve or the distribution of survival times (31).

The null hypothesis considered in this test is that there is no difference between the populations in the probability of an event (relapse or death) at any time point. The analysis is based on the times of events (31). It calculates the observed number of deaths and the expected deaths in each group for each observed time. It thus compares the estimates of the hazard functions of the two groups at each observed time.

When the risk of an event is consistently greater for one group than the other, the log-rank test most likely detects a difference between the groups. However, it is unlikely to detect a difference when survival curves cross. When there is no censorship (loss to follow-up) the log rank test gets reduced to Mann-Whitney test (two-sample Wilcoxon test) for two groups of survival times and Kruskal-Wallis test for more than two groups of survival times[10].

The log-rank test compares each observed event in group i at a distinct time j; $O_{i\,j}$ to its expectation E_j under the null hypothesis and is defined as

[8] http://www.statsdirect.com/help/survival_analysis/logrank.htm

[9] http://en.wikipedia.org/wiki/Logrank_test

[10] http://www.statsdirect.com/help/survival_analysis/logrank.htm

$$Z = \frac{\sum_{j=1}^{J}(Oij - Ej)}{\sqrt{\sum_{j=1}^{J} Vj}} \tag{2.22}$$

where Vj is the variance of the distribution at the distinct time j[11].

2.6 Cox Proportional Hazard Model

Cox proportional hazard model, introduced by Sir Cox in 1972 (32) is a highly-well recognized statistical technique that finds application in the field of medical statistics. It introduces a modeling approach to analyze the survival data and enables one to explore the relationship between the survival of a patient and several explanatory variables (28). It is extremely useful in the cases that involve more than one explanatory variable. It provides an estimate of the risk of a patient towards an event (relapse or death) for individuals, given their prognostic variables. This model generates the coefficients for each explanatory variable. The interpretation of the coefficients is as follows: a negative value for the regression coefficient implies lower hazard risk or better prognosis; a positive value for the regression coefficient implies higher hazard risk or poor prognosis.

The main factor governing the survival function is the hazard function *h(t)* which is defined as the probability that an individual experiences an event in a small time interval given that the individual has survived to a time up to the beginning of the interval (33). It is expressed as

$$h(t) = lim_{s \to 0} P(t \leq T \leq t + s \,|T \geq t) \tag{2.23}$$

[11] http://en.wikipedia.org/wiki/Logrank_test

Or

$$h(t) = \frac{number\ of\ patients\ experiencing\ an\ event\ in\ a\ time\ interval\ beginning\ at\ t}{(number\ of\ patients\ surviving\ at\ time\ t)x(width\ of\ time\ interval)} \qquad (2.24)$$

Let us consider that there are n explanatory variables expressed as x_1,x_2, x_3, ... x_n. The hazard function is then expressed as

$$h(t) = h_0(t)\ x\ exp(\beta_1.x_1 + \beta_2.x_2 + \beta_3.x_3\ ....\ + \beta_n.x_n) \qquad (2.25)$$

where β_1 to β_n represent the regression coefficients that are generally estimated by a statistical method called maximum likelihood (28).

Taking the natural logarithms on both sides of the equation, we get

$$ln\ h(t) = ln\ h_0(t) + \beta_1.x_1 + \beta_2.x_2 + \beta_3.x_3\ ....\ + \beta_n.\ x_n \qquad (2.26)$$

The factor $h_0(t)$ is known as baseline hazard function which represents the hazard function for the patients when all the explanatory variables are zero. The exponential of the regression coefficients; exp (β_j) provides the relative risk change that is associated with the increase of the covariate x_j by one unit. The significance of each variable is obtained by dividing the regression coefficient by its standard error (measure of uncertainty of an estimate), and comparing this value with the standard normal distribution. A value > 1.96 is implies that the variable is statistically significant (28).

This model yields an equation for the hazard as a function of several explanatory variables and forces the hazard relation between the two patients to be constant over time. Thus, it is said to be a proportional hazard model.

2.7 Related Studies

There have been numerous studies in the past, related to the application of machine learning in breast cancer prediction and prognosis. This section discusses in brief about such studies.

Sotiriou et al. (34) adopted a population-based approach to determine the genes associated with improved relapse-free survival. They identified a set of 485 probe elements from a total of 7650 probe elements by performing Cox proportional hazard regression analysis. This set could separate the relapse-free survival in 99 patients with a $p<0.05$. 16 probe elements were significantly associated with relapse-free survival at a stringent significance level of $p<0.001$. To identify a minimal number of the most important prognostic genes, the list of 485 probe elements was overlapped with 231 genes present in the prognostic gene set in van't Veer et al. (9). This overlap resulted in 11 unique genes represented by 14 probe elements.

van't Veer et al. (9) reported the development data for the 70-gene set which was identified from 117 primary breast tumors, analyzed on DNA microarray platform that contained 25000 genes. This gene set is the basis for the MammaPrint test (1). Van de Vijver et al. (10) performed the first major validation of the 70-gene signature that was reported in van't Veer et al. (9) by classifying a series of 295 patients with primary breast carcinomas to poor prognosis or good prognosis using microarray analysis.

Sorlie et al. (35;36) classified breast carcinomas that were based on variations in gene expression patterns derived from cDNA microarrays, and correlated the tumor characteristics to clinical outcome. They found that this classification could be used as a prognostic marker for relapse-free and overall survival in a subset of patients who received uniform therapy. It was reported previously that cancers could be classified into groups such as, basal epithelial-like group, ERBB2-overexpressing group and normal breast-like group based on variations in gene expression. Sorlie et al. (36) found that the previously characterized luminal epithelial/estrogen receptor positive group could be divided into at least two subgroups, each with a distinctive expression profile. They further refined the previously defined subtypes of breast tumors that could be distinguished by their distinct patterns of gene expression, by analyzing 115 malignant breast tumors using hierarchical clustering.

Wang et al. (37) identified a 76-gene signature from a training set of 115 tumors, of which there were 80 estrogen receptor (ER)-positive and 35 ER-negative tumors. The patients were grouped on the basis of the ER status, and each subgroup was analyzed separately for the selection of biomarkers. The 76-gene signature was obtained from a combination of 60 genes that were identified from ER-positive subgroup, and 16 genes from ER-negative subgroup. This signature served as a powerful tool for identification of patients at high-risk of distant recurrence.

Bild et al. (38) identified gene expression signatures that reflected the activation status of several oncogenic pathways. They evaluated these gene expression signatures in several large collections of human cancers, resulting in identification of patterns of pathway deregulation in tumors and association with disease outcomes.

Miller et al. (39) identified a 32-gene expression signature that distinguishes p53-mutant and wild-type tumors of different histologies, by analyzing transcript profiles of 251 p53-sequenced primary breast tumors.

Ivshina et al. (40) identified 264 robust grade-associated markers from a study of expression profiles of 347 primary invasive breast tumors analyzed on affymetrix microarrays. Class prediction algorithms were used, six of which could accurately classify Grade 1 and Grade 3 tumors, and separate Grade 2 tumors into two highly discriminate classes: namely, G2a and G2b.

Loi et al. (41) assigned ER-positive breast cancer patients to either high or low genomic grade subgroups by using the gene expression grade index (GGI) algorithm which defines histological grade based on gene expression profiles. These subgroups were compared with previously reported ER-positive molecular classifications.

2.8 Summary

This chapter reviewed the various machine learning algorithms and the statistical techniques that were employed in our research. The related work done in other previous studies were briefly discussed. These studies have adopted the common practice of identifying a gene signature by using only one machine learning algorithm. Nevertheless, the high dimensionality of DNA microarray data requires integrating multiple feature selection algorithms at different stages of gene selection to obtain better performance. We thereby developed a scheme to combine several feature selection algorithms to identify novel disease biomarkers. The data from various publications were used in the identification and validation of the gene signatures. The subsequent chapters describe our work in detail.

Chapter 3

Developing a gene expression signature of genomic instability that serves as an independent predictor of clinical breast cancer outcomes

3.1 Introduction

In this study, we recognized that the level of genomic instability (as reflected by the variability of DNA content; aneuploidy) is strongly associated with breast cancer prognosis (4;5). In predicting breast cancer outcomes, the observations resulting from nuclear DNA content and genomic instability were found to be similar to those resulting from gene expression signatures (4-7). Moreover, Kaplan-Meier curves that assessed the recurrence-free survival were identical when using either gene expression signatures or genomic instability as independent variables. These facts provoked the assumption that the two are connected. We sought to determine the nature of the connection by considering the following hypotheses: first, the degree of genomic instability is reflected by an aneuploidy-specific gene signature (set of genes with large variability in DNA content); second, this signature is robust and allows breast cancer prediction of clinical outcomes in independent datasets. The first hypotheses was tested by gene expression profiling of 48 breast cancer carcinomas with defined patterns and varying degrees of genomic instability; whereas, the second hypothesis was tested through classification of published breast cancer datasets using the gene expression signature of genomic instability.

In this chapter, we discuss in detail about the development and validation of a 12-gene genomic instability signature that resulted from an attempt to explore the differences between different genomic instability groups. It was identified from a dataset that contained gene expression profiling of 7657 genes on 48 breast cancer patients having varying degrees of genomic instability. A supervised machine learning approach was adopted in multiple settings for performing feature selection using a data-mining tool, WEKA[12].

Next, the extent to which the gene expression signature (that defined genomic instability) could predict the breast cancer outcomes in previously published datasets was determined. This was accomplished by using the nearest centroid classification algorithm to perform patient stratification in the validation datasets. The significance of this stratification was tested using Kaplan-Meier analyses followed by log-rank test in R[13].

The association between the genomic instability-defined risk groups and traditional prognostic factors of breast cancer such as lymph node status, tumor grade, NIH consensus criteria (42) and St.Gallen criteria (43) was evaluated. Furthermore, it was investigated whether the 12-gene genomic instability signature could provide additional prognostic information, within the subgroups defined by traditional clinical-pathological factors.

The remainder of this chapter is organized as follows: Section 3.2 briefly introduces the datasets that were used in this study, Section 3.3 discusses the process of biomarker identification, Section 3.4 elucidates the validation of the genomic instability gene signature for disease-free and overall survival prediction on previously published datasets, Section 3.5 discusses the association of the gene expression-defined groups and clinical parameters, and finally Section 3.6 summarizes the chapter.

[12] http://www.cs.waikato.ac.nz/ml/weka/

[13] http://www.r-project.org/

3.2 Acquisition of data

The data that was used as a training dataset was obtained from Dr. Thomas Ried, NCI. This data contained 7657 genes and 48 primary breast cancer specimens collected at the Karolinska Institute and Hospital, Stockholm, Sweden during 2000 and 2001. The data was analyzed using global gene expression profiling on cDNA arrays. This analysis was complemented by mapping of genomic imbalances using comparative genomic hybridization. 17 of these tumors were classified as diploid; genomically stable (dGS), 15 as aneuploid; yet genomically stable with a defined stemline (aGS), and 16 as aneuploid and genomic unstable (aGU). This dataset was subjected to quality assessment and 4 samples were discarded as they did not pass the quality assessment criteria. Finally, there were a total of 44 samples; 14 dGS, 14 aGS, and 16 aGU.

The validation datasets were obtained from various patient cohorts mentioned in previous related publications. The following datasets were used as validation datasets.

1) Sotiriou et al. (34) (PMID: 12917485) - This cohort contains 99 node-negative and node-positive breast cancer patients. All of the tumor samples were invasive ductal carcinomas; 46 individuals were node-negative and 53 were node-positive; 16 patients with tumor grade 1, 38 patients with tumor grade 2, and 45 patients with tumor grade 3; 65 estrogen receptor (ER)-positive and 34 ER-negative patients. Two patients received PMF Chemotherapy; 30 patients received CMF; and two received Adr, CMF chemotherapy. The dataset is publically available at the PNAS website[14].

[14] http://www.pnas.org/cgi/content/full/1732912100/DC1

2) Sorlie et al. (35) (PMID: 11553815) - This cohort contains 75 breast carcinomas (66 ductal, five lobular, 1 pleomorph, 1 mucinous, 1 papillary and 1 DCIS). Fifty-one patients were treated with doxorubicin monotherapy before surgery followed by adjuvant tamoxifen in the case of positive ER and/or progesterone receptor (PgR) status. This cohort contains 56 ER-positive patients and 17 ER-negative patients. Nine patients had Grade 1, 33 patients had Grade 2, and 32 patients had Grade 3. The cohort contains 23 lymph node-positive and 52 lymph node- negative patients. The dataset is publically available at the Gene Expression Omnibus database with the accession number GSE3193.

3) Van de Vijver et al. (10) (PMID: 12490681) - There were 295 consecutive patients with primary breast carcinomas, 151 with lymph node-negative disease, and 144 with lymph node-positive disease. The dataset is publically available at the Rosetta Inpharmatics website[15]

3.3 Biomarker Identification

This section describes the identification of the genomic instability signature. A supervised machine learning approach involving a combination of feature selection algorithms was adopted in two sample settings, to identify a genomic instability signature from the expression profiles of 7657 genes on 44 breast cancer samples. In the first setting, a binary classification was done to explore the differences between the genomic stable and the genomic unstable group. To accomplish this, feature selection was performed on the dataset using Random Forests with the help of varSelRF package in R (33). This algorithm builds trees upon a bootstrap sample. The performance of the algorithm was evaluated by one-third of the cases that were not used for

[15] http://www.rii.com/publications/2002/nejm.html

growing the trees, namely, out-of-bag (OOB) cases. This algorithm was carried out in a series of following steps:

1. A large forest with 2000 trees was built based on all of the 7657 genes and the importance measure of the genes was obtained.

2. 20% of the least important genes were filtered out and a forest was built with the remaining number of genes to get the OOB error estimate.

3. The step 2 was repeated until one or two genes were left, and

4. The gene set which had the smallest OOB error rate was selected.

In this process, a set of 7 genes was selected that had the smallest OOB error rate. Table 3.1 enlists the 7-gene signature.

Table 3.1: List of 7-gene Signature

GENE NAME	MAP	CLONE ID
HNF3A—hepatocyte nuclear factor 3, al	14q12-q13	1711594
Homo sapiens mRNA; cDNA DKFZp762M127	11	1822809
STK15—serine/threonine kinase 15	20q13.2-q13.3	2007691
KIAA0882—KIAA0882 protein	4q31.1	2190664
MYB—v-myb avian myeloblastosis viral oncog	6q22-q23	2555590
RERG—RAS-like, estrogen-regulated, gr	12p13.1	644989
Incyte EST	NaN	88935

The classification accuracy of the selected gene set was evaluated using the Random Committee algorithm in WEKA (44) with a leave-one-out cross validation. We performed several experiments using the algorithms implemented in WEKA, and selected Random Committee as the most appropriate for the classification of the breast cancer patients into the genomic stability (dGS + aGS) vs. instability (aGU). This algorithm generated a classification accuracy of **93.18%.** The percentage of correctly classified genomically stable (aGS +dGS) patients was **92.86%** and that of the genomically unstable (aGU) patients was **93.75%.** The results of the binary classification are represented graphically in Figure 3.1.

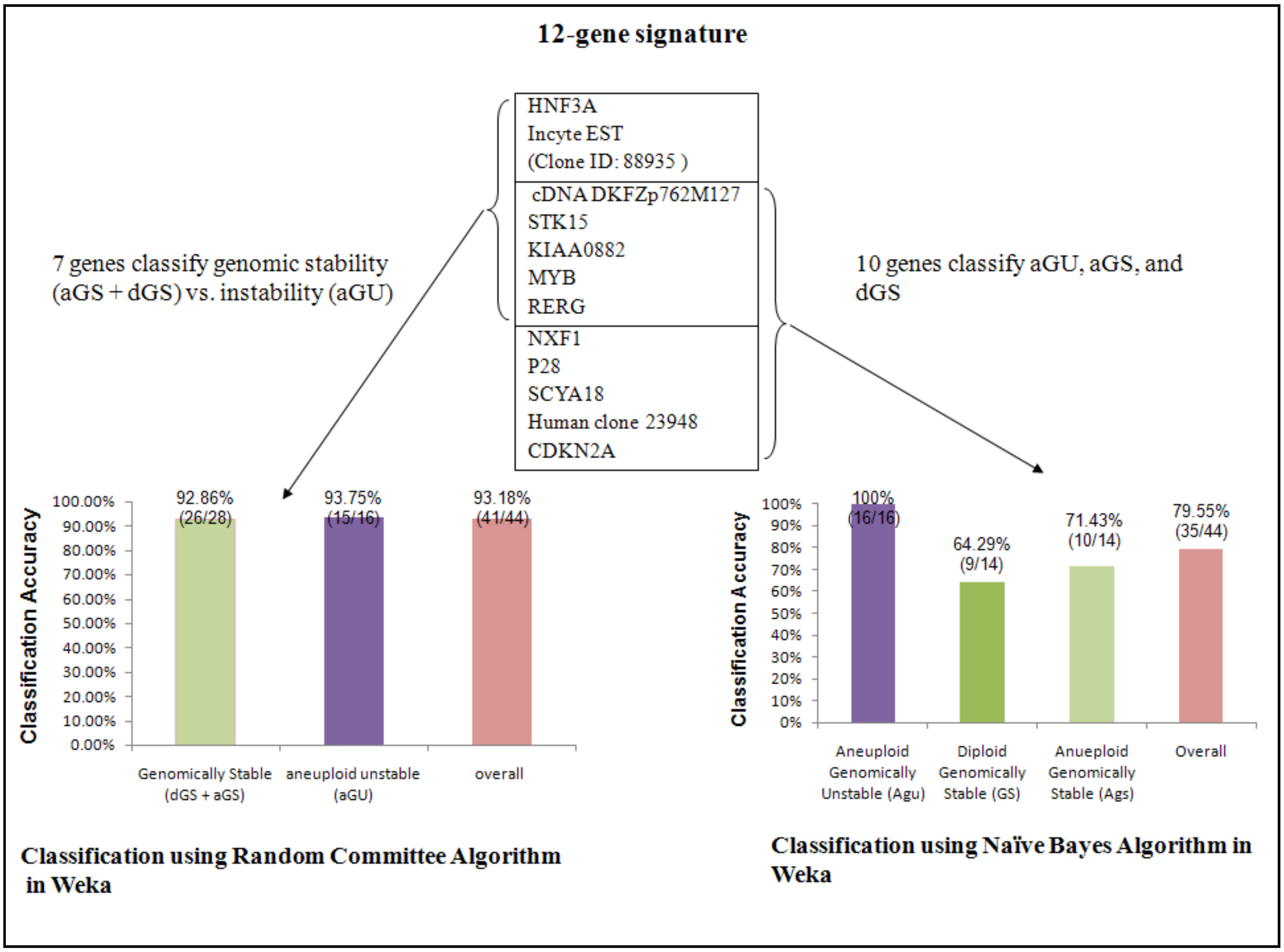

Figure 3.1: The results of binary and multi-classification of the patients (n=44) in the training dataset.

