BIOINFORMATICS ORIGINAL PAPER
Vol. 24 no. 19 2008, pages 2200–2208
doi:10.1093/bioinformatics/btn374

Gene expression

A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all?

B. Haibe-Kains[1,2,*], C. Desmedt[2], C. Sotiriou[2] and G. Bontempi[1]

[1]Machine Learning Group, Department of Computer Science and [2]Functional Genomics Unit, Department of Medical Oncology, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium

Received on January 18, 2008; revised on May 30, 2008; accepted on July 15, 2008

Advance Access publication July 17, 2008

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Survival prediction of breast cancer (BC) patients independently of treatment, also known as prognostication, is a complex task since clinically similar breast tumors, in addition to be molecularly heterogeneous, may exhibit different clinical outcomes. In recent years, the analysis of gene expression profiles by means of sophisticated data mining tools emerged as a promising technology to bring additional insights into BC biology and to improve the quality of prognostication. The aim of this work is to assess quantitatively the accuracy of prediction obtained with state-of-the-art data analysis techniques for BC microarray data through an independent and thorough framework.

Results: Due to the large number of variables, the reduced amount of samples and the high degree of noise, complex prediction methods are highly exposed to performance degradation despite the use of cross-validation techniques. Our analysis shows that the most complex methods are not significantly better than the simplest one, a univariate model relying on a single proliferation gene. This result suggests that proliferation might be the most relevant biological process for BC prognostication and that the loss of interpretability deriving from the use of overcomplex methods may be not sufficiently counterbalanced by an improvement of the quality of prediction.

Availability: The comparison study is implemented in an R package called `survcomp` and is available from http://www.ulb.ac.be/di/map/bhaibeka/software/survcomp/.

Contact: bhaibeka@ulb.ac.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

During the last two decades, several clinical and pathological indicators such as histological grade, tumor size and lymph node involvement have been used for the survival prediction of breast cancer (BC) patients independently of treatment, also known as prognostication. Examples of clinical guidelines to the selection of patients who should receive adjuvant therapy are the St Gallen consensus criteria (Goldhirsh *et al*., 2003), the NIH guidelines (Eifel *et al*., 2001), the Nottingham prognostic index (NPI, Galea *et al*., 1992) and Adjuvant! Online (AOL, Olivotto *et al*., 2005). Although BC prognostication has been the object of intense research, a still open challenge is how to detect patients who needs adjuvant systemic therapy.

The advent of array-based technology and the sequencing of the human genome brought new insights into breast cancer biology and prognosis. Interestingly, several research teams conducted comprehensive genome-wide assessments of gene expression profiling and identified prognostic gene expression signatures. Examples of gene signatures which were obtained by studying the relationship between gene expression profiles and clinical outcome, are the 70-gene (van't Veer *et al*., 2002) and 76-gene (Wang *et al*., 2005) signatures. With respect to clinical guidelines, these signatures were shown to correctly identify a larger group of low-risk patients not requiring treatment. This is particularly relevant for clinicians, since reducing treatments means also reducing potential side effects and cutting costs. Another example of gene signature is reported in Sotiriou *et al*. (2006b). This study is focused on histological grade, a well-established pathological indicator rooted in the cell biology of breast cancer. In fact, clinicians encounter problems when confronted with patients with intermediate-grade tumors (Grade 2). These tumors, which represent 30–60% of cases, are a major source of inter-observer discrepancy and may display intermediate phenotype and survival, making treatment decisions for these patients a great challenge, with subsequent under- or over-treatment. By means of a supervised analysis, the authors developed a gene expression grade index based on 128 probes. The associated genes were mainly involved in cell-cycle regulation and proliferation and were consistently differentially expressed between low- and high-grade breast carcinomas. This signature, which essentially quantifies the degree of similarity between the tumor expression pattern of these genes and the tumor grade, was able to separate patients labeled with histological Grade 2 tumors into two groups having distinct clinical outcomes similar to those of histological Grades 1 and 3, respectively.

Other research groups have proposed gene expression signatures that are predictive of the clinical outcome in breast cancer [see Sotiriou and Piccart (2007) for a review]. However, since different risk prediction methods, different accuracy measures and

*To whom correspondence should be addressed.

different validation sets were used, it is not easy to compare their performance in terms of BC prognostication.

The purpose of this work is 2-fold: first set up a common and independent assessment framework to compare the performance of existing gene signatures and several state-of-the-art risk prediction methods; second, compare the prediction accuracy of these methods in several BC microarray prognostication tasks to elucidate the key characteristics of a successful risk prediction method and to bring additional insights into BC biology.

Every risk prediction model aims to assign risk values or survival probabilities to patients on the basis of the information that is available at the time of diagnosis. This is known to be a difficult task because of several issues specific to survival microarray data. First of all, censored information cannot be exploited by traditional supervised classification and regression methods, but demands the adoption of specific survival analysis techniques, like the semi-parametric Cox's proportional hazards model (Cox, 1972). A second issue is the high dimensionality of microarray data. When the number of explanatory variables exceeds by far the number of patients in the sample cohort (high feature-to-sample ratio), overfitting of naively applied data mining methods and overoptimistic performance assessment lie in wait. At the same time, it is very difficult to select the most relevant variables for prediction, because of their interdependency and the reduced power of the statistical inference procedure for high feature-to-sample ratio datasets (Bontempi, 2007). As a consequence, it is common to select variables that fit nicely the training set and fail dramatically on independent validation sets, thus leading to unstable gene signatures (Ein-Dor *et al*., 2005; Michiels *et al*., 2005) and poor prediction models. A third issue is the lack of standards in performance assessment for risk prediction models. Indeed, there exist few accuracy measures for risk prediction, and, to the best of our knowledge, no articles studied their agreement on the same set of methods and datasets. Lastly, the validation and the comparison of BC microarray prognostication methods are made difficult due to the lack of independent data.

In this work, we compare the performance of 13 risk prediction methods on more than 1000 patients. This is made possible thanks to the recent publications of several large microarray datasets in gene expression databases, such as the Gene Expression Omnibus (GEO, Barrett *et al*., 2005) An important outcome of the analysis is that, in spite of the large number of samples, there is no statistical evidence that complex methods outperform the simplest BC prognostication techniques. This result suggests that the loss of interpretability deriving from the use of overcomplex data analysis strategies may not be sufficiently counterbalanced by an improvement in the quality of prediction.

Finally, it is worth to mention that the present article complies with the *research reproducibility* guidelines proposed in Gentleman (2005) in terms of availability of the code and reproducibility of results and figures.[1] A list of acronyms used throughout the article is given in Supplementary Table 1.

[1]Raw gene expression and clinical data are publicly available in the GEO public database and the Sweave version of the article including the standalone R code (R Development Core Team, 2007) is available from http://www.ulb.ac.be/di/map/bhaibeka/survcompaper/.

2 METHODS

2.1 Notations for survival analysis

Throughout the article we will adopt the following notation: upper case and lower case letters represent random variables and their realization, respectively, while bold letters denote vectors or matrices. Let us denote the time as t. We suppose that a sample cohort of n patients is available and that for each patient we observed a p-dimensional vector of covariates $\mathbf{x}_i$ with $1 \leq i \leq n$ at the time of diagnosis $t=0$, as well as the evolution of the survival status. Since we limit our study to microarray data, the covariate $\mathbf{x}_i$ denotes the expression of the whole genome of the i-th patient. Survival data for the i-th patient are denoted as follows: t_i stands for the event time, c_i for the censoring time and δ_i for the censoring indicator ($\delta_i=1$ if $t_i \leq c_i$ and $\delta_i=0$ if $t_i > c_i$). We introduce the counting process $d_i(t)=1$ if $t_i \leq t$ and $d_i(t)=0$ if $t_i > t$ to denote survival status at any time t where $d_i(t)=1$ indicates that patient i experienced an event prior to time t.

2.2 Risk prediction methods

The aim of a risk prediction model is to predict future survival status for all patients in the cohort. All the risk prediction models considered in this study return a *risk score* denoted by R, that is a continuous value which quantifies the risk of a patient to experience an event. Clinicians often use the risk score to derive *risk groups*, denoted by G, on the basis of quantiles of the risk score distribution. Although the discretization of individual risk scores into a finite (and often small) set of risk groups may introduce bias (Gerds and Schumacher, 2001), this approach is very intuitive and conforms to the daily doctors' decision making process, e.g. the attribution of either low or high risk to patients. In the following, the quantity r_i and g_i will denote the risk score and the risk group for patient i, respectively. G is either 0 or 1 for a low- or high-risk patient, respectively.

In this experimental study, we decided to focus on a set of 13 state-of-the-art methods (summarized in Table 1) with the ambition of being representative of a large number of risk prediction strategies. The first risk prediction method is also the simplest one and defines the risk score as the expression of a single proliferation gene (AURKA) well studied in literature (Hanahan and Weinberg, 2000). The following 10 methods (from 2 to 11) are characterized by the type of observed genotype (input data), the dimension reduction strategy, the structure of the model, the learning algorithm and the predicted phenotype (outcome variable).

Genotype: it can be the expression of a single proliferation gene (AURKA), the expression of a biologically driven selection of genes of interest (BD) or the expression of the whole genome (GW). AURKA and the small set of genes in BD were selected to represent several biological processes in BC (Hanahan and Weinberg, 2000). The selected genes were AURKA (also known as STK6, 7 or 15), PLAU (also known as uPA), STAT1, VEGF, CASP3, ESR1 and ERBB2, representing the proliferation, tumor invasion/metastasis, immune response, angiogenesis, apoptosis phenotypes and the ER and HER2 signaling, respectively.

Dimension reduction strategy: we use either a simple univariate ranking (RANK) of the k most relevant features or a selection of the first k principal components (PCA). Univariate ranking uses Wilcoxon rank sum test (Wilcoxon, 1945) in the case of binary outcome or Cox's proportional hazards model (Cox, 1972) in the case of survival outcome. The signature size k is either fixed or tuned by cross-validation (CV) as described in Section 2.2.1. It is worth to note that no dimension reduction was performed for BD input data due to the low dimensionality of the input space.

Structure of the model: we adopt either a multivariate (MULTIV) model or a linear combination of univariate models (COMBUNIV), particularly interesting in a high-dimensional setting (Haibe-Kains *et al*., 2008; Kittler *et al*., 1998).

Learning algorithm: we consider four types of learning algorithms: (i) the linear combination of gene expressions weighted by the significance computed from the Wilcoxon rank sum test (WILCOXON) that allowed for identifying the most relevant genes to discriminate the patients with

Table 1. Characteristics of the risk prediction methods studied in this work

	Genotype	Dim. reduction	Structure	Learning algo.	Phenotype
1	**AURKA**				
2	BD		COMBUNIV	WILCOXON	HG
3	BD		COMBUNIV	COX	SURV
4	BD		MULTIV	LM	TOE
5	BD		MULTIV	COX	SURV
6	GW	RANK (CV)	COMBUNIV	WILCOXON	HG
7	GW	RANK (CV)	COMBUNIV	COX	SURV
8	GW	RANK (CV)	MULTIV	RCOX	SURV
9	GW	PCA (CV)	COMBUNIV	WILCOXON	HG
10	GW	PCA (CV)	COMBUNIV	COX	SURV
11	GW	PCA (CV)	MULTIV	RCOX	SURV
12				GENE76	
13				GGI	

We will use the words in bold to refer to the models that were fully defined in previous publications. Otherwise, the model name is a concatenation of all its characteristics separated by '.'.

histological Grades 1 and 3 tumors, (ii) the multivariate linear regression model (LM), (iii) the linear combination of gene expressions weighted by the significance computed from the univariate Cox's proportional hazards model (COX) and (iv) the multivariate Cox's model with $L1$ regularization (RCOX) as implemented in Park and Hastie (2007).