In the second setting, the differences among the three groups were explored by performing a multi-classification. Feature selection was performed in a similar manner using Random forests. This generated a set of 70 genes that had the smallest OOB error rate. Table 3.2

enlists the list of 70-gene signature. This gene signature was further subjected to feature selection using Relief algorithm in WEKA (24) which ranked the 70 genes in the order of importance. The top 10 genes were selected from the 70-gene set. Table 3.3 enlists the list of 10-gene signature.

Table 3.2: List of 70 gene signature

Clone ID	Gene Name
1269591	Homo sapiens mRNA; cDNA DKFZp434E033
1304879	DAPP1—dual adaptor of phosphotyrosine
1309376	Homo sapiens cDNA FLJ13092 fis, clone
1349857	IARS—isoleucine-Trna synthetase
1402715	ESDN—endothelial and smooth muscle ce
142949	STK17B—serine/threonine kinase 17b (a
1453049	SCNN1A—sodium channel, nonvoltage-gat
1481225	ADD3—adducin 3 (gamma)
1506093	GOLGB1—golgi autoantigen, golgin subf
157510	CHI3L1—chitinase 3-like 1 (cartilage
1611623	PLSCR1—phospholipid scramblase 1
1624206	PRP17—pre-mRNA splicing factor 17
1648517	ATP12—homolog of yeast ATP12
1662893	C18orf1—chromosome 18 open reading fr
1674405	ABCE1—ATP-binding cassette, sub-famil
1690295	potassium voltage-gated channel, subfa
1711594	HNF3A—hepatocyte nuclear factor 3, al
1722870	NXF1—nuclear RNA export factor 1
1724982	TP53BP1—tumor protein p53 binding pro
1793853	ALCAM—activated leucocyte cell adhesi
1796576	calcium channel, voltage-dependent, L
1803418	KIAA0089—KIAA0089 protein
1808121	KIAA1324—KIAA1324 protein
1809315	NISCH—nischarin
1813269	CES1—carboylesterase 1 (monocyte/mac
1822809	Homo sapiens mRNA; cDNA DKFZp762M127
1844691	ALE2—armadillo repeat protein ALE2
1850249	Homo sapiens cDNA FLJ11375 fis, clone
1879041	MYBL1—v-myb myeloblastosis viral onco
1967307	CG005—hypothetical protein from BCRA2
1968576	FBP1—fructose-1,6-bisphosphatase 1
1985366	PPP1CA—protein phosphatase 1, catalyt
1998792	P28—dynein, aonemal, light intermedi
2007691	STK15—serine/threonine kinase 15
2013673	CTNS—cystinosis, nephropathic
2045455	C20orf12—chromosome 20 open reading f
2057823	E2-EPF—ubiquitin carrier protein
2133608	TTK—TTK protein kinase
2190664	KIAA0882—KIAA0882 protein

2242817	Homo sapiens, clone MGC:22588 IMAGE:46
2285109	Homo sapiens, Similar to RIKEN cDNA 17
2288855	microtubule-associated protein tau
2366522	MGC4251—hypothetical protein MGC4251
2414624	MAD2L1—MAD2 mitotic arrest deficient-
2444942	CENPA—centromere protein A (17Kd)
2498968	KIAA0753—KIAA0753 gene product
2500225	Homo sapiens clone 24405 mRNA sequence
2555590	v-myb avian myeloblastosis viral oncog
2590131	CD3G antigen, gamma polypeptide (TiT3
2608629	cytochrome P450, subfamily IA (stero
2716261	Homo sapiens cDNA FLJ20115 fis, clone
2740235	CDKN2A—cyclin-dependent kinase inhibi
2791936	TRIM28—tripartite motif-containing 28
2833929	ESTs
2875922	Homo sapiens, clone IMAGE:3448367, Mrn
3123244	Human clone 23948 mRNA sequence
3127171	Human glucocorticoid receptor alpha Mr
3242480	DLG5—discs, large (Drosophila) homolo
3251982	PTPRT—protein tyrosine phosphatase, r
3279439	GOSR2—golgi SNAP receptor comple mem
3451473	ESTs
3970665	microseminoprotein, beta-
447148	Homo sapiens cDNA: FLJ23005 fis, clone
515453	KRAS2—v-Ki-ras2 Kirsten rat sarcoma 2
553251	QDPR—quinoid dihydropteridine reducta
644989	RERG—RAS-like, estrogen-regulated, gr
690231	SCYA18—small inducible cytokine subfa
740878	DUSP4—dual specificity phosphatase 4
88935	Incyte EST
962043	Homo sapiens clone 23736 mRNA sequence

Table 3.3: List of 10 gene signature

Clone ID	Gene Name	MAP
1722870	NXF1—nuclear RNA export factor 1	11q12-q13
1822809	Homo sapiens mRNA; cDNA DKFZp762M127	11
1998792	P28—dynein, aonemal, light intermedi	1p35.1
2190664	KIAA0882—KIAA0882 protein	4q31.1
2555590	v-myb avian myeloblastosis viral oncog	6q22-q23
2740235	CDKN2A—cyclin-dependent kinase inhibi	9p21
3123244	Human clone 23948 mRNA sequence	15q22.32
644989	RERG—RAS-like, estrogen-regulated, gr	12p13.1
690231	SCYA18—small inducible cytokine subfa	17q11.2
2007691	STK15—serine/threonine kinase 15	20q13.2-q13.3

The classification accuracy of this set of genes was determined using the Naïve Bayes algorithm in WEKA with leave-one-out cross validation. This algorithm generated a classification accuracy of **79.55%**. The percentage of correctly classified aGS was **71.43%**, dGS was **64.29%,** and aGU was **100%.** The results of the multi-classification are graphically represented in Figure 3.1.

Thus the resulting two largely concordant signatures from both the approaches confirmed the relevance of the identified signature genes as descriptors of genomic instability. These two signatures (that had 5 genes in common) were combined resulting in the 12-gene genomic instability signature list. Among the 12-gene genomic instability signature, SCYA18, STK15 and CDKN2A were over expressed in genomically unstable breast carcinomas, while the remaining genes were under expressed in genomically unstable tumors ($p < 0.001$, two sided t-tests). This gene signature was then used to predict breast cancer outcomes in previously published independent datasets. Table 3.4 enlists the list of 12-gene signature.

Table 3.4: List of 12 gene signature

Clone ID	Gene Name	Expression in Genomic Unstable vs. Stable	p-value (two sided t-test)
1722870	NXF1—nuclear RNA export factor 1	Under Expressed	0.001154
1822809	Homo sapiens mRNA; cDNA DKFZp762M127	Under Expressed	1.11E-05
1998792	P28—dynein, aonemal, light intermedi	Under Expressed	2.12E-06
2190664	KIAA0882—KIAA0882 protein	Under Expressed	6.13E-08
2555590	v-myb avian myeloblastosis viral oncog	Under Expressed	5.07E-06
2740235	CDKN2A—cyclin-dependent kinase inhibit	Over Expressed	0.000116
3123244	Human clone 23948 mRNA sequence	Under Expressed	1.05E-05
644989	RERG—RAS-like, estrogen-regulated, gr	Under Expressed	9.56E-07
690231	SCYA18—small inducible cytokine subfa	Over Expressed	9.85E-05
2007691	STK15—serine/threonine kinase 15	Over Expressed	2.48E-08
171194	HNF3A—hepatocyte nuclear factor 3, al	Under Expressed	7.42E-05
88935	Incyte EST	Under Expressed	8.28E-05

The classification accuracy of the 12-gene signature in classifying 44 breast cancer tumors into genomic stability (GS) and genomic unstability (aGU) was determined using the Naïve Bayes algorithm in WEKA with leave-one-out cross validation. This algorithm generated a classification accuracy of **97.73%**. The percentage of correctly classified GS was **96.43%** and aGU was **100%**. This accuracy was higher than that generated individually by the 7-gene signature and the 10-gene signature, thereby confirming the improvement in the performance of the combined gene signatures over the individual gene signatures. The confusion matrix generated as a result of the classification is as shown below in Table 3.5.

Table 3.5: Classification accuracy of 12-gene signature

Classified as =>	**GS**	**aGU**	**Accuracy for each class**
GS	27	1	96.43%
aGU	0	16	100%

Overall Accuracy = **97.73%**

The 12-gene signature was further subjected to unsupervised validation by performing a hierarchical clustering analysis with CIMminer (45), to group 44 breast carcinomas. The gene expression was aggregated based on Euclidean distance with average linkage. The distance of the samples was computed based on correlation and the cluster method was complete linkage. The cluster analysis as shown in Figure 3.2, represents the aggregation into two groups separating genomically stable (dGS and aGS) from unstable tumors (aGU). It confirms a linkage between the degree of genomic instability and gene-expression patterns.

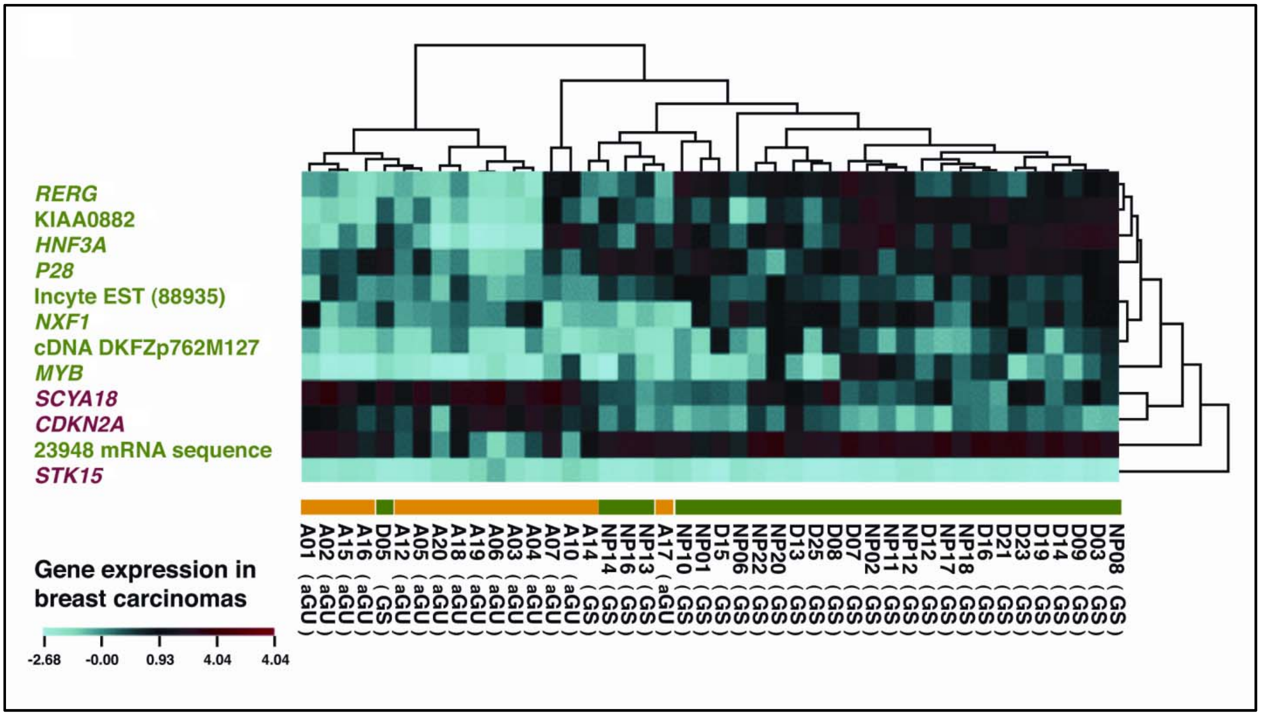

Figure 3.2: Hierarchical clustering analyses with CIMminer performed on 44 breast carcinomas using the 12-gene signature.

3.4 Validation of the genomic-instability gene signature for breast cancer prognosis

This study sought to explore the extent to which the gene expression signature (that defines genomic instability in the breast cancer) could be used for prediction of disease outcome in previously published independent datasets.

Various classification algorithms were tried in WEKA using a leave-one-out cross validation technique to classify the samples in each of the datasets. The samples in each of the datasets were first stratified into low risk and high risk depending on the survival information and status of the clinical outcome. The criteria that were considered for the stratification were:

RFS ≤ 5 years and Status =1 -> High risk

RFS > 5 years and Status = 0 -> Low risk

The classification algorithms in WEKA were used to obtain the classification accuracy. Since the datasets were generated on diverse microarray platforms and had incompatible expression profiles, the classification algorithms in WEKA failed to give consistent results across the datasets. A single classification model could not be used across all the datasets. Moreover, the results of these classification algorithms gave an estimate over a specific period of 5 years. The need to analyze the clinical outcomes over the time-course was identified which was possible by Kaplan-Meier survival curves. The results obtained by performing a leave-one-out cross validation using various classification algorithms in WEKA are as shown in Table 3.6.

Table 3.6: Classification analyses using various classification algorithms in WEKA

Datasets	**Van de Vijver**				**Sotiriou**				**Sorlie**			
Algorithms	Specificity (Low Risk) (%)	Sensitivity (High Risk) (%)	Overall Accuracy	(Specificity + Sensitivity) /2	Specificity (Low Risk) (%)	Sensitivity (High Risk) (%)	Overall Accuracy	(Specificity + Sensitivity) /2	Specificity (Low Risk) (%)	Sensitivity (High Risk) (%)	Overall Accuracy	(Specificity + Sensitivity) /2
Naïve Bayes	62.05	74.71	66.8	68.38	100	0	57.95	50	36.36	82.35	71.11	59.36
Neural Network	77.51	54.02	69.53	65.77	68.63	56.76	63.64	**62.7**	36.36	85.29	73.33	60.83
IBk	72.78	41.38	62.11	57.08	58.82	56.76	57.95	57.79	36.36	85.29	73.33	60.83
Random Committee	81.66	22.99	61.72	52.33	82.35	40.54	64.77	61.45	18.18	94.12	75.56	47.15
HyperPipes	95.86	9.2	66.41	52.53	86.27	13.51	55.68	49.89	0	91.18	68.89	45.59
Random Forest	81.66	28.74	63.67	55.2	72.55	40.54	59.09	56.55	36.36	91.18	77.78	**63.77**
Decision Table	62.72	74.71	66.8	**68.72**	100	0	57.95	50	0	91.18	68.89	45.59
Average	76.32	43.68	65.29	60.00	81.23	29.73	59.58	55.48	23.37	88.66	72.7	54.73

Considering the inconsistent results of the above analyses and the need for a single classification algorithm, the nearest centroid classification method was adopted for evaluating the accuracy of the identified 12-gene genomic instability signature on other datasets used for validation. The accuracy of the gene signature was evaluated on 496 tumor profiles in breast cancer that were obtained from the published datasets:

Matchminer (46) was used to obtain the gene names for those genes that had either clone id or affymetrix id as a gene-identifier in the validation datasets. The genes in the signature were identified in each of the datasets used for validation. The average gene expression of each gene in each group namely, genomic stable (GS) and genomic unstable (aGU), was computed from the training dataset and was considered as a standardized centroid. Each patient in the validation cohorts was classified into GS group or aGU group based on the Pearson correlation of the patient's gene expression profiles with the average expression profiles (centroids) of the GS and aGU group in the training data. Table 3.7 contains the gene expression centroids obtained for each gene in each of the two groups.

Table 3.7: Average gene expression profiles (centroids) for 12 genes

Gene names	aGU	GS
CDKN2A--cyclin-dependent kinase inhibi	1.06749	0.046581
HNF3A--hepatocyte nuclear factor 3, al	-0.46792	1.393824
Homo sapiens mRNA; cDNA DKFZp762M127	-0.70024	0.119753
Human clone 23948 mRNA sequence	0.940516	2.622091
Incyte EST	0.061511	0.55565
KIAA0882--KIAA0882 protein	-0.49695	0.955434
NXF1--nuclear RNA export factor 1	0.015858	0.639736
P28--dynein, axonemal, light intermedi	0.199474	1.202058
RERG--RAS-like, estrogen-regulated, gr	-0.30429	0.925177
SCYA18--small inducible cytokine subfa	2.343289	0.710686
STK15--serine/threonine kinase 15	-1.04827	-1.90629
v-myb avian myeloblastosis viral oncog	-1.46513	-0.03942

In the external validation, patients were classified as GS (genomically stable) if the correlation of the gene expression with the average GS centroid was higher than that with the average aGU centroid. Similarly, patients were classified as aGU (genomically unstable) if the correlation of the gene expression with the average aGU centroid was higher than that with the average GS centroid. If there were multiple probes for the same annotated gene, the average of the gene expressions was computed for all the probes and used in the correlation analysis. In order to compare the performance with other signatures, no threshold was set on correlation coefficients in patient classification and no patient was removed. Thus, using the nearest centroid classification algorithm, the patients in each validation cohort were classified to the group with the centroid, to which the gene expression profile of the new sample was closest to in squared distance[16]. The distribution of the correlation coefficients for each validation dataset is as shown in Figures 3.3-3.5.