Phenotype: we use three different phenotypical informations to fit the prediction models: (i) the binary class defined by histological grades 1 and 3 (HG), (ii) the censored survival data (SURV) and (iii) the time of events (TOE), i.e. the times from diagnosis until the patients experienced an event. In the following, we will denote each of the 10 models with a unique label obtained by concatenating the acronyms referring to its characteristics (Table 1). For instance, BD.COMBUNIV.COX.SURV refers to a combination of univariate Cox's proportional hazards models fitted from a biologically driven selection of genes.

The last two models taken into consideration are the published GENE76 (Wang *et al*., 2005) and GGI (Sotiriou *et al*., 2006b) models. The GENE76 model is defined as a hierarchical model using two linear combinations of the top gene expressions with respect to a ranking based on Cox's proportional hazards model. The choice of the linear combination to compute the risk score depends on the estrogen receptor status of the patient. The GGI model consists of a linear combination of the expressions of the top probes ranked according to their standardized mean difference (Hedges and Olkin, 1987) between patients with histological grades 1 and 3 tumors. The weights of the linear combination are simply the signs of the ranking statistics.

2.2.1 Tuning of hyperparameters The GGI and GENE76 models did not require any tuning of hyperparameters since they were fully defined in previous publications. Only the models based on dimension reduction and regularization required the tuning of an hyperparameter. For GW models, a simple ranking or a principal components analysis was used to select the k most relevant features. The hyperparameter k was either set to 30, this signature size being reported as a good trade-off between relevance and model complexity in the comparison study of Dudoit *et al*. (2002), or tuned using a 5-fold CV procedure (see Section 1 in Supplementary Material). The dimension reduction strategies using the latter procedure are referred as RANKCV and PCACV for the univariate ranking and the principal component analysis, respectively. For methods using RCOX as learning algorithm, the hyperparameter for the penalty term was tuned by using a 5-fold CV as in Park and Hastie (2007).

2.3 Performance assessment

In order to assess the performance of the risk prediction methods, we used five accuracy measures: the time-dependent receiver operating characteristic (ROC) curve (Heagerty *et al*., 2000), the sensitivity and specificity, the concordance index (Harrell *et al*., 1996), the Brier score (Brier, 1950; Graf *et al*., 1999) and the traditional hazard ratio (HR) from Cox's proportional hazards model (Cox, 1972).

2.3.1 Time-dependent ROC Curve The ROC curve is a standard technique for assessing the performance of a continuous variable for binary classification (Sweets, 1988). A ROC curve is a plot of sensitivity versus $1-$specificity for all the possible cutoff values of the continuous variable, denoted by c. In survival analysis, the continuous variable is the risk score, denoted by R, and the binary class to predict is the event occurrence, denoted by $D(t)$. As the event occurrence is time-dependent, time-dependent ROC curves are more appropriate than conventional ones. In Heagerty *et al*. (2000), the authors proposed to summarize the discrimination potential of a risk score R, estimated at the diagnosis time $t=0$, by calculating ROC curves for cumulative event occurrence by time t. Once we define the sensitivity SE and the specificity SP as follows

$$\mathrm{SE}(c,t,r) = \Pr\{r > c | d(t)=1\} \quad (1)$$

$$\mathrm{SP}(c,t,r) = \Pr\{r \leq c | d(t)=0\} \quad (2)$$

the ROC curve $\mathrm{ROC}(t)$ at time t is the plot of $\mathrm{SE}(c,t,r)$ versus $1-\mathrm{SP}(c,t,r)$, where the cutoff point c is the parameter. In order to estimate the conditional probabilities in (1) and (2), accounting for possible censoring, we used the nearest neighbor estimator for the bivariate distribution function proposed by Akritas (1994).

From the ROC curve $\mathrm{ROC}(t)$ we can derive the area under the curve (AUC) quantity, denoted by $\mathrm{AUC}(t)$. Since AUC depends on time t, we define the *integrated area under the curve* (IAUC) as the area under $\mathrm{AUC}(t), \forall t \in T$. Note that, in this study, the larger the AUC at time t, the better is the predictability of time to event (TTE) at time t. Similarly, the larger IAUC, the better is the average predictability of TTE.

2.3.2 Sensitivity and specificity A widely used performance criterion for a clinical test is the pair {sensitivity, specificity} (Simon, 2005). However, the calculation of these values from survival data requires estimators accounting for TTE and possible censoring. We used the estimators defined in (1) and (2) for sensitivity and specificity, respectively. For risk score prediction, we estimated the specificity for a sensitivity of 90% in accordance with the St Gallen (Goldhirsh *et al*., 2003) and National Institutes of Health (Eifel *et al*., 2001) treatment guidelines. For risk group prediction, we estimated both the sensitivity and the specificity of the binary classification returned by all the methods. Note that the larger the sensitivity and the specificity, the better is the predictability of TTE.

2.3.3 Concordance index The concordance index (C-index) computes the probability that, for a pair of randomly chosen comparable patients, the patient with the higher risk prediction will experience an event before the lower risk patient. The C-index takes the form

$$C\text{-index} = \frac{\sum_{i,j\in\Omega} 1\{r_i > r_j\}}{|\Omega|} \quad (3)$$

where r_i and r_j stand for the risk predictions of the i-th and the j-th patient, respectively, and Ω is the set of all the pairs of patients $\{i,j\}$ who meet one of the following conditions: (i) both patients i and j experienced an event and time $t_i < t_j$ or (ii) only patient i experienced an event and $t_i < c_j$. In the case of risk group prediction, an additional condition must be met, that is the risk predictions are different for patients i and j (no ties in r).