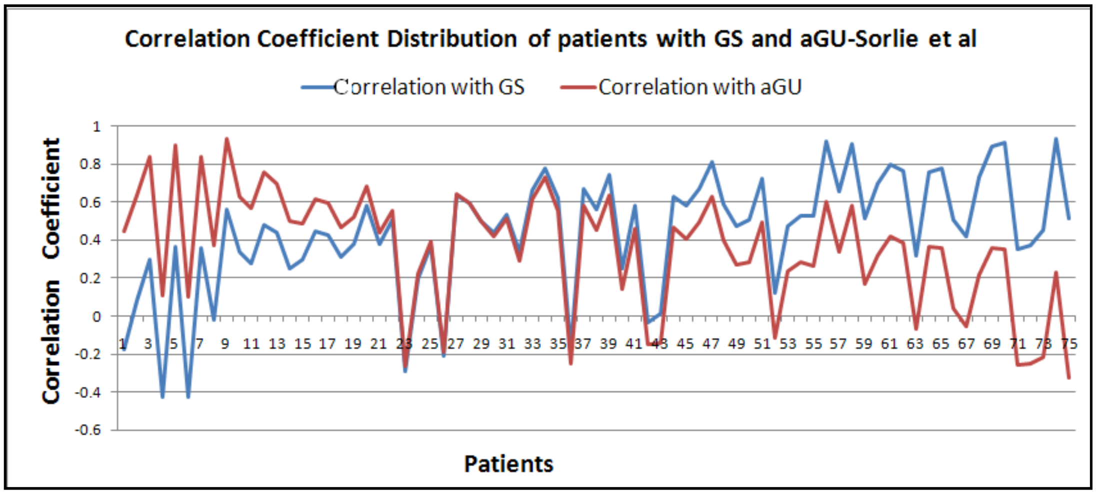

Figure 3.3: Correlation coefficients with the GS and aGU centroids in patients from Sorlie et al.

[16] http://www-stat.stanford.edu/~tibs/PAM/Rdist/howwork.html

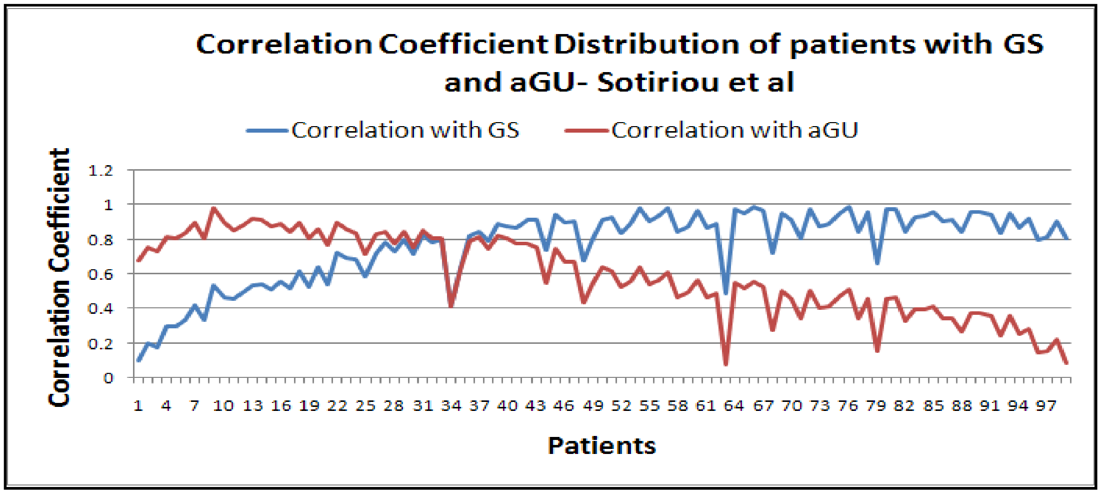

Figure 3.4: Correlation coefficients with the GS and aGU centroids in patients from Sotiriou et al.

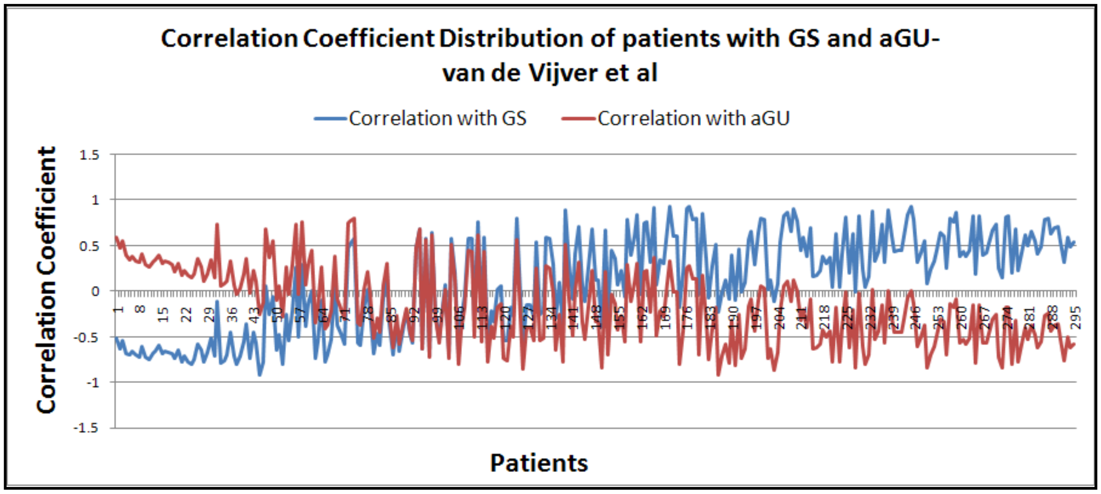

Figure 3.5: Correlation coefficients with the GS and aGU centroids in patients from Van de Vijver et al.

To test the statistical significance of this classification, Kaplan-Meier analysis was performed. For each validation dataset, the predictive class that was obtained from the nearest centroid classification was taken and survival curves were plotted using the survival package in R (33). Statistical significance of the difference between the survival curves for different prognostic groups was assessed using likelihood ratio tests and log-rank tests.

Kaplan-Meier analyses showed that genomic instability-defined prognostic groups were associated with distinct relapse-free and overall survival (*p < 0.05*, log-rank tests) despite the fact that about 50% of the patients died without having suffered from breast cancer recurrence (8). Patients with GS signature had longer relapse-free survival and overall survival than those with the aGU signature. The survival curves for each validation dataset are as shown in the Figures 3.6-3.8.

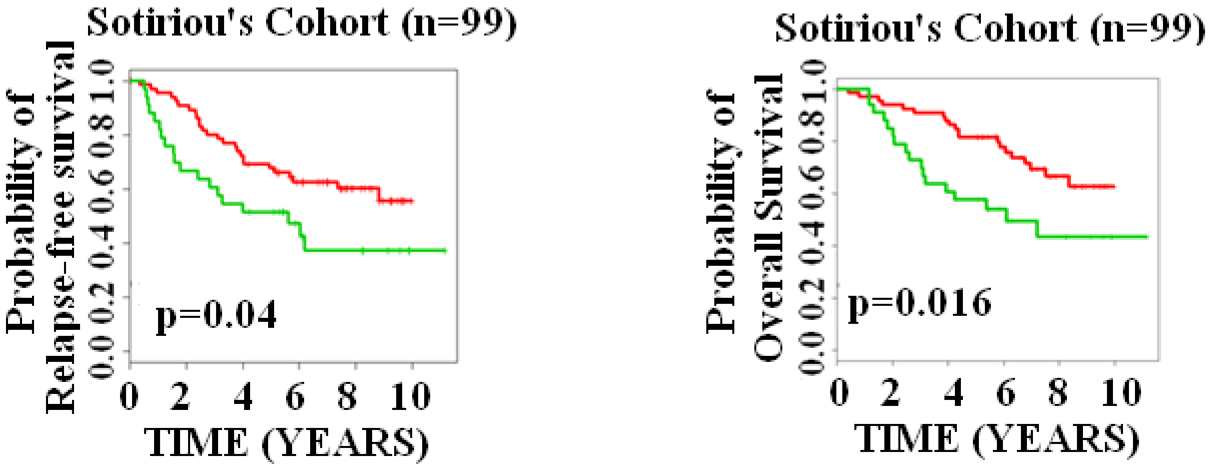

Figure 3.6: The 12-gene signature classifies breast cancer patients from Sotiriou's cohort into prognostic subgroups with distinct relapse-free survival and overall survival in Kaplan-Meier analysis (*p <0.05*, log-rank test). The curves in red represent the genomically stable group and the curves in green represent the genomic instability group.

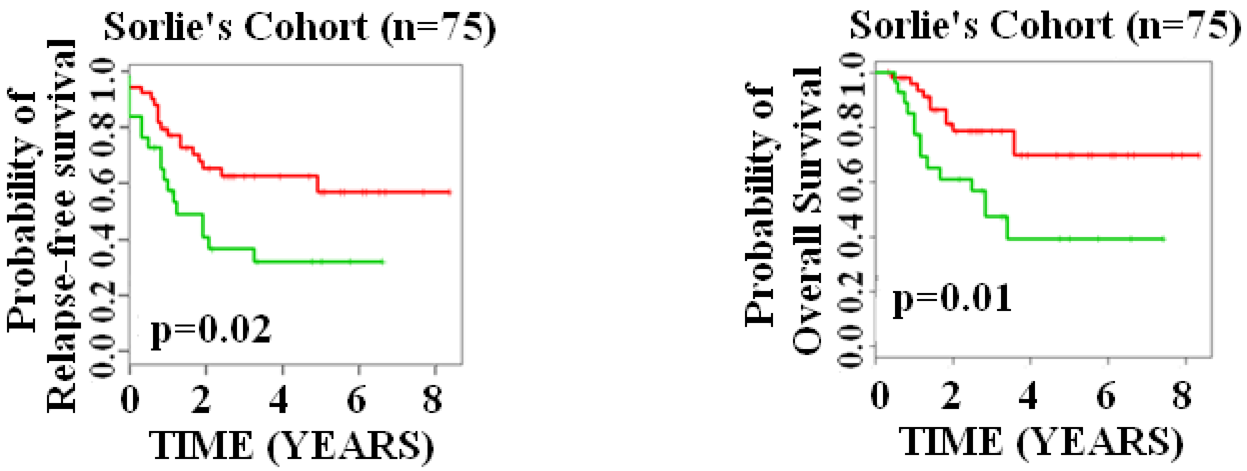

Figure 3.7: The 12-gene signature classifies breast cancer patients from Sorlie's cohort into prognostic subgroups with distinct relapse-free survival and overall survival in Kaplan-Meier analysis (*p <0.05*, log-rank test). The curves in red represent the genomically stable group and the curves in green represent the genomic instability group.

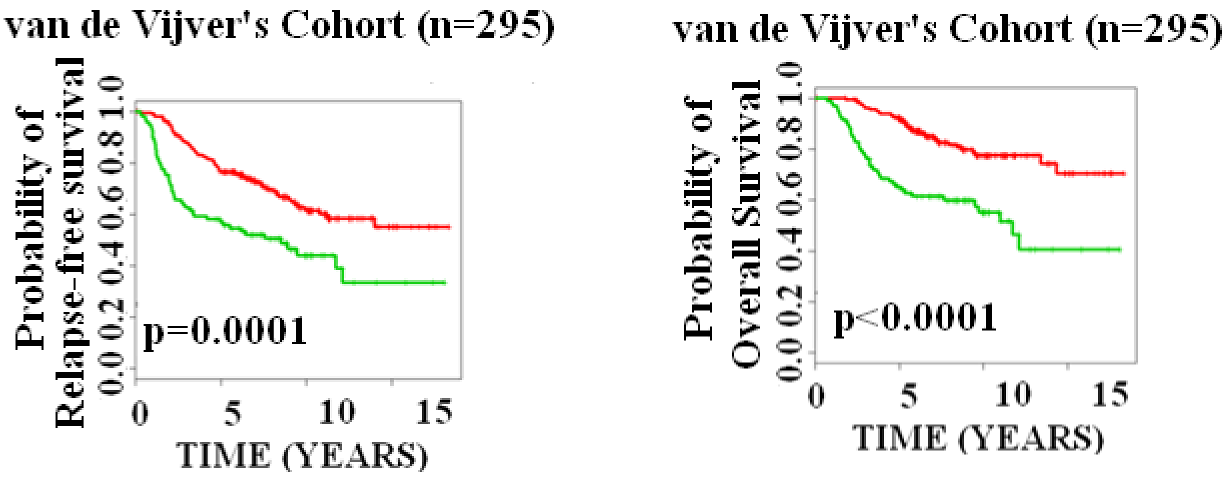

Figure 3.8: The 12-gene signature classifies breast cancer patients from van de Vijver's cohort into prognostic subgroups with distinct relapse-free survival, overall survival in Kaplan-Meier analysis (*p <0.05*, log-rank test). The curves in red represent the genomically stable group and the curves in green represent the genomic instability group.

3.5 Association of genomic instability-defined risk groups and clinical parameters

In this study, the association between genomic instability-defined risk groups and traditional prognostic factors of breast cancer was evaluated by combining all the validation datasets together and performing either Pearson's Chi-squared test or Fisher's exact test (two-sided) between the genomic instability-defined risk groups and each of the traditional (clinical) factors such as lymph node status, tumor grade, estrogen receptor status, and age. Chi-squared test was performed using the 'chisq.test' function in R (33). Chi-squared test was used if its assumptions were satisfied. Otherwise, Fisher's exact test was applied using the R function 'fisher.test' (33). Table 3.8 reports the *p* values resulted from the tests. A $p < 0.05$ indicates a significant association between the genomic instability signature and the corresponding clinical-pathological parameter. Thus, from Table 3.8, the risk groups defined by the 12-gene genomic instability signature were found to be closely associated with ER status and tumor grade.

Table 3.8: Association between the 12-gene genomic instability signature and clinic-pathologic parameters in patients (n=469) combined from Sorlie's cohort (n=75), Sotiriou's cohort (n=99), and Van de Vijver's cohort (n=295).

Clinical Parameters	**GS Group**	**aGU Group**	***p*-value**
Age			
≤53	245/316	118/153	$p < 0.98$
>53	71/316	35/153	
Lymph Node			
Status	172/316	77/153	$p< 0.46$
Positive	144/316	76/153	
Negative			
ER Status			
Positive	289/316	58/153	$p < 2.2e-16$
Negative	25/316	95/153	
Unknown	2/316		
Tumor Grade			
I	22/114	3/60	$p< 0.00020$
II	53/114	18/60	
III	38/114	39/60	
Unknown	1/114		

Having studied the association, we were curious to investigate whether the 12-gene genomic instability signature could further stratify patients belonging to certain subgroups into genomic instability-defined prognostic risk groups. This was achieved by first, combining together patients having the same subgroup of clinical parameters (such as lymph node-positive, lymph node-negative, tumor grade II, NIH high risk, and St.Gallen high risk) from the three validation datasets, and then plotting Kaplan-Meier survival curves by using the predictive class obtained from the nearest centroid method. These survival curves depicted the statistical significance of the genomic instability-defined risk groups.

3.5.1. The 12-gene signature is independent of lymph node status in breast cancer prognosis

To investigate whether the 12-gene signature was independent of lymph node status, the three validation cohorts were combined and the lymph node-negative patients and lymph node-positive patients were analyzed separately. For lymph node-positive patients, the GS and the aGU groups had distinct disease-free survival (log-rank tests; *p=0.001*; *n=249*; Figure 3.9A) and distinct overall survival (log-rank test; *p<0.0001*; *n=249*; Figure 3.9B). For lymph node-negative patients, the GS and the aGU groups had distinct disease-free survival (log-rank tests; *p=0.0002*; *n=220*; Figure 3.9C) and distinct overall survival (log-rank test; p<0.0001; n=220; Figure 3.9 D). It was seen that the 12-gene signature quantified breast cancer outcomes in Kaplan-Meier analyses independent of lymph node status in a combination of the three patient cohorts. Thus the 12-gene genomic instability signature could provide additional prognostic information within subgroups defined by lymph node status.

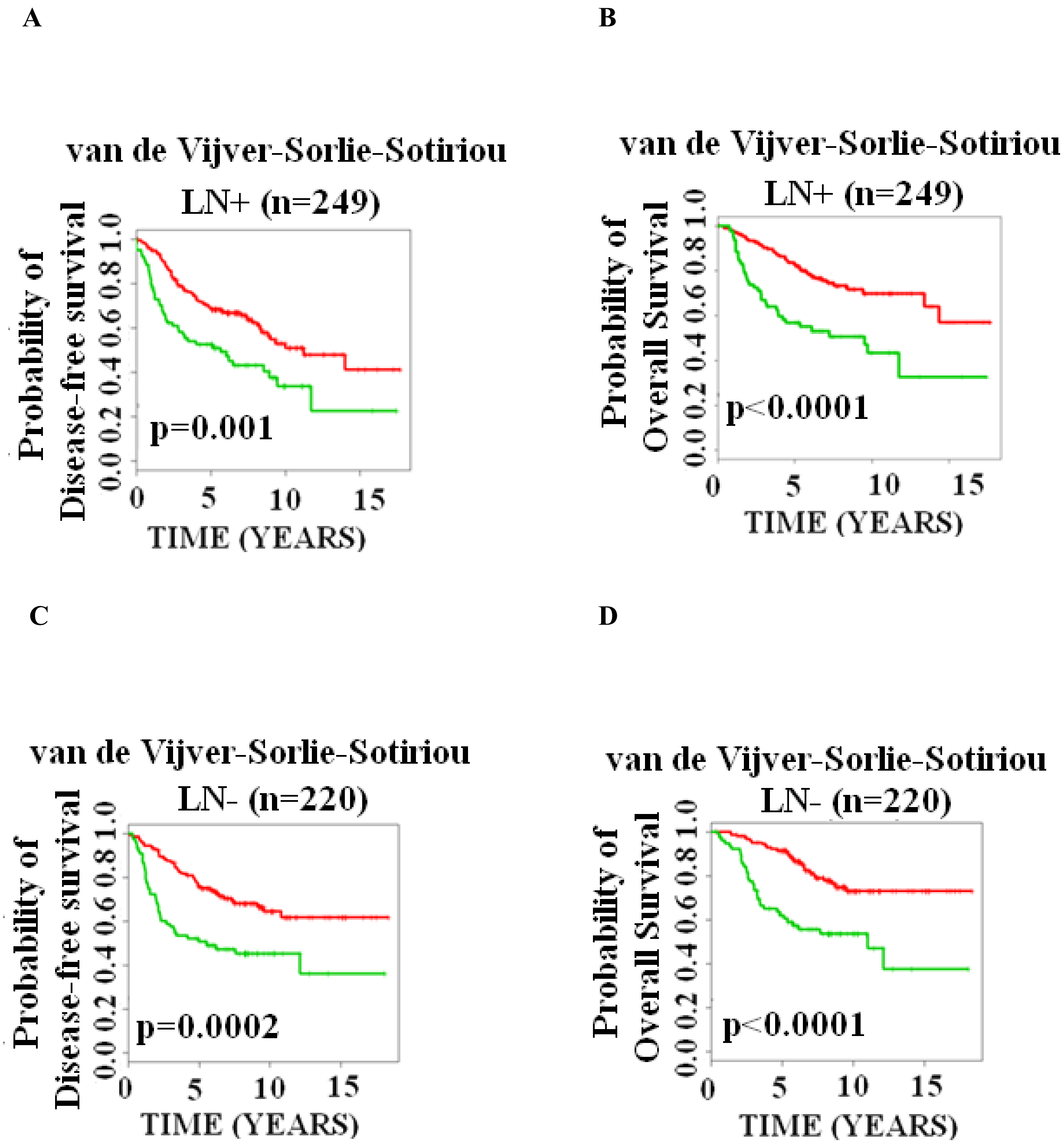

Figure 3.9: The 12-gene signature stratifies patients with lymph node status into subgroups with distinct disease-free survival (A and C) and overall survival (B and D) in Kaplan-Meier analysis. The curves in red represent the genomically stable group and the curves in green represent the genomic instability group.

3.5.2 The 12-gene signature is independent of tumor grade II in breast cancer prognosis

In order to investigate whether the 12-gene signature is independent of tumor grade II, we combined the three external validation cohorts and analyzed all the patients with tumor grade II. The GS and the aGU groups were found to have distinct disease-free survival (log-rank test; *p<0.0001*; *n=172*; Figure 3.10A) and overall survival (log-rank test; *p=0.0001*; *n=172*; Figure 3.10B). It was observed that in grade II tumors, application of the 12-gene signature allowed for improved prognostic classification. The results of this analysis are as shown in the Figure 3.10.

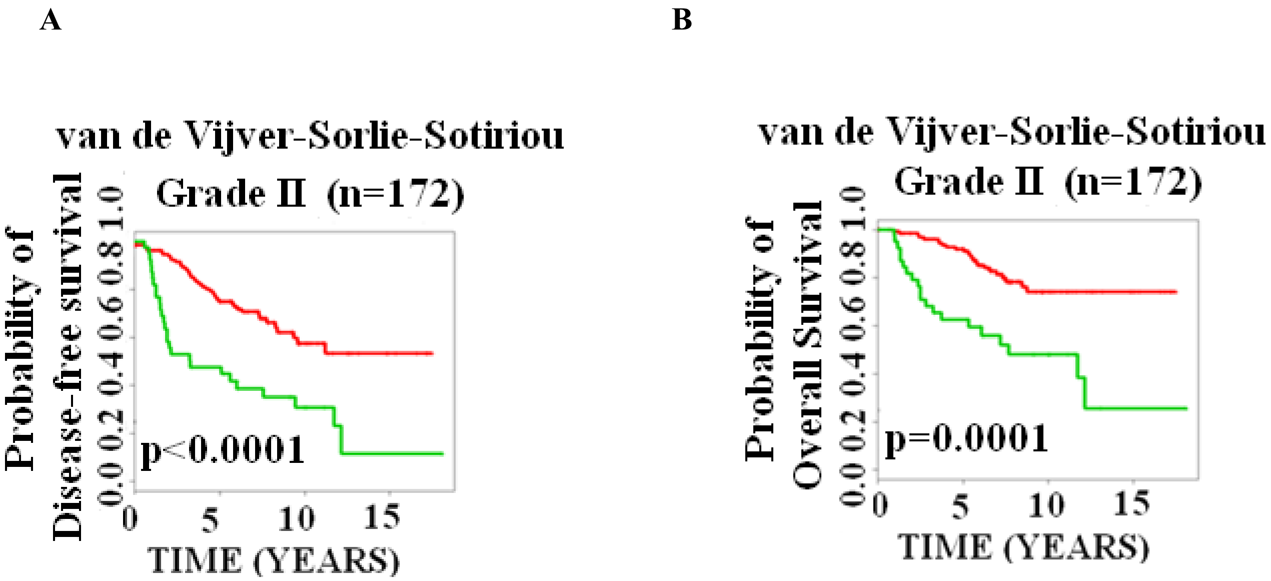

Figure 3.10: The 12-gene signature stratifies patients with tumor grade II into subgroups with distinct disease-free survival (A) and overall survival (B) in Kaplan-Meier analyses. The curves in red represent the genomically stable group and the curves in green represent the genomic instability group.

3.5.3 The 12-gene signature is independent of other predictors of high- risk (NIH Criteria and St.Gallen Criteria)

According to van't Veer et al.(9), a patient was considered as high-risk by NIH criteria, if the tumor size was greater that 1cm, and was considered high-risk by St.Gallen criteria, if one or more of the following conditions were true: estrogen (negative) or tumor size (>2 cm) or tumor grade (Grade II or III) or patient age (< 35 years). This study sought to investigate if the 12-gene genomic instability signature could provide additional prognostic information within the high-risk groups defined by the NIH criteria (42) and the St.Gallen criteria (43). Patients from the three cohorts were combined and the ones who were defined high-risk were analyzed for each criterion.

It was found that, among the high-risk patients defined by the NIH criteria (*n=377*), those with the GS signature had significantly better prognosis than those with the aGU signature for both disease-free survival (log-rank test; *p=0.0001*; *n=377*; Figure 3.11A) and overall survival (log-rank test; *p<0.0001*; *n=377*; Figure 3.11B). Similarly in the case of high risk patients defined by the St.Gallen criteria (*n=439*), those with the GS signature had significantly better prognosis than those with the aGU signature for both disease-free survival (log-rank test; *p<0.0001*; *n=439*; Figure 3.12A) and overall survival (log-rank tests; *p<0.0001*; *n=439*; Figure 3.12B).

It was seen that the 12-gene signature quantified breast cancer outcomes in Kaplan-Meier analyses independent of high-risk groups defined by NIH and St.Gallen criteria, in a combination of the three patient cohorts. Thus the 12-gene genomic instability signature could provide additional prognostic information within high-risk subgroups defined by NIH and St.Gallen criteria.

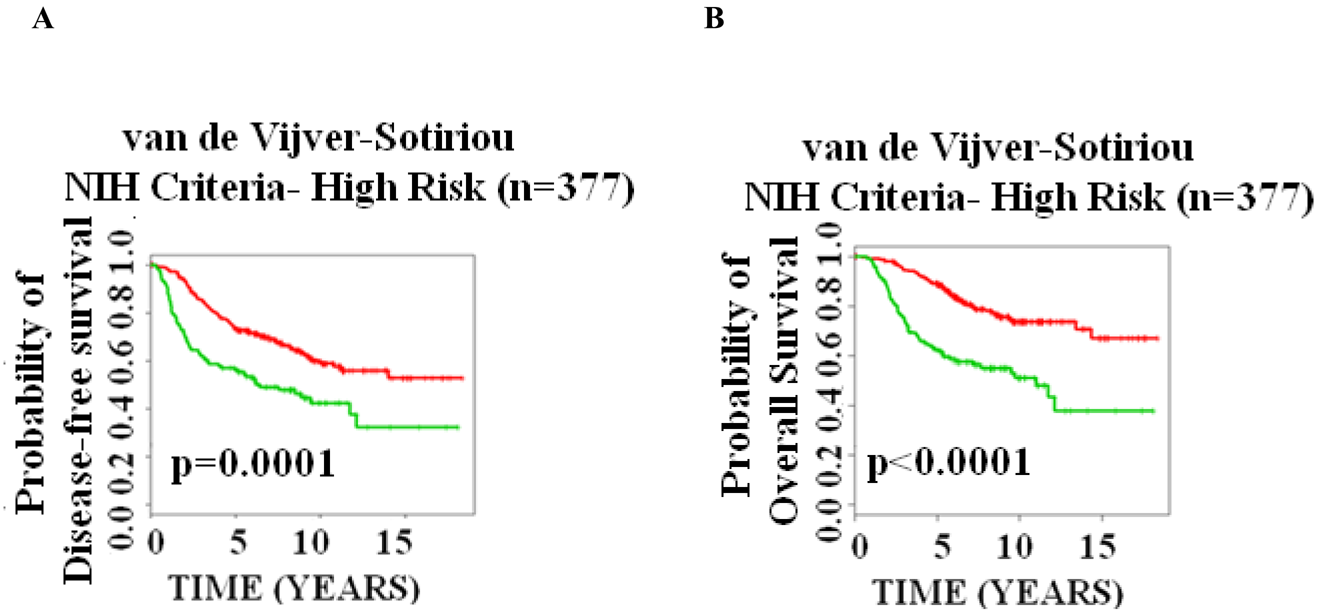

Figure 3.11: The 12-gene signature stratifies high risk patients defined by NIH Criteria into subgroups with distinct disease-free survival (A) and overall survival (B) in Kaplan-Meier analyses. The curves in red represent the genomically stable group and the curves in green represent the genomic instability group.

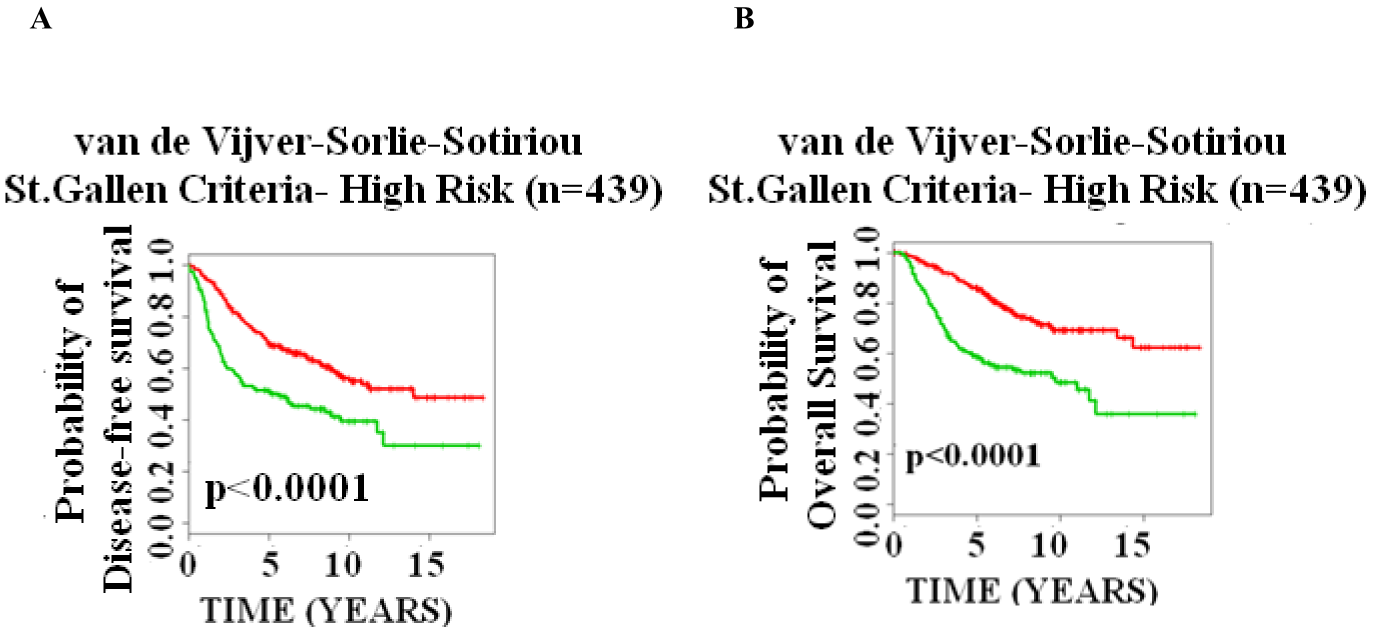

Figure 3.12: The 12-gene signature stratifies high risk patients defined by St. Gallen Criteria into subgroups with distinct disease-free survival (A) and overall survival (B) in Kaplan-Meier analyses. The curves in red represent the genomically stable group and the curves in green represent the genomic instability group.

3.6 Summary

In this chapter, we sought to explore if there is a linkage between the degree of genomic instability, gene expression patterns, and clinical outcomes. This was achieved by first, identifying a 12-gene genomic instability signature from a training dataset that contained 7657 genes measured on 48 breast cancer patients with varying degrees of genomic instability. A combination of different feature selection algorithms was employed to identify the 12-gene genomic instability signature. By using the nearest centroid classification algorithm, the 12-gene genomic instability signature could significantly stratify patients from multiple validation datasets, into prognostic groups. The significance of this stratification was tested by Kaplan-Meier analyses and log-rank tests. It was observed that the gene expression-defined groups had distinct relapse-free and overall survival independent of traditional prognostic factors ($n=469$, $p < 0.05$, log-rank tests).

Thus the degree of genomic instability which is measured by the nuclear DNA content, directly impacts on a breast cancer patient's prognosis and serves as one of the most powerful predictors of clinical outcome independent from established parameters (4;5). Patients with breast carcinomas having a relatively stable genome have considerably longer disease-free survival times compared to the ones having genomically unstable tumors (2;3). Therefore, prognostication based on gene expression signatures could be augmented by the quantitative measurements of nuclear DNA content.

Chapter 4

Validating a population-based signature for breast cancer prognosis

4.1. Introduction

In this study, the importance of selecting patients at high-risk for recurrence for additional chemotherapy was realized at the light of the fact that, breast cancer patients with advanced stages receive chemotherapy, but only half of them benefit from it (11). The two available gene signatures, namely, Oncotype (8) and Mammaprint (9) are based on specific subgroups of breast cancer patients. Specifically, Oncotype was designed for tamoxifin-treated, node-negative, estrogen receptor-positive breast cancer patients, and Mammaprint was used for lymph node-negative women under the age of 61 with a tumor size less than 5 cm. A population-based approach was needed to the molecular prognosis of breast cancer for predicting breast cancer recurrence in broader clinical settings.

Our previous studies (13) have shown that breast cancer recurrence and metastases can be predicted at the individual level, based on a 28-gene recurrence signature that was developed from Sotiriou et al. (34) on a population-based approach. This dataset contained 7650 genes on 99 node-negative and node-positive patient samples. In this study, we validated the 28-gene expression signature on several independent datasets that had different clinical-pathological characteristics and were generated on various DNA microarray platforms. The nearest centroid classification algorithm (10) was employed for stratifying the patients from the validation datasets into different prognostic groups. The association of the gene signature-defined prognostic groups and traditional clinical-pathological factors was estimated in quantifying breast cancer disease-free survival and overall survival. The ability of the 28-gene signature to further stratify the clinical subgroups was investigated.

The significance of this work in the clinical management is that, it could enable physicians to take proper decisions regarding the need for additional chemotherapy for patients (47). They could identify the high-risk patients based on this molecular classification scheme. Moreover, this study also sought to investigate if a common gene set could predict a poor outcome in breast cancer and ovarian cancer (12). According to epidemiological studies (48), breast cancer patients have an increased risk of primary ovarian cancer.

The remainder of the chapter is organized as follows: Section 4.2 discusses the previous work done, related to the development of the 28-gene signature, Section 4.3 elucidates the process of validation of the 28-gene signature in multiple DNA microarrays using the nearest centroid classification method, Section 4.4 illustrates the association of the gene expression-defined risk groups and clinical parameters, Section 4.5 explains the analysis of the 28-gene signature on ovarian cancer, and finally Section 4.6 summarizes the chapter.

4.2 Previous Work-Identification of 28 gene signature

The 28-gene expression signature was previously identified in our lab (13), from Sotiriou et al. (34) which comprised of a gene expression data containing 7650 genes assayed by cDNA microarray on 99 patient samples, 53 of which were node-positive and 46 of node-negative patient samples. The data is publicly available as the supporting information on the PNAS website[17]

[17] http://www.pnas.org/cgi/content/full/100/18/10393

The data was first pre-processed to remove genes that had more than 5 missing values. 559 genes were eliminated in this step, and the remaining missing values were replaced by using the EMV package in software R[18]. The k-nearest-neighbor algorithm (*k=20*) was used to estimate the missing values. There were 7091 genes in the dataset after data pre-processing.