Note that the C-index is a generalization of the $\mathrm{AUC}(t)$, though it is unable to represent the evolution of performance with respect to time (Harrell *et al*., 1996).

Standard errors, confidence intervals and P-values for the C-index are computed by assuming asymptotic normality (Pencina and D'Agostinno, 2004). Note that, in this study, the larger C-index, the better is the predictability of TTE.

2.3.4 Brier score The Brier score, denoted by BSC, is defined as the squared difference between an event occurrence and its predicted probabilities at time t. Probabilities of event, denoted by Q, can be derived from Cox's proportional hazards model fitted with the risk score R or risk group G predictions. Intuitively, if a patient experiences no event at time t, the event predicted probability should be close to zero. Symmetrically, if the patient experiences an event the probability should be close to one. The BSC formalizes this intuition by computing the time dependent quantity

$$\mathrm{BSC}(t,q)=\sum_{i=1}^{n}(d_i(t)-q_i(t))^2 W \tag{4}$$

where the weights W are used to remove a large sample censoring bias (Gerds and Schumacher, 2006; Graf *et al*., 1999).

A summary of the predictability error over times is returned by the integrated Brier score, denoted by IBSC. Note that the lower the BSC, the better is the predictability of TTE at time t. Similarly, the lower the IBSC, the better is the average predictability of TTE.

For judging the (I)BSC, we will rely on the score of a benchmark risk prediction model which is obtained with the overall Kaplan–Meier estimator (Kaplan and Meier, 1958) for the survival function (this model is called KM in further sections). This simple risk prediction model corresponds to a model which assigns the same risk prediction to all patients. It ignores the information contained in explanatory variables completely and thus provides a suitable benchmark value similar as the one obtained with the null model in linear regression.

2.3.5 Hazard ratio In this work we used the HR as an accuracy measure for the risk group prediction in order to keep it interpretable and comparable between different risk prediction methods as the scale of predictions is well defined (see Section 2.1). HR is a summary of the risk difference between several survival curves estimated by Cox's proportional hazards model (Therneau and Grambsch, 2000). Cox's model assumes that the relative risk of event between groups is constant at each interval of time. The hazard function for a patient i as defined by Cox's proportional hazards model, can be written as

$$\lambda_i(t)=\lambda_0(t)\exp(\beta g_i) \tag{5}$$

Given the nature of the variable G, $\lambda_0(t)$ is the hazard function for a patient in the low-risk group. Moreover, the hazard function for any patient in the high-risk group is $\psi\lambda_0(t)$ (proportional hazards), so ψ is the HR with $\psi=\exp(\beta)$. Note that, in this study, the larger the HR, the larger is the difference in survival probabilities between the groups of patients, and consequently the better is the discrimination between low- and high-risk groups.

2.4 Performance comparison

To test whether a method performs significantly better than another one, we used two types of statistical tests: (i) a paired Student t-test based on the assumption of normality for the natural logarithm of the hazard ratio (i.e. the coefficient β in Cox's proportional hazards model) and the concordance index; (ii) a paired Wilcoxon rank sum test of event occurrence with respect to the time t for the AUC(t) and BSC(t). We considered that a method performs significantly better than another one if its performance is significantly better with a $P<0.05$ and the difference between the performance estimates of the two methods is larger than 1% of the lowest value. Note that we did not statistically compare the estimations of sensitivity and specificity due to the lack of standard statistical test.

The concordance indices and the hazard ratios for all the risk prediction methods are represented using a forest plot (Lewis and Clarke, 2001). The accuracy measures are shown as squares centered on the point estimate of the performance of each method. A horizontal line runs through the square to show its 95% confidence interval.

3 RESULTS

3.1 Breast cancer datasets

In order to compare different risk prediction methods with published gene signatures, we used four large microarray BC datasets collected with Affymetrix microarray platform (22 283 common probes), called VDX (Wang *et al*., 2005), TBG (Desmedt *et al*., 2007), TAM (Haibe-Kains *et al*., 2008; Loi *et al*., 2007) and UPP (Miller *et al*., 2005). These datasets are publicly available from the GEO database[2] through accession numbers GSE2034, GSE7390, GSE6532/GSE9195 and GSE3494, respectively. VDX includes the gene expressions of 286 untreated node-negative BC patients and was used to build GENE76 and to validate GGI (see end of Section 2.2). Only TBG exhibited the same criteria for the selection of patients (198) than VDX, i.e. untreated node-negative BC patients, and was used as an official validation of GENE76 and GGI. TAM was composed of 354 ER-positive BC patients (the largest molecular group of BC) being homogeneously treated by tamoxifen therapy. UPP was composed of 251 patients being treated with heterogeneous therapies. Although the selection of patients was different for VDX and TBG, TAM and UPP datasets might contain important prognostic information as well and could, therefore, be used as additional validation sets. Due to their homogeneity in selection of patients, we should consider VDX and TBG as the most important datasets in this comparative study. The results obtained with TAM and UPP made possible a more thorough assessment of the performance.

We considered the distant metastasis free survival of BC patients as the survival endpoint for VDX, TBG and TAM. This endpoint refers to the appearance of distant metastasis only. We considered the relapse free survival of BC patients, i.e. appearance of local, regional or distant relapses, in UPP as the information on distant metastasis was not available. All the survival data were censored at 10 years as in Desmedt *et al*. (2007).

We used VDX as training set and TBG, TAM and UPP as validation sets. Although this choice was guided by the original publications in BC prognostication, we also performed all the analyses using TBG as training set to ensure that our results were not driven by the choice of the training set (Michiels *et al*., 2005). We obtained similar results that let the conclusions of this study unchanged (see Section 10 in Supplementary Material).

We assessed the performance in the training set and in the three validation sets using all the risk prediction methods summarized in Table 1. The AURKA model was used as reference because of its low complexity. Both risk score and risk group predictions were compared.

3.2 Risk score prediction

This section presents the results for the performance assessment of the risk score predictions using the five performance criteria presented in Section 2.3.

The specificity for a sensitivity of 90%, is reported for all the risk prediction methods in Table 2. We observed values consistent with

[2]http://www.ncbi.nlm.nih.gov/geo/.

Table 2. Specificity for a sensitivity of 90% for risk score prediction in the training set (VDX) and the three validation sets (TBG, TAM and UPP)

Model	Specificity			
	VDX	TBG	TAM	UPP
AURKA	0.253[a]	0.348	0.394	0.293
BD.COMBUNIV.WILCOXON.HG	0.247	0.311	0.362	0.258
BD.COMBUNIV.COX.SURV	0.268	0.360	0.394	0.293
BD.MULTIV.LM.TOE	0.268	0.460	0.220	0.217
BD.MULTIV.COX.SURV	0.205	0.118	0.372	0.131
GW.RANK.COMBUNIV.WILCOXON.HG	0.258	0.373	0.277	0.227
GW.RANK.COMBUNIV.COX.SURV	0.400	0.360	0.362	0.162
GW.RANK.MULTIV.RCOX.SURV	0.468	0.242	0.326	0.242
GW.PCA.COMBUNIV.WILCOXON.HG	0.147	0.298	0.067	0.091
GW.PCA.COMBUNIV.COX.SURV	0.426	0.379	0.450	0.217
GW.PCA.MULTIV.RCOX.SURV	0.405	0.509	0.358	0.141
GENE76	0.626	0.391	0.309	0.088
GGI	0.258[a]	0.522	0.422	0.308

[a]As AURKA and GGI models were not fitted on VDX, this dataset can be considered as a validation set.

the literature (Buyse *et al*., 2006; Desmedt *et al*., 2007; Foekens *et al*., 2006; van de Vijver *et al*., 2002). GGI was the best method in two validation sets, yielding larger specificity values than the simplest model AURKA. The increase was estimated to 17.4, 2.8 and 1.5% for TBG, TAM and UPP datasets, respectively.

The performance assessment using the concordance index, the time-dependent ROC curve and the Brier score are given in Supplementary Figures 1, 2–5 and 6–9, respectively.

Looking at the most complex models, i.e. multivariate (MULTIV) survival (SURV) models using genome-wide data (GW) and GENE76, we observed overoptimistic performance estimates in training set compared to the validation sets. Although we used advanced machine learning techniques to control overfitting, i.e. linear combination of univariate models or $L1$ regularization in Cox's proportional hazards model, these complex models failed to outperform simpler ones in a validation setting.

The simple AURKA model was competitive in all the datasets. As mentioned earlier, we statistically compared the performance of all the models with AURKA (Table 3). Only GGI was significantly better in at least two validation sets whatever the accuracy measure. It is worth to note that AURKA outperformed KM, the benchmark model for the Brier score (see Section 2.3.4), in all the datasets except for TBG.

As we did not observe a significant improvement using cross-validated dimension reduction strategies (see Section 2.2.1), we reported the performance of the methods using either RANKCV or PCACV in Supplementary Tables 3 and 4.

We computed all the pairwise performance comparisons (see Section 9.1 in Supplementary Material) and observed again that complex models performed poorly in validation sets compared to simpler ones. According to the IAUC performance criterion in TBG dataset, the GW methods using SURV phenotype are significantly better than the other risk prediction methods. However, these significant results are not confirmed in the other validation sets. We noticed also that, unlike GGI, GENE76 was consistently worse than most of the other considered methods.

3.3 Risk group prediction

This section presents the results for the performance assessment of the risk group predictions using the five performance criteria presented in Section 2.3. We used the tertile to define the risk groups (see Section 2.2), leaving 33% of the patients in the low-risk group and the remaining 66% in the high-risk group. This proportion is usually observed in BC prognostication (Buyse *et al*., 2006; Desmedt *et al*., 2007; van't Veer *et al*., 2002; Wang *et al*., 2005).

The sensitivity and the specificity for all the risk prediction methods are reported in Table 4. Again, we observed values for sensitivity and specificity that are consistent with the literature. GGI was the best method in two validation sets, yielding larger sensitivity and specificity values than the simplest model AURKA. However, the difference is small except for TBG dataset.

The performance assessment of the risk group predictions using the concordance index, the HR and the Brier score are given in Supplementary Figures 10, 11 and 12–15, respectively.

Similarly to the risk score prediction, the simple AURKA model was competitive in all the datasets. We statistically compared the performance of all the models with AURKA (see Table 5). In order to illustrate the gain of using GGI over AURKA for risk groups prediction, we compared the survival curves for each dataset separately (see Supplementary Figs 16–19). Because neither AURKA nor GGI were fitted on VDX, this dataset is not considered as a training set. We observed a substantial improvement in risk group prediction in terms of survival probabilities in the low-risk group in using the risk groups predicted by GGI compared to AURKA. At 5 years, the increase in survival probabilities in the low-risk group was estimated to 2, 6, 0 and 1% for VDX, TBG, TAM and UPP, respectively.

As we did not observe a significant improvement using cross-validated dimension reduction strategies, we reported the results of the methods using either RANKCV or PCACV in Supplementary Tables 5 and 6.