The marker genes were identified by using a combination of random forests employing the VarSelRF package of software R (33), and linear discriminant analysis (LDA) of software SAS[19]. The VarSelRF package in R (33) was used in a series of steps. In the first step, a forest with N trees was built and the features were ranked according to the importance of the variables. In the second step, 20% of the variables that were least important were removed and a new forest was constructed with K trees. This step was repeated till there were two genes left. The gene subset with the smallest OOB error rate was selected. In the experiments, a value of $N = 3000$ and K =1000 were considered, because a large number of trees in the initial forests is likely to produce stable importance measures (19). The "0-Standard Error (0-SE) rule" was observed that identifies the gene subset with the smallest OOB error rate. The 28-gene signature that was obtained as a result of the feature selection is shown in Table 4.1.

[18] http://www.r-project.org

[19] http://www.sas.com/

Table 4.1: List of 28-gene signature

Gene	Spot ID	Clone ID	UniGene Cluster ID
Homo sapiens GT212 mRNA	3912	198917	Hs.463079
TOMM70A	4919	198312	Hs.227253
MCF2	2370	268412	Hs.387262
RAD52 homolog	418	1377154	Hs.552577
MCM2	1881	239799	Hs.477481
C18B11	5984	131988	Hs.173311
SEC13L	6497	757210	Hs.301048
SLC25A5	5182	291660	Hs.522767
PLSCR1	6959	268736	Hs.130759
TXNRD1	7296	789376	Hs.434367
RAD50	2925	261828	Hs.242635
-	6498	46196	
INPPL1	1987	703964	Hs.523875
-	583	501651	Hs.439445
TXNRD1	6736	789376	Hs.434367
PBX2	536	80549	Hs.509545
SSBP1	3434	125183	Hs.490394
HSPCB (heat shock 90kD protein 1, beta)	2403	34396	Hs.448229
PDGFRA	6674	376499	Hs.74615
ACOT4	6555	488202	Hs.49433
DDOST	2416	50666	Hs.523145
Immunoglobulin alpha (1 or 2) heavy chain constant region	2276	182930	Hs.497723
S100P	5593	135221	Hs.2962
FAT	7009	591266	Hs.481371
FGF2	3514	324383	Hs.284244
INSM1	3061	22895	Hs.89584
IRF5	5962	260035	Hs.521181
SMARCD2	2923	741067	Hs.250581
MAP2K2	1652	769579	Hs.465627

4.3 Validation of the 28-gene expression signature in multiple DNA microarrays

The predictive power of the 28 genes was investigated in assessing breast cancer outcomes. We designed a prognostic categorization scheme for DNA microarray datasets that were generated on various platforms. We adopted the nearest centroid classification method for estimating the predictive power of the identified gene signature on other datasets that were used for validation. These datasets contained various DNA microarray platforms such as DNA microarrays, Affymetrix U95, U133A, and U133 plus 2.0. The examined outcomes include relapse-free survival (RFS), metastases-free survival (MFS), disease-free survival (DFS; where a clinical event refers to either a local recurrence or distant metastases of breast cancer), disease-specific survival (DSS; an event refers to death from breast cancer), and overall survival (OS). The previously published datasets used for validation in this experiment were:

1) Bild et al. (38) (PMID: 16273092) - This cohort contained a total of 157 patients; 110 with estrogen receptor (ER)-level 1, and 47 with ER-level 0. The dataset is publically available at the Gene Expression Omnibus database with an accession number GSE3143.

2) Sorlie et al. (36) (PMID: 12829800) - This cohort contained a total of 122 tissue samples of which, 77 carcinomas and 7 nonmalignant tissues were previously published. There were 83 ER-positive patients and 32 ER-negative patients. There were 34 lymph node-negative patients and 46 lymph node-positive patients. The cohort contains 11 patients with Grade I, 49 patients with Grade II, and 53 patients with Grade III. The dataset is publically available at the Gene Expression Omnibus database with an accession number GSE4335.

3) Wang et al. (37) (PMID: 15721472) - This cohort contained 286 lymph node-negative patients of which, 146 were of stage T1, 132 of stage T2, and 8 of stage T3/4. This cohort contained 209

ER-positive and 77 ER-negative patients. There were 165 progesterone receptor (PR)-positive, 111 PR-negative and 10 with unknown PR status. 148 patients had poor grade, 42 had moderate grade, 7 had good grade, and 89 had unknown grade. There were 139 pre-menopausal and 147 post-menopausal patients. The dataset is publically available at the Gene Expression Omnibus database with an accession number GSE2034.

4) Van de Vijver et al. (10) (PMID: 12490681) - There were a total of 295 consecutive patients with primary breast carcinomas; 151 with lymph node-negative disease, and 144 with lymph node-positive disease. The dataset is publically available at the Rosetta Inpharmatics website[20].

5) Miller et al. (39) (PMID: 16141321) - This cohort contained a total of 236 patients; 62 patients with Grade I, 121 with Grade II, 51 with Grade III and 2 patients with unknown grade information. 201 patients were ER-positive and 31 patients were ER-negative. There were 179 PR-positive patients and 57 PR-negative patients. The cohort contained 78 lymph node-positive patients and 149 lymph node-negative patients. The dataset is publically available at the Gene Expression Omnibus database with an accession number GSE3494.

6) Loi et al. (41) (PMID: 17401012) - This cohort contained 137 untreated patients and 277 tamoxifen treated patients. Gene expression profiles of 327 patients were screened on GPL96 Affymetrix Gene Chip Human Genome U133 Array Set HG-U133A platform and 87 patient expression profiles were generated on GPL570 Affymetrix GeneChip Human Genome U133 plus 2.0 Array. The cohort contained 250 lymph node-negative patients and 143 lymph node-positive patients. There were 82 patients with Grade I, 182 patients with Grade II, and 76 patients with Grade III. There were 349 ER-positive patients and 45 ER-negative patients. The

[20] http://www.rii.com/publications/2002/nejm.html

dataset is publically available at the Gene Expression Omnibus database with an accession number GSE6532.

7) Ivshina et al. (40) (PMID: 17079448) - This cohort contained patient and tumor samples of the Uppsala and Singapore sets. The Uppsala set was composed of 249 patients. The Singapore set contained 40 patients. There were 211 ER-positive patients and 34 ER-negative patients. This cohort contained 81 lymph node-positive patients and 159 lymph node-negative patients. The dataset is publically available at the Gene Expression Omnibus database with an accession number GSE4922.

Various classification algorithms were tried in WEKA using a leave-one-out cross validation technique to classify the samples in each of the datasets. The samples in each of the datasets were first stratified into low risk and high risk depending on the survival information and status of the clinical outcome. The criteria that was considered for the stratification were

RFS ≤ 5 years and Status =1 -> High risk

RFS > 5 years and Status = 0 -> Low risk

The classification algorithms in WEKA were used to obtain the classification accuracy. A cross-cohort validation was also performed by considering datasets from the same platform namely, Wang et al.(37), Ivshina et al. (40), and Loi et al. (41). The best classification model was identified in WEKA by performing a 10-fold cross validation on Wang et al. (37). Logistic was found to be the classification algorithm that produced the highest classification accuracy. This model was applied to the testing datasets, Ivshina et al.(40) and Loi et al.(41). It was found that the classification model could not identify any high risk patients in the testing datasets. The results of the classification accuracies generated in the cross-cohort validation analyses are shown in Table 4.2.

Table 4.2: Results of cross-cohort validation using Logistic algorithm in WEKA

Dataset	Sensitivity	Specificity	Overall Accuracy	(Sensitivity+Specificity)/2
Wang	43.16	82.74	68.44	**62.95**
Ivshina(Testing)	0	100	66.5	**50**
Loi(Testing)	0	100	64.86	**50**

Since the datasets were generated on diverse microarray platforms and had incompatible expression profiles, the classification algorithms in WEKA failed to give consistent results across the datasets. A single classification model could not be used across all the datasets. Moreover, the results of these classification algorithms gave an estimate over a specific period of 5 years. The need to analyze the clinical outcomes over the time-course was identified which was possible by Kaplan-Meier survival curves. The results obtained by performing a leave-one-out cross validation using various classification algorithms in WEKA are as shown in Table 4.3.

Table 4.3: Classification analyses using various classification algorithms in WEKA

Datasets	**Van de Vijver**				**Wang**				**Miller**			
Algorithms	Specificity (Low Risk) (%)	Sensitivity (High Risk) (%)	Overall Accuracy	(Specificity + Sensitivity) /2	Specificity (Low Risk) (%)	Sensitivity (High Risk) (%)	Overall Accuracy	(Specificity + Sensitivity) /2	Specificity (Low Risk) (%)	Sensitivity (High Risk) (%)	Overall Accuracy	(Specificity + Sensitivity) /2
Naïve Bayes	77.51	51.72	68.75	**64.62**	63.1	62.11	62.74	**62.61**	92.41	0	74.87	46.2
Neural Network	75.74	41.38	64.06	58.56	64.88	45.26	57.79	55.07	82.28	21.62	70.77	51.95
IBk	76.33	37.93	63.28	57.13	77.98	31.58	61.22	54.78	88.6	16.22	74.87	**52.41**
Random Committee	89.35	29.89	69.14	59.62	89.29	18.95	63.88	54.12	99.37	5.4	81.54	52.39
Hyperpipes	95.86	5.75	65.23	50.81	91.07	10.53	61.98	50.8	99.37	0	80.51	49.68
Random Forest	85.8	36.78	69.14	61.3	85.12	18.95	61.22	52.04	98.10	5.41	80.51	51.76
Decision Table	76.33	32.18	61.33	54.26	99.4	0	63.5	49.7	92.41	0	74.87	46.2
Average	82.42	33.66	65.84	58.04	81.55	26.77	61.76	54.16	93.22	6.95	76.85	50.08

Thus, considering the inconsistent results of the above analyses and the need for a single classification algorithm, the nearest centroid classification method was adopted for evaluating the accuracy 28-gene signature on other datasets used for validation. In the process of validation, the patients in the training dataset from Sotiriou et al. (34) were classified into two subgroups, namely, good-prognosis and poor-prognosis based on their survival information which included relapse-free survival, and status (that indicates if the patient developed metastases or not). A patient was classified into good-prognosis group if the patient survived longer than five years after the primary treatment; otherwise, the patient was classified into poor-prognosis group. The criteria used for the classification is expressed as:

RFS $\geq$ 5 years, Status =0 → good-prognosis

RFS < 5 years, Status =1 → poor-prognosis

The average expression centroids (profiles) of the patients with good-prognosis and poor-prognosis were computed separately in the training dataset from Sotiriou et al. (34). Table 4.4 contains the gene expression centroids obtained for each gene in each of the two groups.

Table 4.4: Gene Expression profiles (centroids) of 28 genes

GENE	Poor prognosis	Good prognosis
C18B11	-0.00243	-0.08241
DDOST	-0.36212	-0.625
FAT	0.331797	0.243982
FGF2	-0.14305	-0.04102
Immunoglobulin alpha (1 or 2) heavy chain constant region	0.823265	1.205533
Homo sapiens GT212 mRNA	0.496081	0.678341
HSPCB (heat shock 90kD protein 1, beta)	-0.59895	-0.77243
IMAGE:46196	-0.38726	-0.53131
ACOT4	0.480262	0.592882
IMAGE:501651	0.149646	0.291806
INPPL1	0.397124	0.541886
INSM1	0.285381	0.367837
IRF5	-0.31068	-0.50189
MAP2K2	-0.06515	0.007757
MCF2	0.1877	0.258102
MCM2	-1.25522	-1.46739
PBX2	-0.17896	-0.2586
PDGFRA	0.10173	0.203139
PLSCR1	0.173373	-0.12427
RAD50	0.182354	0.258133
RAD52 homolog	0.085389	0.189869
S100P	0.2262	-0.55354
SEC13L	-0.84496	-1.28998
SLC25A5	-0.7837	-1.2255
SMARCD2	0.127054	0.236324
SSBP1	-0.48885	-0.68036
TOMM70A	0.232762	0.112048
TXNRD1	-1.07283	-1.41502

Each patient in the validation cohorts was categorized into good-prognosis group or poor-prognosis group based on the Pearson correlation of the patient's gene expression profiles with the average expression profiles of the good-prognosis centroid in the training set. If there were multiple probes for the same annotated gene, the average of the gene expressions for all the probes was computed and used in the correlation analysis. As the validation sets contain DNA microarrays that were generated on heterogeneous platforms, different cut-off values were chosen for patient stratification based on the correlation coefficients with the average good-prognosis centroid. A patient was classified as good-prognosis if the correlation was greater than the corresponding cut-off value; otherwise, this patient was classified as poor- prognosis.

A cut-off value of -0.3 was taken for predicting overall survival (OS) and disease-specific survival (DSS). This cut-off value was applied consistently in patient stratification for three different platforms: Affymetrix HG-133A [Miller et al. (39)], Affymetrix HG-U95 [Bild et al. (38)]. For relapse-free survival and disease-free survival prediction, several cut-off values were chosen for different platforms as follows: A cut-off value of 0.15 was considered for cDNA microarrays [van de Vijver et al.(10) and Sorlie et al.(36)], -0.4 for Affymetrix HG-U133A [Wang et al.(37) , Ivshina et al.(40), and Loi et al.(41)], and -0.5 for Affymetrix U133 Plus 2.0 Array [Loi et al. (41)]. The different cut-offs taken for different platforms and clinical endpoint is as shown below in Table 4.5.

Table 4.5: Table of various cut-offs taken for various platforms

Platform	**cDNA microarray**		**GPL96 Affymetrix HG-U133A**				**GPL91 Affymetrix HG-U95A**	**GPL570 Affymetrix HG-U133 Plus 2.0 Array**
Cut-off value	0.15		-0.4			-0.3	-0.3	-0.5
Datasets	Van de Vijver	Sorlie	Wang	Ivshina	Loi	Miller	Bild	Loi
Clinical End Points	RFS MFS OS	RFS OS	RFS	DFS	RFS	OS DSS	OS	RFS MFS

The significance of this stratification scheme was tested by Kaplan-Meier analyses and log-rank tests. For each validation dataset, the predictive class obtained from the nearest centroid classification was taken and survival curves were plotted using Kaplan-Meier analyses. Kaplan-Meier analyses showed that gene expression-defined groups had distinct relapse-free and overall survival ($p < 0.05$, log-rank tests). The patients belonging to good-prognosis group had higher survival probabilities than those belonging to the poor-prognosis groups.

The results of the analyses for each validation dataset are as shown in the Figures 4.1-4.7.

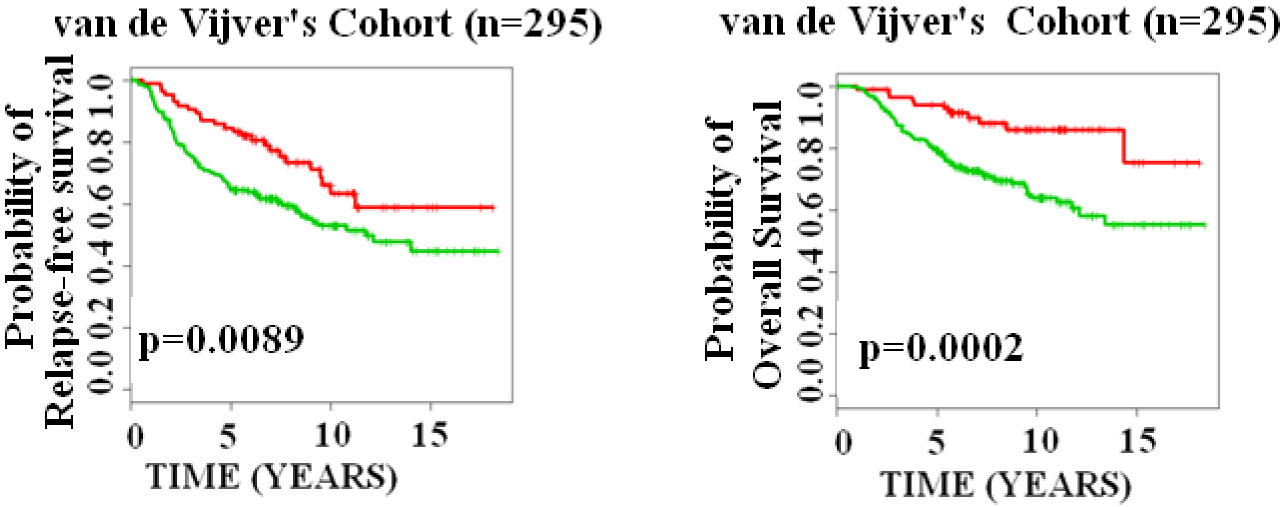

Figure 4.1: The 28-gene signature classifies breast cancer patients from van de Vijver's cohort into prognostic subgroups with distinct relapse-free survival and overall survival in Kaplan-Meier analysis by taking 0.15 as a cut-off for stratifying patients into each subgroup. The cut-off is based on the correlation coefficients with the average good-prognosis centroid in the training set. The curves in red represent the good-prognosis group and the curves in green represent the poor-prognosis group.