We computed all the pairwise comparisons (see Section 9.2 in Supplementary Material) and made similar observations than for the risk score prediction. This holds true if we restrict the analysis to TBG dataset. However, less significant differences between methods were detected. This can be explained by the loss of prognostic information due to the risk group creation.

4 DISCUSSION AND CONCLUSIONS

We assisted recently to intense research in BC prognostication due to a growing availability of high-dimensional genomic information that could potentially be used for risk prediction (Simon, 2005). The situation is often characterized by a relatively small number of patients and a large number of explanatory variables. This high-dimensional setting can be seen as an opportunity to create better risk prediction models compared to those solely based on clinical data and/or single markers. At the same time this prevents from a straightforward use of classical approaches of statistical modeling and data analysis.

To the best of our knowledge, the present study is the first that has systematically compared the performance of state-of-the-art survival methods for BC prognostication from gene expression data by using a training/validation framework and several accuracy measures for survival prediction. The public availability of large microarray BC datasets and the recently introduced measures for performance

Table 3. Performance for risk score prediction in the training set (VDX) and the three validation sets (TBG, TAM and UPP)

Model	*C*-index				IAUC				IBSC			
	VDX	TBG	TAM	UPP	VDX	TBG	TAM	UPP	VDX	TBG	TAM	UPP
KM									0.189	0.145	0.141	0.151
AURKA	0.636[a]	0.609	0.683	0.637	0.636[a]	0.601	0.674	0.63	**0.178**[a]	0.144	**0.132**	**0.146**
BD.COMBUNIV.WILCOXON.HG	0.606	0.618	0.687	0.629	0.602	**0.643**	**0.682**	0.619	0.185	0.143	0.131	0.146
BD.COMBUNIV.COX.SURV	0.638	0.613	0.684	0.638	0.638	0.607	0.675	0.632	0.178	0.143	0.131	0.146
BD.MULTIV.LM.TOE	0.601	0.645	0.683	0.622	0.602	**0.681**	**0.682**	0.63	0.186	**0.141**	0.132	0.147
BD.MULTIV.COX.SURV	0.649	0.603	0.657	0.598	0.649	0.596	0.642	0.6	**0.172**	0.15	0.132	0.149
GW.RANK.COMBUNIV.WILCOXON.HG	0.619	0.624	0.691	0.653	0.639	**0.617**	**0.684**	**0.662**	0.182	**0.141**	0.131	0.146
GW.RANK.COMBUNIV.COX.SURV	**0.742**	**0.665**	0.65	0.637	**0.774**	**0.686**	0.638	**0.651**	**0.148**	0.153	0.158	0.172
GW.RANK.MULTIV.RCOX.SURV	**0.774**	0.663	0.638	0.63	**0.823**	**0.715**	0.635	**0.654**	**0.136**	0.151	0.175	0.16
GW.PCA.COMBUNIV.WILCOXON.HG	0.586	0.591	0.566	0.579	0.617	0.616	0.565	0.561	0.186	**0.14**	0.136	0.148
GW.PCA.COMBUNIV.COX.SURV	**0.726**	**0.676**	0.695	0.594	**0.749**	**0.705**	0.672	0.589	**0.154**	0.147	0.153	0.177
GW.PCA.MULTIV.RCOX.SURV	**0.75**	**0.694**	0.69	0.591	**0.779**	**0.733**	0.667	0.598	**0.143**	0.155	0.171	0.176
GENE76	**0.754**	0.64	0.667	0.557	**0.794**	**0.632**	0.633	0.558	**0.158**	0.153	0.149	0.182
GGI	0.613[a]	**0.652**	**0.718**	0.67	0.611[a]	**0.671**	**0.717**	**0.686**	0.183[a]	**0.14**	**0.13**	**0.142**

The accuracy measures in bold are significantly better than the accuracy of AURKA model. In case of IBSC, the accuracy measures of AURKA are in **bold** if they are significantly better than KM, the benchmark model, whatever the performance improvement.
[a]As AURKA and GGI models were not fitted on VDX, this dataset can be considered as a validation set.

Table 4. Sensitivity and specificity for risk group prediction in the training set (VDX) and the three validation sets (TBG, TAM and UPP)

Model	Sensitivity				Specificity			
	VDX	TBG	TAM	UPP	VDX	TBG	TAM	UPP
AURKA	0.802[a]	0.892	0.900	0.806	0.389[a]	0.379	0.365	0.354
BD.COMBUNIV.WILCOXON.HG	0.792	0.892	0.880	0.806	0.389	0.379	0.365	0.354
BD.COMBUNIV.COX.SURV	0.812	0.892	0.900	0.833	0.400	0.379	0.369	0.359
BD.MULTIV.LM.TOE	0.833	0.946	0.840	0.778	0.411	0.391	0.358	0.348
BD.MULTIV.COX.SURV	0.792	0.784	0.900	0.806	0.389	0.354	0.369	0.354
GW.RANK.COMBUNIV.WILCOXON.HG	0.812	0.892	0.840	0.833	0.400	0.379	0.358	0.359
GW.RANK.COMBUNIV.COX.SURV	0.885	0.892	0.860	0.778	0.437	0.379	0.362	0.348
GW.RANK.MULTIV.RCOX.SURV	0.927	0.892	0.840	0.806	0.458	0.379	0.358	0.354
GW.PCA.COMBUNIV.WILCOXON.HG	0.740	0.838	0.760	0.750	0.363	0.366	0.344	0.343
GW.PCA.COMBUNIV.COX.SURV	0.896	0.892	0.940	0.778	0.442	0.379	0.376	0.348
GW.PCA.MULTIV.RCOX.SURV	0.896	0.919	0.880	0.722	0.442	0.385	0.365	0.338
GENE76	0.958	0.919	0.840	0.722	0.474	0.385	0.358	0.335
GGI	0.844[a]	1.000	0.900	0.861	0.416[a]	0.404	0.369	0.359

[a]As AURKA and GGI models were not fitted on VDX, this dataset can be considered as a validation set.

assessment in survival analysis allow to perform an in-depth comparative study in order to elucidate the key characteristics of a successful risk prediction method and to bring new insights into BC prognostication.