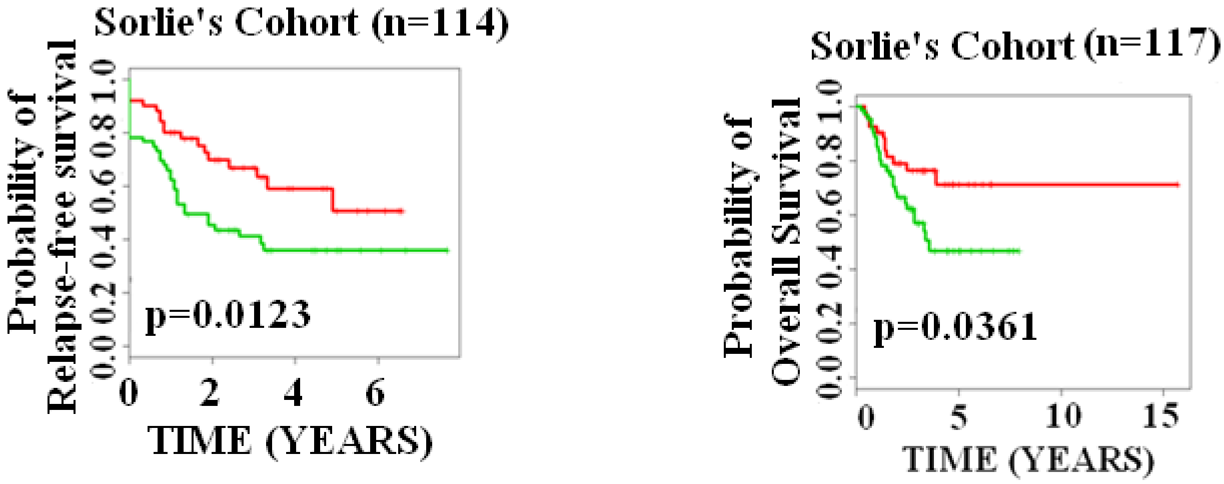

Figure 4.2: The 28-gene signature classifies breast cancer patients from Sorlie's cohort into prognostic subgroups with distinct relapse-free survival and overall survival in Kaplan-Meier analysis by taking 0.15 as a cut-off for stratifying patients into each subgroup. The cut-off is based on the correlation coefficients with the average good-prognosis centroid in the training set. The curves in red represent the good-prognosis group and the curves in green represent the poor-prognosis group.

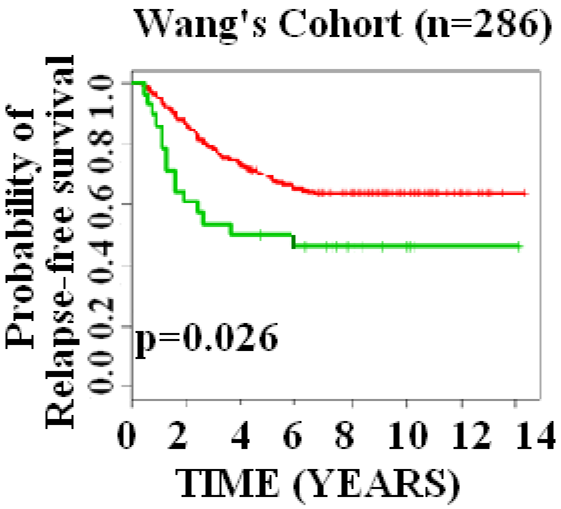

Figure 4.3: The 28-gene signature classifies breast cancer patients from Wang's cohort into prognostic subgroups with distinct relapse-free survival in Kaplan-Meier analysis by taking -0.4 as a cut-off for stratifying patients into each subgroup**.** The cut-off is based on the correlation coefficients with the average good-prognosis centroid in the training set. The curves in red represent the good-prognosis group and the curves in green represent the poor-prognosis group.

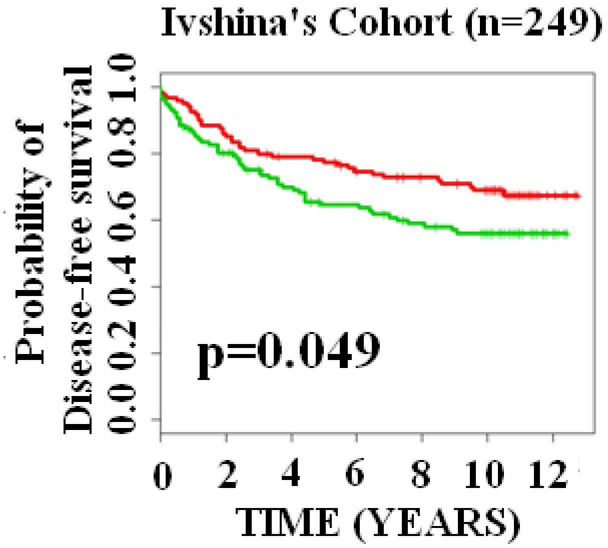

Figure 4.4: The 28-gene signature classifies breast cancer patients from Ivshina's cohort into prognostic subgroups with distinct disease-free survival in Kaplan-Meier analysis by taking -0.4 as a cut-off for stratifying patients into each subgroup. The cut-off is based on the correlation coefficients with the average good-prognosis centroid in the training set. The curves in red represent the good-prognosis group and the curves in green represent the poor-prognosis group.

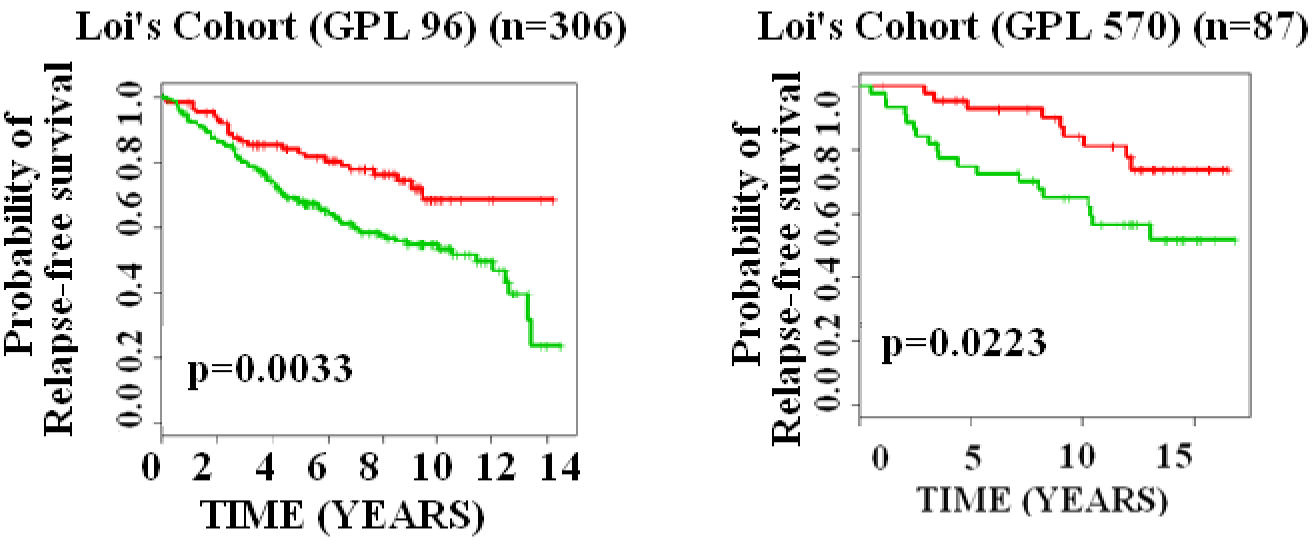

Figure 4.5: The 28-gene signature classifies breast cancer patients from Loi's cohort into prognostic subgroups with distinct relapse-free survival in Kaplan-Meier analysis by taking -0.4 as a cut-off for GPL 96 and -0.5 as a cut-off for GPL 570. The cut-off is based on the correlation coefficients with the average good-prognosis centroid in the training set. The curves in red represent the good-prognosis group and the curves in green represent the poor-prognosis group.

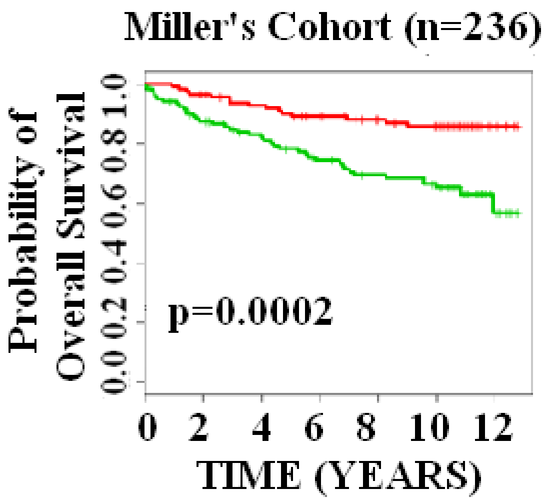

Figure 4.6: The 28-gene signature classifies breast cancer patients from Miller's cohort into prognostic subgroups with distinct overall survival in Kaplan-Meier analysis by taking -0.3 as a cut-off for stratifying patients into each subgroup. The cut-off is based on the correlation coefficients with the average good-prognosis centroid in the training set. The curves in red represent the good-prognosis group and the curves in green represent the poor-prognosis group.

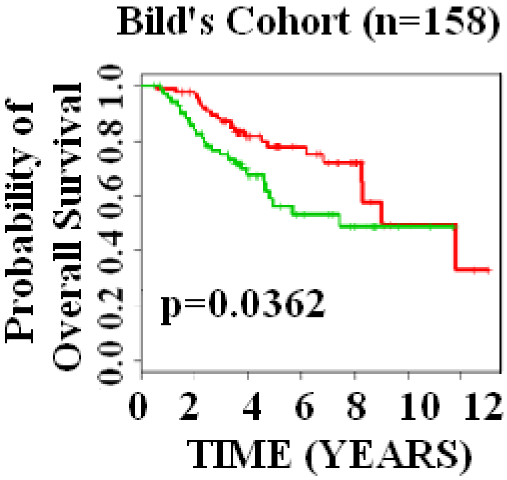

Figure 4.7: The 28-gene signature classifies breast cancer patients from Bild's cohort into prognostic subgroups with distinct overall survival in Kaplan-Meier analysis by taking -0.3 as a cut-off for stratifying patients into each subgroup. The cut-off is based on the correlation coefficients with the average good-prognosis centroid. The curves in red represent the good- prognosis group and the curves in green represent the poor-prognosis group.

4.4 Association of gene expression-defined risk groups and clinical parameters

In this study, the association between prognostic groups and clinical-pathological parameters was determined on the validation datasets by using either Pearson's Chi-squared test or Fisher's exact test (two-sided) between the two parameters. Chi-squared test was performed using the "chisq.test" function in R. Chi-squared test was used if its assumptions were satisfied. Otherwise, Fisher's exact test was applied using the R function "fisher.test". The clinical parameters analyzed in this study were lymph node status, estrogen receptor (ER) status, age, and tumor grade. Patient data with disease-free survival information was combined from Van de Vijver's cohort (*n=295*), Sorlie's cohort (*n=114*), Wang's cohort (*n=286*), Ivshina's cohort (*n=249*), Loi's cohort (*n=393*). Patient data with overall survival information was combined from Van de Vijver's cohort (*n=295*), Sorlie's cohort (*n=117*), Miller's cohort (*n=236*), Bild's cohort (*n=158*).

Table 4.6 and Table 4.7 report the *p* values resulting from the tests for datasets with disease-free survival information and overall survival information respectively. A $p<0.05$ indicates a significant association between the gene expression signature and the corresponding clinical-pathological parameter. It was observed that the prognostic groups defined by the 28-gene expression signature were significantly associated with all the clinical parameters in relapse-free survival prediction ($p<0.05$). In case of overall survival prediction, the prognostic groups were significantly associated with ER status and tumor grade ($p<0.05$) but were not associated with patient age and lymph node status ($p>0.05$).

Table 4.6: Association between the 28-gene signature and clinic pathologic parameters in patients with disease-free survival (n=1337)

Clinical Parameters	**Good Signature Group(n=653)**	**Poor Signature Group(n=684)**	***p*-value**
Age			
<=50 (n=430)	143/653	287/684	0.019
>50 (n=621)	252/653	369/684	
Unknown (n=286)	258/653	28/684	
Lymph Node Status			
Positive(n=444)	163/653	281/684	6.633e-10
Negative(n=870)	484/653	386/684	
Unknown(n=23)	6/653	17/684	
ER Status			
Positive(n=1075)	549/653	526/684	0.001325
Negative(n=248)	98/653	150/684	
Unknown(n=14)	6/653	8/684	
Grade			
I (n=168)	88/653	80/684	9.802e-14
II(n=327)	124/653	203/684	
III(n=245)	41/653	204/684	
Unknown(n=597)	400/653	197/684	

Table 4.7: Association between the 28-gene signature and clinic pathologic parameters in patients with overall survival (n= 806)

Clinical Parameters	**Good Signature Group(n=336)**	**Poor Signature Group(n=470)**	***p*-value**
Age			
<=50 (n=300)	96/336	204/470	0.5515
>50 (n=112)	40/336	72/470	
Unknown (n=394)	200/336	194/470	
Lymph Node Status			
Positive(n=300)	108/336	192/470	
Negative(n=334)	135/336	199/470	0.2887
Unknown(n=172)	93/336	79/470	
ER Status			
Negative(n=179)	57/336	122/470	
Positive(n=618)	274/336	344/470	0.003723
Unknown(n=9)	5/336	4/470	
Grade			
I (n=148)	82/436	66/505	
II(n=270)	117/436	153/505	8.549e-13
III(n=223)	43/436	180/505	
Unknown(n=165)	94/436	71/505	

We sought to investigate whether the 28-gene signature could further refine the subgroups defined by these clinical parameters. Patients with available clinical parameters and outcomes from all the cohorts were considered in this analysis. For analyzing each clinical parameter, patients having the same subgroup of clinical parameters were combined together from all the validation datasets, and Kaplan-Meier survival curves were plotted based on the prognostic categorization obtained from the 28-gene signature. These survival curves depicted the statistical significance of the gene expression-defined prognostic risk groups.

4.4.1 The 28-gene signature is independent of lymph node status in breast cancer prognosis.

In order to investigate whether the 28-gene signature is independent of lypmh node status, patients from all the external validation cohorts were combined and the lymph node-positive and lypmh node-negative patients were analyzed separately. The results of this analysis are shown in Figure 4.8. For lymph node-positive patients, the prognostic groups had distinct disease-free survival (log-rank test; *p<0.0001*; *n=444*; Figure 4.8A) and distinct overall survival (log-rank test; *p=0.0008*; *n=300*; Figure 4.8B). For lymph node-negative patients, the prognostic groups had distinct disease-free survival (log-rank test; *p=0.0029*; *n=870*; Figure 4.8C) and distinct overall survival (log-rank test; *p=0.0001*; *n=334*; Figure 4.8D). It was seen that the 28-gene signature quantified breast cancer outcomes in Kaplan-Meier analyses independent of lymph node status in the combination of different patient cohorts.

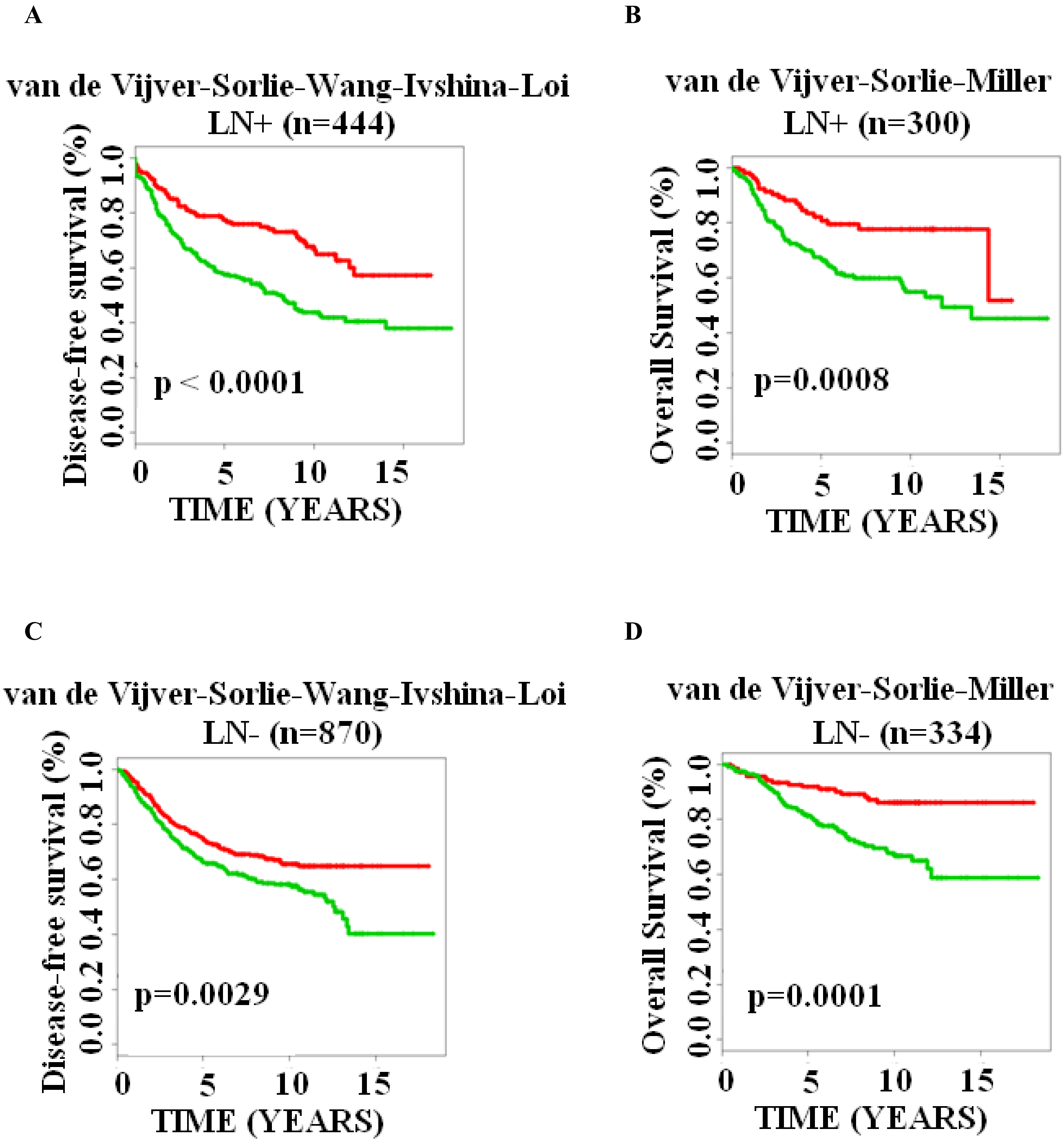

Figure 4.8: The 28-gene signature stratified subgroups defined by lymph node status in predicting breast cancer disease-free survival (A and C) and overall survival (B and D) using Kaplan-Meier analysis. The curves in red represent the good-prognosis group and the curves in green represent the poor-prognosis group.