We used four large microarray BC datasets (one for training and three for validation) in order to compute unbiased estimates of five accuracy measures (see Section 2.3) for 13 risk prediction methods (see Section 2.2). As expected, we observed that complex methods, e.g. multivariate survival models fitted using dimension reduction from genome-wide data or GENE76, performed very well in the training set. However, the performances in the validation sets were poorer and they failed to outperform consistently the simplest model, i.e. AURKA, in spite of the use of machine learning strategies (namely combination of univariate models or regularization) to reduce the risk of overfitting. These results highlighted the fact that the loss of interpretability deriving from the use of overcomplex methods in survival analysis of BC microarray data might be not sufficiently counterbalanced by an improvement in the quality of prediction.

Interestingly, AURKA, the simplest model defining the risk score as the expression of a single proliferation gene, performed well in all the survival prediction tasks. From Tables 3 and 5, we noticed that AURKA was significantly better than KM, the benchmark model ignoring all genetic information, except only for the risk score prediction in TBG. Several other methods outperformed consistently KM as shown in Supplementary Figures 22 and 25. These results are very encouraging for BC prognostication as it was shown that, in diffuse large B-cell lymphoma prognostication for instance, most

Table 5. Performance for risk group prediction in the training set (VDX) and the three validation sets (TBG, TAM and UPP)

Model	*C*-index				HR				IBSC			
	VDX	TBG	TAM	UPP	VDX	TBG	TAM	UPP	VDX	TBG	TAM	UPP
KM									0.189	0.145	0.141	0.151
AURKA	0.685[a]	0.729	0.834	0.673	2.04[a]	2.43	4.64	1.84	**0.182**[a]	**0.14**	**0.133**	**0.147**
BD.COMBUNIV.WILCOXON.HG	0.675	0.728	0.804	0.673	1.86	2.39	4.06	1.89	0.184	0.14	0.134	0.147
BD.COMBUNIV.COX.SURV	0.698	0.729	0.834	**0.705**	2.17	2.43	4.58	**2.11**	0.181	0.141	0.133	0.146
BD.MULTIV.LM.TOE	0.721	0.811	0.716	0.647	2.26	3.7	2.52	1.77	**0.18**	**0.137**	0.138	0.149
BD.MULTIV.COX.SURV	0.685	0.611	0.828	0.66	2.21	1.59	4.89	1.84	0.182	0.146	0.132	0.148
GW.RANK.COMBUNIV.WILCOXON.HG	0.694	0.785	0.775	0.733	1.99	3.61	3.42	2.42	0.182	**0.139**	0.136	**0.146**
GW.RANK.COMBUNIV.COX.SURV	**0.836**	0.77	0.778	0.632	**4.69**	2.96	3.53	1.53	**0.168**	0.143	0.139	0.156
GW.RANK.MULTIV.RCOX.SURV	**0.906**	0.765	0.749	0.696	**9.62**	3.28	3	2.18	**0.159**	0.15	0.147	0.157
GW.PCA.COMBUNIV.WILCOXON.HG	0.616	0.69	0.589	0.586	1.46	1.94	1.3	1.37	0.187	0.142	0.14	0.15
GW.PCA.COMBUNIV.COX.SURV	**0.843**	0.734	0.909	0.63	**5.13**	2.62	9.5	1.53	**0.167**	0.147	0.133	0.174
GW.PCA.MULTIV.RCOX.SURV	**0.826**	0.749	0.818	0.564	**4.3**	2.6	4.64	1.15	**0.169**	0.142	0.136	0.177
GENE76	**0.903**	0.756	0.754	0.548	**8.37**	2.79	3.52	1.16	**0.16**	0.146	0.145	0.17
GGI	0.706[a]	**0.906**	0.824	**0.769**	2.12[a]	**7.24**	4.03	**2.88**	0.181[a]	**0.133**	0.134	**0.145**

The accuracy measures in bold are significantly better than the accuracy of AURKA model. In case of IBSC, the accuracy measures of AURKA are in **bold** if they are significantly better than KM, the benchmark model, whatever the performance improvement.
[a]As AURKA and GGI models were not fitted on VDX, this dataset can be considered as a validation set.

prediction models were not better than KM in a validation setting [see discussion of Schumacher *et al.* (2007)]. Moreover, GGI was the only model that outperformed AURKA in at least two validation sets whatever the accuracy measure for risk score and risk group predictions. As GGI is a linear combination of proliferation gene expressions (see Section 2.2), these results highlight the importance of proliferation measured by gene expression profiling in BC prognostication, and confirm the results of Sotiriou *et al.* (2006a, b).

In order to go further in the comparison of the different risk prediction methods, we computed all the pairwise performance comparisons for all the methods (see Section 9 in Supplementary Material) and we observed that models using only the biologically driven selection of genes of interest (BD) led to similar performance than models using genome-wide data (GW) with ranking (RANK) or principal components analysis (PCA). This suggests that finding a combination of relevant variables in a high-dimensional setting is a difficult task since simple dimension reduction methods did not succeed to significantly improve the models from genome-wide data compared to simpler models using a very small set of genes selected from literature. Our results are consistent with earlier studies focused on stability of feature selection for GW methods (Ein-Dor *et al.*, 2005; Michiels *et al.*, 2005). Moreover, we observed that models fitting the histological grade (HG) as phenotype performed globally better in validation sets than models fitting survival data (SURV or TOE) suggesting that we did not succeed to catch additional information about prognostication in using survival models. The fact that the performance of GGI was the best in validation sets reinforced this observation as the GGI was built using a method similar to our weighted combination of relevant genes for the histological grade (see Section 2.2).