4.4.2 The 28-gene signature is independent of estrogen receptor (ER) status in breast cancer prognosis

In order to investigate whether the 28-gene signature is independent of ER status, the patients from all the external validation cohorts were combined and the ER-positive and ER-negative patients were analyzed separately. The results of this analysis is shown in Figure 4.9. For ER-positive patients, the prognostic groups had distinct disease-free survival (log-rank test; *$p<0.0001$*; *$n=1075$*; Figure 4.9A) and overall survival (log-rank test; *$p<0.0001$*; *$n=618$*; Figure 4.9B). For ER-negative patients, the prognostic groups had distinct disease-free survival (log rank test; *$p=0.0062$*; *$n=248$*; Figure 4.9C) and overall survival (log-rank test; *$p=0.06$*; *$n=179$*; Figure 4.9D). It was seen that the 28-gene signature quantified breast cancer outcomes in Kaplan-Meier analyses independent of ER status in the combination of different patient cohorts.

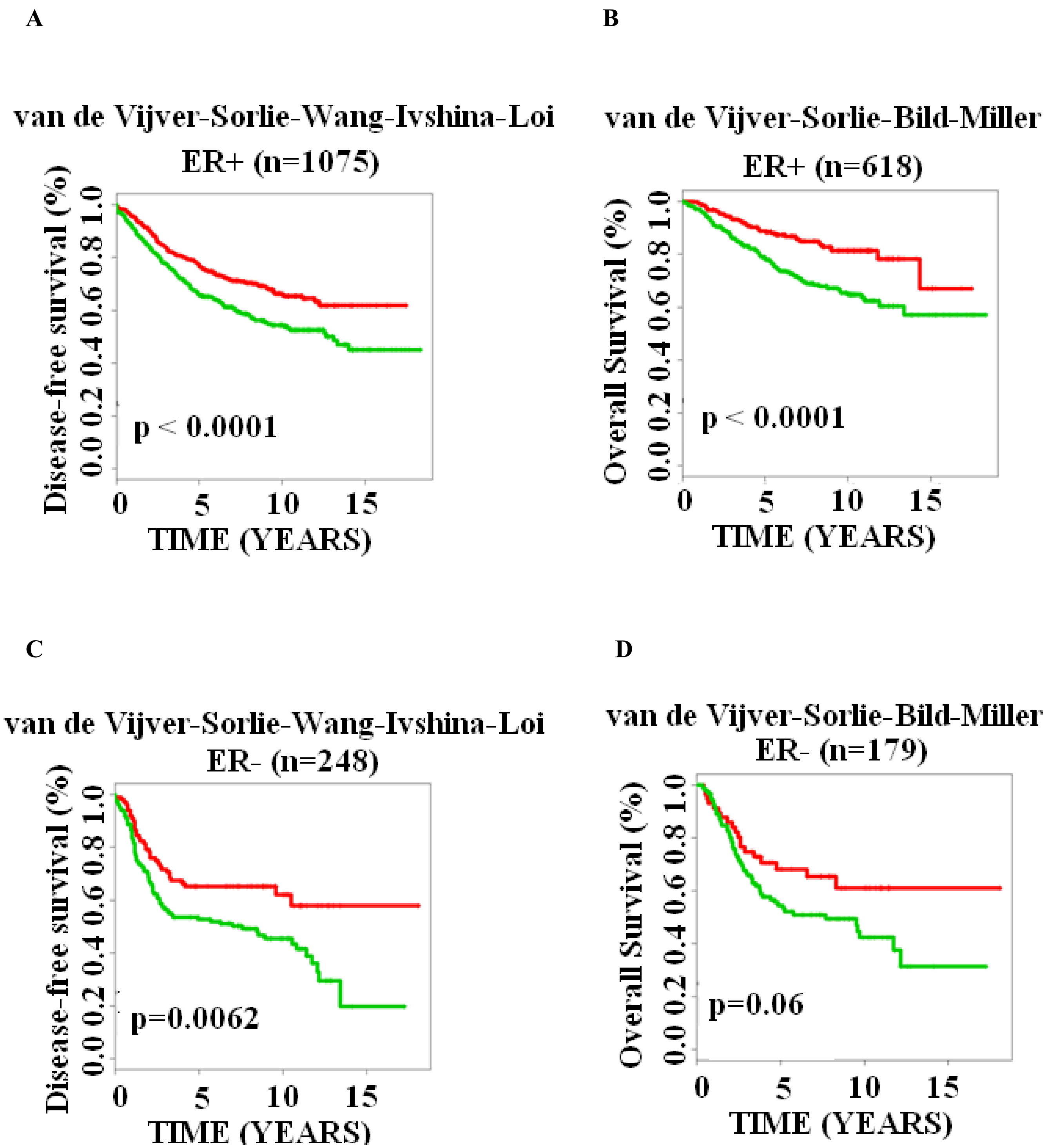

Figure 4.9: The 28-gene signature stratified subgroups defined by ER status in predicting breast cancer disease-free survival (A and C) and overall survival (B and D) using Kaplan-Meier analysis. The curves in red represent the good-prognosis group and the curves in green represent the poor-prognosis group.

4.4.3 The 28-gene signature is independent of tumor grade II in breast cancer prognosis

In order to investigate whether the 28-gene signature could further stratify patients with tumor grade II, the patients from all the external validation cohorts were combined and the patients having tumor grade II were analyzed. The results of this analysis is shown below in Figure 4.10. It was found that the prognostic groups had distinct disease-free survival (log-rank test; *p=0.0197*; *n=327*; Figure 4.10A) and overall survival (log-rank test; *p=0.0024*; *n=270*; Figure 4.10B). It was seen that the 28-gene signature quantified breast cancer outcomes in Kaplan-Meier analyses independent of tumor grade II in the combination of different patient cohorts.

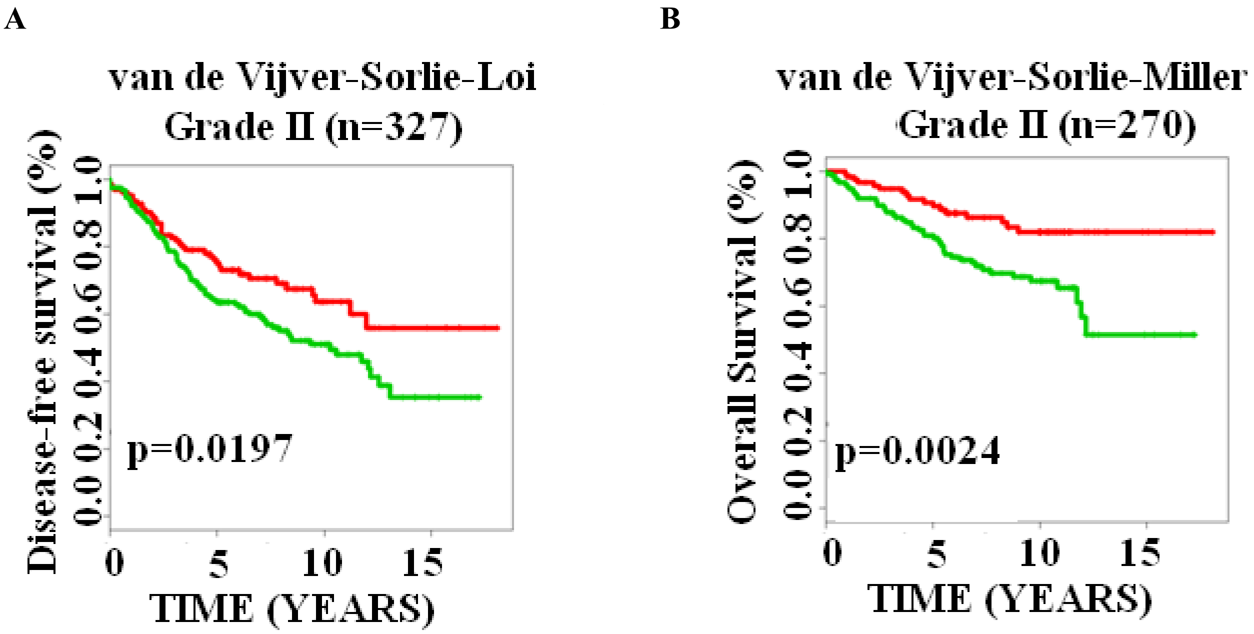

Figure 4.10: The 28-gene signature stratified subgroups defined by Tumor grade II in predicting breast cancer disease-free survival (A) and overall survival (B) using Kaplan-Meier analysis. The curves in red represent the good-prognosis group and the curves in green represent the poor-prognosis group.

4.4.4 Performance of the 28-gene signature on all combined patient cohorts

The patients having disease-free survival information and those having overall survival information from all the cohorts were combined and a survival plot was plotted using Kaplan-Meier analysis for patients with the groups obtained in the correlation analysis. It was seen that the 28-gene signature could stratify the patients into the two subgroups with distinct disease-free survival (log-rank test; *p<0.0001*; *n=1337*; Figure 4.11A) and distinct overall survival (log-rank test; *p<0.0001*; *n=806*; Figure 4.11B) with Kaplan-Meier analysis. The patients belonging to good-prognosis group had higher survival probabilities than those belonging to the poor-prognosis groups. These results confirm that the 28-gene signature is applicable to prognostic categorization for the clinical management of breast cancer based on the expression profiles generated on diverse DNA microarray platforms.

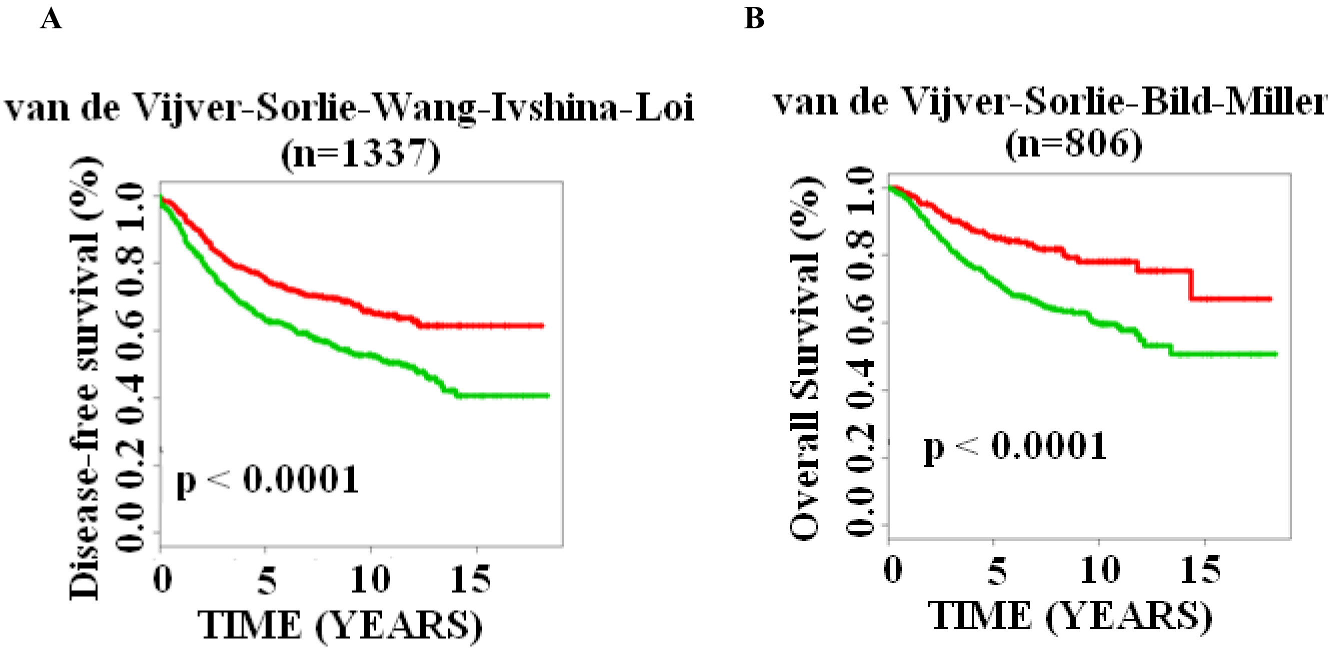

Figure 4.11: The 28-gene signature stratifies patients into subgroups with distinct disease-free survival (A) and overall survival (B) in Kaplan-Meier analysis (*p <0.001*, log-rank test). The curves in red represent the good-prognosis group and the curves in green represent the poor-prognosis group.

4.5 Analysis of the 28-gene signature on Ovarian Cancer

This study sought to explore whether the 28-gene signature revealed common molecular features affecting breast and ovarian cancer. The ovarian cancer dataset from Bild et al. (38), that contained 124 patients was taken and the signature genes were identified in the dataset using Matchminer (46). To avoid over-fitting in the validation, the dataset was randomly split into training and testing sets in the ratio of 2:1. The gene expression data contained in the training set was fitted in a Cox proportional hazard model, independent of traditional clinical-pathological parameters. The survival risk score for each patient in the training set was generated and the median of these scores was determined. This value (risk score: 0.3011433) was used as a cut-off to stratify patients in the training set and testing set into good-prognosis (low-risk) and poor-prognosis (high-risk) groups. A patient with a risk score higher than the cut-off was classified into poor-prognosis (high-risk); whereas a patient with a risk score lower than the cut-off was classified into good-prognosis (low-risk). The significance of this stratification scheme was tested by Kaplan-Meier analysis and log-rank test.

The result of this analysis is shown in figure 4.12. This model generated significant patient stratification (log-rank test; $p < 0.0001$; Kaplan-Meier analyses; $n=82$; Figure 4.12A) for ovarian cancers in the training set. The prognostic model and the cut-off value identified from the training set were applied to the test set which resulted in significant patient stratification (log-rank test; $p=0.0075$; Kaplan-Meier analyses; $n=42$; Figure 4.12B). Thus the 28-gene signature reflected common biological processes involved in breast cancer and ovarian cancer metastases and relapse. The coefficients, z-score, and the p-value of the variables (genes), obtained from the Cox proportional hazard model are as listed in Table 4.8. Signature genes with multiple probes were included in the model.

Table 4.8: The coefficients in the Cox proportional hazard model

Gene / Clone ID	Affymertrix ID	Coef	exp(coef)	se(coef)	z-score	p-value
C18B11	221940_at	-3.71E-03	0.996	0.003764	-0.9853	3.20E-01
DDOST	208674_x_at	-3.09E-03	0.997	0.00096	-3.2159	1.30E-03
DDOST	208675_s_at	1.73E-03	1.002	0.000459	3.7655	1.70E-04
FAT	201579_at	1.36E-03	1.001	0.00069	1.9761	4.80E-02
FGF2	204421_s_at	7.03E-03	1.007	0.010632	0.661	5.10E-01
FGF2	204422_s_at	-1.57E-02	0.984	0.011846	-1.3219	1.90E-01
IMAGE.182930	211868_x_at	5.67E-04	1.001	0.003814	0.1487	8.80E-01
IMAGE.182930	215118_s_at	-1.78E-04	1	0.003685	-0.0484	9.60E-01
IMAGE.182930	216318_at	1.39E-02	1.014	0.01185	1.1696	2.40E-01
IMAGE.182930	216541_x_at	1.05E-02	1.011	0.005294	1.9797	4.80E-02
IMAGE.182930	216542_x_at	3.24E-03	1.003	0.001631	1.9872	4.70E-02
IMAGE.182930	216557_x_at	-8.85E-03	0.991	0.00317	-2.7927	5.20E-03
IMAGE.182930	217022_s_at	-1.79E-04	1	0.000185	-0.9654	3.30E-01
IMAGE.182930	211636_at	1.22E-01	1.13	0.053242	2.2986	2.20E-02
IMAGE.182930	214916_x_at	2.65E-03	1.003	0.001172	2.26	2.40E-02
IMAGE.198917	212697_at	-1.81E-03	0.998	0.001032	-1.7522	8.00E-02
INPPL1	201598_s_at	5.61E-04	1.001	0.0008	0.7008	4.80E-01
INSM1	206502_s_at	2.68E-03	1.003	0.00078	3.4405	5.80E-04
IRF5	205468_s_at	-1.43E-02	0.986	0.015558	-0.9184	3.60E-01
MAP2K2	213487_at	4.60E-02	1.047	0.040941	1.1226	2.60E-01
MAP2K2	213490_s_at	-1.02E-03	0.999	0.003071	-0.3332	7.40E-01
MCF2	208017_s_at	-2.28E-02	0.977	0.009433	-2.4179	1.60E-02
MCF2	217004_s_at	-1.02E-02	0.99	0.008655	-1.1837	2.40E-01
MCM2	202107_s_at	-2.18E-03	0.998	0.000619	-3.5249	4.20E-04
PBX2	202875_s_at	-1.13E-02	0.989	0.003333	-3.388	7.00E-04
PBX2	202876_s_at	2.58E-04	1	0.000756	0.3411	7.30E-01
PBX2	211097_s_at	2.73E-02	1.028	0.008733	3.1296	1.80E-03
PDGFRA	215305_at	-4.87E-02	0.952	0.020093	-2.4255	1.50E-02
PLSCR1	202430_s_at	1.50E-03	1.002	0.001342	1.1204	2.60E-01
PLSCR1	202446_s_at	-6.67E-05	1	0.000381	-0.1749	8.60E-01
RAD50	208393_s_at	-1.02E-02	0.99	0.002219	-4.5823	4.60E-06
RAD50	209349_at	1.51E-02	1.015	0.006131	2.4621	1.40E-02
RAD52	210630_s_at	1.23E-02	1.012	0.005932	2.0698	3.80E-02
RAD52	211904_x_at	-3.45E-02	0.966	0.013025	-2.6465	8.10E-03
S100P	204351_at	-6.75E-04	0.999	0.001109	-0.6086	5.40E-01
SEC13L	221931_s_at	1.06E-03	1.001	0.001731	0.61	5.40E-01
SLC25A5	200657_at	-1.31E-04	1	0.00015	-0.8735	3.80E-01
SMARCD2	201827_at	1.82E-03	1.002	0.000866	2.1042	3.50E-02
SSBP1	202591_s_at	1.82E-03	1.002	0.000599	3.0458	2.30E-03
TOMM70A	201512_s_at	-7.12E-04	0.999	0.001645	-0.4326	6.70E-01
TXNRD1	201266_at	9.83E-05	1	0.000636	0.1546	8.80E-01

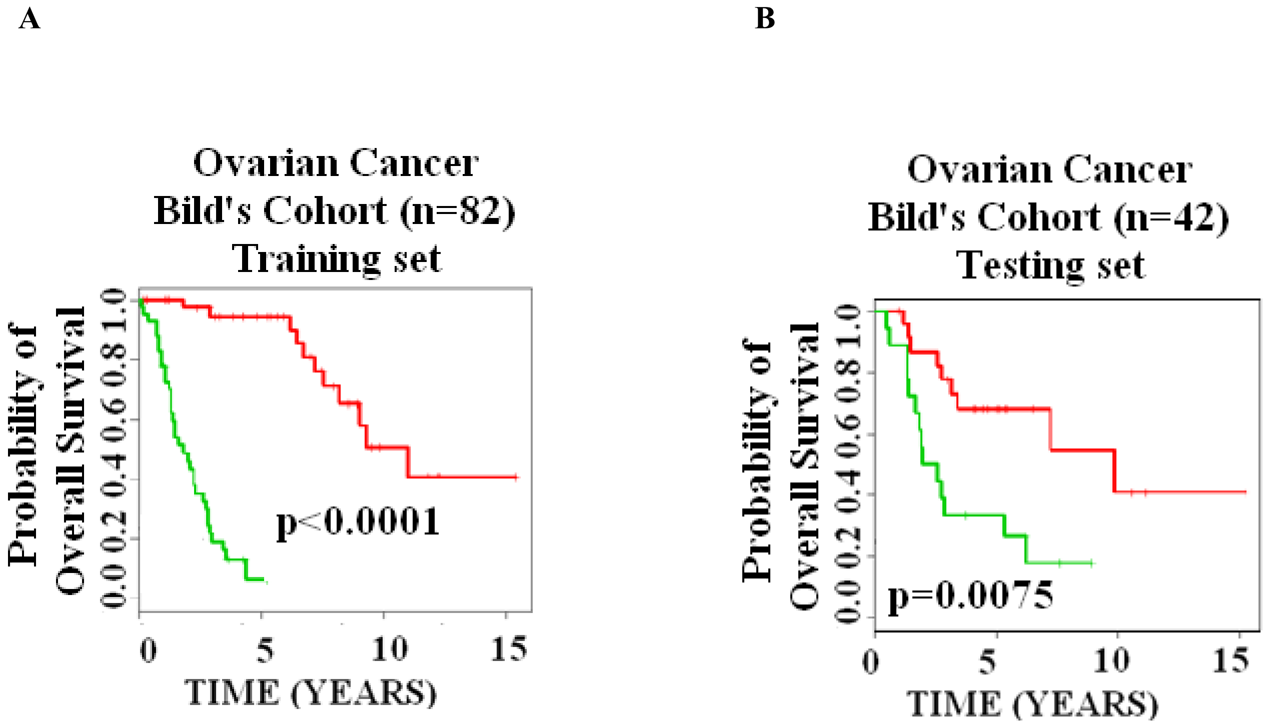

Figure 4.12: The 28-gene signature stratifies patients from the training set (A) and testing set (B) of ovarian cancer dataset Bild et al. into subgroups with overall survival in Kaplan-Meier analysis (*p≤0.0075*, log-rank test). The median of the scores (0.301) generated by fitting the Cox proportional hazard model on the training set was taken as the cut-off. The curves in red represent the low-risk and the curves in green represent high-risk.

4.6 Summary

In this chapter we described how we used the population-based 28-gene expression signature to predict a poor outcome in breast cancer and ovarian cancer. The 28-gene signature was identified in our previous study (13) using the dataset from Sotiriou et al. (34), which contained 7,650 genes assayed by cDNA microarrays on 99 patient samples. The 28-gene signature was validated on multiple published datasets that were generated on different microarray platforms. The nearest centroid classification algorithm (NCC) was used to stratify patients into gene signature-defined prognostic groups by considering different cut-off values for different microarray platforms and clinical endpoints.

The 28-gene expression signature generated significant patient stratification for breast cancer patients in both disease-free survival prediction (*$p < 0.0001$*; log-rank test; *$n=1337$*) and overall survival prediction (*$p < 0.0001$*; log-rank test; *$n=806$*) in Kaplan-Meier analyses. The 28-gene expression-defined prognostic risk groups had distinct clinical outcomes (log-rank test; *$p < 0.05$*; Kaplan-Meier analyses) within each clinical-pathological factor-defined subgroup and were significant in providing additional prognostic information within each of the subgroups (such as lymph node-negative, lymph node-positive, ER-, ER+, and tumor grade II). The signature also generated significant prognostic categorization in ovarian cancers in both training set (log-rank tests; $p< 0.0001$; $n=82$) and test set (log-rank tests; *$p=0.0075$*; *$n=42$*) in Kaplan-Meier analyses.

One of the challenges faced in this study was to design a uniform prognostic mapping scheme for the data from all the studied cohorts. Since the datasets used for validation contained data generated on diverse DNA microarray platforms and clinical end-points, a single cut-off value could not be obtained for stratification of patients across all the datasets. This problem was solved by selecting different cut-off values for different platforms and endpoints. It was observed that the cut-off values identified, were consistently validated in multiple patient cohorts, except for one cut-off value that was selected for predicting relapse-free survival prediction for Loi et al. (41).

This study confirmed the clinical applicability of the population-based 28-gene signature in predicting recurrence in breast cancer and ovarian cancer based on the expression profiles generated on diverse DNA microarray platforms. This is significant in the clinical management of breast cancer, as this molecular classification scheme would help the physicians to take proper decisions related to the risk of the patients to chemotherapy or related treatments.

Chapter 5

Conclusions

The technology of using gene expression as biomarkers for predicting the recurrence of breast cancer provides the potential to refine breast cancer prognosis. Breast cancer patients with the same disease stage may have remarkably different clinical outcome and treatment response. There is a need to develop novel bioinformatic models for biomarker identification.

In this study, the degree of genomic instability was integrated with gene expression patterns and a combination of several feature selection algorithms was used to identify the genomic instability gene signature. A population-based gene expression signature was used to predict breast and ovarian cancer outcomes. A prognostic patient-categorization scheme was designed on the basis of the transcriptional profiles generated on various microarray platforms. Since the datasets contained gene expression data that were generated on various microarray platforms, the cross-validation techniques in WEKA did not give us consistent observations in all the datasets. An innovative method was adopted for validation, namely, nearest centroid classification method (NCC), for classifying unknown samples in an effort to validate the performance of the identified gene signature on numerous datasets. The NCC algorithm is efficient and robust with respect to irrelevant or novel attributes (14). It was a challenging task to design a uniform prognostic mapping scheme for the data from all the studied cohorts. Since the datasets used for validation contained data generated on diverse DNA microarray platforms and clinical end-points, a single cut-off value could not be obtained for stratification of patients across all the datasets. This problem was solved by selecting different cut-off values for different platforms and endpoints. The significance of our research is that, both the gene signatures that were identified, namely, the 28-gene population-based signature as well as the 12-gene genomic instability signature, could be used to classify a new breast cancer patient into different prognostic risk groups.

The first part of our study suggests that, prognostication based on gene expression signatures is significant in the clinical management of breast cancer and could be augmented by quantitative measurement of nuclear DNA content. The second part of our study confirmed the practical applicability of the population-based gene signature in predicting recurrence in breast and ovarian cancer. The results of our study indicate that, stratification of patients into different subgroups on the basis of the prognosis profile, may be a useful means of guiding therapy in patients with breast cancer.

Previous studies declare that the performance of a gene signature could be enhanced by combining it with other gene signatures (49). In the future analysis, we will explore whether the prediction accuracy of our gene signatures could be improved by integrating them with other gene signatures. Moreover, the identified gene signatures could be tested in other epithelial cancer types in addition to breast and ovarian cancer. Thus, gene-expression profiling opens up a new era in diagnosis, prognosis and treatment and helps to understand clearly, many of the pathogenesis processes involved in the disease (50).

Reference List

(1) Marchionni L, Wilson RF, Marinopoulos SS, Wolff AC, Parmigiani G, Bass EB, et al. Impact of gene expression profiling tests on breast cancer outcomes. Evid Rep Technol Assess (Full Rep) 2007 Dec;(160):1-105.

(2) Kronenwett U, Huwendiek S, Ostring C, Portwood N, Roblick UJ, Pawitan Y, et al. Improved grading of breast adenocarcinomas based on genomic instability. Cancer Res 2004 Feb 1;64(3):904-9.

(3) Kronenwett U, Ploner A, Zetterberg A, Bergh J, Hall P, Auer G, et al. Genomic instability and prognosis in breast carcinomas. Cancer Epidemiol Biomarkers Prev 2006 Sep;15(9):1630-5.

(4) Auer GU, Caspersson TO, Wallgren AS. DNA content and survival in mammary carcinoma. Anal Quant Cytol 1980 Sep;2(3):161-5.

(5) Auer G, Eriksson E, Azavedo E, Caspersson T, Wallgren A. Prognostic significance of nuclear DNA content in mammary adenocarcinomas in humans. Cancer Res 1984 Jan;44(1):394-6.

(6) Fallenius AG, Franzen SA, Auer GU. Predictive value of nuclear DNA content in breast cancer in relation to clinical and morphologic factors. A retrospective study of 227 consecutive cases. Cancer 1988 Aug 1;62(3):521-30.

(7) Fallenius AG, Auer GU, Carstensen JM. Prognostic significance of DNA measurements in 409 consecutive breast cancer patients. Cancer 1988 Jul 15;62(2):331-41.

(8) Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004 Dec 30;351(27):2817-26.

(9) van't Veer LJ, Dai HY, van de Vijver MJ, He YDD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002 Jan 31;415(6871):530-6.

(10) van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AAM, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine 2002 Dec 19;347(25):1999-2009.

(11) Sjostrom J. Predictive factors for response to chemotherapy in advanced breast cancer. Acta Oncol 2002;41(4):334-45.

(12) Wooster R, Weber BL. Breast and ovarian cancer. N Engl J Med 2003 Jun 5;348(23):2339-47.

(13) Ma Y, Qian Y, Wei L, Abraham J, Shi X, Castranova V, et al. Population-based molecular prognosis of breast cancer by transcriptional profiling. Clin Cancer Res 2007 Apr 1;13(7):2014-22.

(14) Joseph A.Cruz, David S.Wishart. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics 2006;59-78.

(15) Li H, Zhang K, Jiang T. Robust and accurate cancer classification with gene expression profiling. Proc IEEE Comput Syst Bioinform Conf 2005;310-21.

(16) Imola K.Fodor. A survey of dimension reduction techniques. 5-9-2002.

(17) Tom Mitchell. Machine Learning. [Edition 2, Chapter 1]. 2005.

(18) George H John, Pat Langley. Estimating Continuous Distributions in Bayesian Classiers. Eleventh Annual Conference on Uncertainty in Artificial Intelligence , 338-345. 1995. San Francisco, Morgan Kaufmann Publishers.

(19) Diaz-Uriarte R, Alvarez dA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7:3.

(20) Ma Y, Ding Z, Qian Y, Shi X, Castranova V, Harner EJ, et al. Predicting cancer drug response by proteomic profiling. Clin Cancer Res 2006 Aug 1;12(15):4583-9.

(21) Milde M.S.Lira, Ronaldo R.B.de Aquino, Aida A.Ferreira, Manoel A.Carvalho Jr, Otoni NóbregaNeto, Gabriela S.M.Santos. Combining Multiple Artificial Neural Networks Using Random Committee to Decide upon Electrical Disturbance Classification. 2007 p. 2863-8.

(22) Kenji Kira, Larry A.Rendell. A practical approach to feature selection. [Proceedings of the ninth international workshop on Machine learning], 249-256. 1992. San Francisco, CA, USA, Morgan Kaufmann Publishers Inc.

(23) Jun Yang, Yue-Peng Li. Orthogonal Relief Algorithm for Feature Selection. Lecture Notes in Computer Science 4113/2006, 227-234. 2006. Springer Berlin / Heidelberg.

(24) Mark A.Hall, Georey Holmes. Benchmarking attribute selection techniques for discrete class data mining. IEEE 15[6], 1437-1447. 2003.

(25) Christoph F.Eick, Alain Rouhana, A.Bagherjeiran, R.Vilalta. Using clustering to learn distance functions for supervised similarity assessment. Machine Learning and Data

Mining in Pattern Recognition 3587/2005, 120-131. 2005. Springer Berlin / Heidelberg.

(26) Levner I. Feature selection and nearest centroid classification for protein mass spectrometry. BMC Bioinformatics 2005;6:68.

(27) Marc Strickert, Udo Seiffert. Correlation-based Data Representation. Similarity-based Clustering and its Application to Medicine and Biology , 1-16. 2007.

(28) Stephen J.Walters. What is a Cox model. What is .....? series 1[1]. 2007. Hayward Group plc.

(29) Utley M, Gallivan S, Young A, Cox N, Davies P, Dixey J, et al. Potential bias in Kaplan-Meier survival analysis applied to rheumatology drug studies. Rheumatology (Oxford) 2000 Jan;39(1):1-2.

(30) Mary Jo Gillespie, Lloyd Fisher. Confidence Bands for the Kaplan-Meier Survival Curve Estimate. Annals of Statistics 2008 Aug;36(4).

(31) J Martin Bland, Douglas G Altman. The logrank test. BMJ 328. 2004.

(32) D.R.Cox. Regression Models and Life-Tables. Journal of the Royal Statistical Society 1972;Vol. 34(No. 2):187-220.

(33) Brian S.Everitt, Torsten Hothorn. A Handbook of Statistical Analyses Using R. 2006. Chapman & Hall/CRC Press Binding.

(34) Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proceedings of the National Academy of Sciences of the United States of America 2003 Sep 2;100(18):10393-8.

(35) Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001 Sep 11;98(19):10869-74.

(36) Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences of the United States of America 2003 Jul 8;100(14):8418-23.

(37) Wang YX, Klijn JGM, Zhang Y, Sieuwerts A, Look MP, Yang F, et al. Gene-expression pro-files to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005 Feb 19;365(9460):671-9.

(38) Bild AH, Yao G, Chang JT, Wang QL, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006 Jan 19;439(7074):353-7.

(39) Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proceedings of the National Academy of Sciences of the United States of America 2005 Sep 20;102(38):13550-5.

(40) Ivshina AV, George J, Senko O, Mow B, Putti TC, Smeds J, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Research 2006 Nov 1;66(21):10292-301.

(41) Loi S, Haibe-Kains B, Desmedt C, Lallemand F, Tutt AM, Gillet C, et al. Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. Journal of Clinical Oncology 2007 Apr 1;25(10):1239-46.

(42) Eifel P, Axelson JA, Costa J, Crowley J, Curran WJ, Jr., et al. National Institutes of Health Consensus Development Conference statement: adjuvant therapy for breast cancer, November 1-3, 2000. J Natl Cancer Inst Monogr 2001;(30):5-15.

(43) Goldhirsch A, Glick JH, Gelber RD, Senn HJ. Meeting highlights: International Consensus Panel on the Treatment of Primary Breast Cancer. J Natl Cancer Inst 1998 Nov 4;90(21):1601-8.

(44) Ian H.Witten, Eibe Frank. Data Mining:Practical Machine Learning tools and techniques. 2nd Edition ed. Morgan Kaufmann; 2005.

(45) Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace AJ, Jr., Kohn KW, et al. An information-intensive approach to the molecular pharmacology of cancer. Science 1997 Jan 17;275(5298):343-9.

(46) Bussey KJ, Kane D, Sunshine M, Narasimhan S, Nishizuka S, Reinhold WC, et al. MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome Biology 2003;4(4).

(47) Dalton WS, Friend SH. Cancer biomarkers--an invitation to the table. Science 2006 May 26;312(5777):1165-8.

(48) Bergfeldt K, Rydh B, Granath F, Gronberg H, Thalib L, Adami HO, et al. Risk of ovarian cancer in breast-cancer patients with a family history of breast or ovarian cancer: a population-based cohort study. Lancet 2002 Sep 21;360(9337):891-4.

(49) Massague J. Sorting out breast-cancer gene signatures. New England Journal of Medicine 2007 Jan 18;356(3):294-7.

(50) Toonen EJ, Barrera P, Radstake TR, van Riel PL, Scheffer H, Franke B, et al. Gene expression profiling in rheumatoid arthritis; current concepts and future directions. Ann Rheum Dis 2008 Feb 4.

(51) Habermann JK, Doering J, Hautaniemi S, Roblick U, Bündgen NK, Nicorici D, Kronenwett U, Rathnagiriswaran S, Mettu RKR, Ma Y, Krüger S, Bruch HP, Auer G, Guo NL, Ried T. The gene expression signature of genomic instability in breast cancer is an independent predictor of clinical outcome. Accepted by International Journal of Cancer.

(52) Guo NL, Rathnagiriswaran S, Turner P, Ducatman B, Apopa P, Abraham J, Flynn DC, Msiska Z, Vallyanthan V, Shi X, Castranova V, Qian Y. Validation of a population-based gene signature for breast cancer prognosis. Submitted to Clinical Cancer Research.