It is worth to mention that all the accuracy measures were in nice agreement in our comparative study. Smaller significant differences in performance estimates were detected by the IAUC and the IBSC criteria, probably due to the type of statistical test [paired Wilcoxon rank sum test for AUC(t) or BSC(t) compared to paired Student t-test for C-index or HR, see Section 2.3].

A final analysis concerns the performance comparison with the classical indicators, such as the histological grade (Scarff and Torloni, 1968), AOL (Olivotto *et al.*, 2005) and NPI (Galea *et al.*, 1992). In our comparative study, GGI was consistently better than the histological grade (data not shown). We were not able to compute the risk scores for AOL and NPI on TAM and UPP datasets due to lack of information. As shown in Supplementary Table 10, AURKA and GGI outperformed consistently AOL in the VDX and TBG datasets, except for the IAUC in TBG. However, this is not the case for NPI. These results suggest that the superiority of microarray-based risk prediction methods is not obvious and need further investigations.

In conclusion, our results challenge the use of microarray technology to screen the whole genome for BC prognostication of global populations of patients. Indeed, we found that models using a single gene or a small set of biologically driven selected genes yielded similar or even better performance than models fitted from genome-wide data. Although GGI, the model yielding the best performances in validation sets, uses a set of 128 probes, it can be considered as a simple extension of AURKA, i.e. a quantification of proliferation using more genes. Moreover the authors recently showed that we could yield similar performance in using only a small subset of these 128 probes (Durbecq *et al.*, 2007). The use of high-sensitivity gene expression profiling technologies such as the reverse transcription polymerase chain reaction, in addition to be cheaper and more user friendly, might improve the performance of these risk prediction models.

The relevance of proliferation for BC prognostication was previously reported by several other research groups. Indeed, Thomassen *et al.* (2007) found that cell cycle and cell proliferation represented the predominant overlaps in gene ontology categories of the nine prognostic signatures they compared. Yu *et al.* (2007)

also conducted pathway analyses of five published prognostic gene signatures and found that the signatures had many pathways in common such as cell cycle, regulation of cell cycle, mitosis, apoptosis, etc. Our group also investigated in a large meta-analysis of publicly available gene expression data, how different gene lists may give rise to signatures with equivalent prognostic performance and found by dissecting these signatures according to the main molecular processes involved in breast cancer, that proliferation may be the common driving force of several prognostic signatures (Sotiriou *et al*., 2006a).

Until now, the generation of the prognostic signatures has been done on global populations of BC patients. However, since it is clear that breast cancer is a molecular heterogeneous disease, with subgroups defined primarily by the estrogen (ER) and HER2 receptors (Perou *et al*., 2000; Sotiriou *et al*., 2003), prognosis could be refined to these molecularly homogeneous subgroups of patients. We showed, for example, in a meta-analysis recently published by our group that proliferation is the strongest parameter predicting clinical outcome in the ER+/HER2 − subgroup of patients only (group of patients representing more than 66% of the global population), whereas immune response and tumor invasion appear to be the main biological processes associated with prognosis in the ER−/HER2− and HER2+ subgroups, respectively (Desmedt *et al*., 2008; Sotiriou *et al*., 2007). These recent results suggest that we could improve BC prognostication by restricting the genome-wide analysis to specific molecular subtypes. This will be the subject of further investigations.

ACKNOWLEDGEMENTS

The authors would like to thank Martin Schumacher and Thomas Gerds for providing the `prodlim` R package, and Yann-Aël Le Borgne for his constructive comments.

Funding: This work was supported by the Belgian National Foundation for Scientific Research FNRS (B.H.-K., C.D., C.S.), and by the MEDIC Foundation (C.S.).

Conflict of Interest: C. Sotiriou, M. Delorenzi and M. Piccart are named inventors on a patent application for the Gene expression Grade Index used in this study. There are no other conflicts of interest.

REFERENCES

Akritas,M.G. (1994) Nearest neighbor estimation of a bivariate distribution under random censoring. *Ann. Stat.*, **22**, 1299–1327.

Barrett,T. *et al*. (2005) NCBI GEO: mining millions of expression profiles – database and tool. *Nucleic Acids Res.*, **33**, D562.

Bontempi,G. (2007) A blocking strategy to improve gene selection for classification of gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **4**, 293–300.

Brier,G.W. (1950) Verification of forecasts expressed in terms of probabilities. *Mon. Weather Rev.*, **78**, 1–3.

Buyse,M. *et al*. (2006) Validation and clinical utility of a 70-gene prognostic signature for patients with node-negative breast cancer. *J. Natl. Cancer Inst.*, **98**, 1183–1192.

Cox,D.R. (1972) Regression models and life tables. *J. R. Stat. Soc. Ser B*, **34**, 187–220.

Desmedt,C. *et al*. (2007) Strong time-dependency of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multi-centre independent validation series. *Clin. Cancer Res.*, **13**, 3207–3214.

Desmedt,C. *et al*. (2008) Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin. Cancer Res.* (in press).

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Durbecq,V. *et al*. (2007) Transforming genomic grade index (GGI) into a user-friendly qRT-PCR tool which will assist clinicians and patients in optimizing treatment of early breast cancer. *Journal of Clinical Oncology*, **25**, 21058.

Eifel,P. *et al*. (2001) National institutes of health consensus development conference statement: adjuvant therapy for breast cancer. *J. Natl. Cancer Inst.*, **93**, 979–989.

Ein-Dor,L. *et al*. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171–178.

Foekens,J.A. *et al*. (2006) Multicenter validation of a gene expression–based prognostic signature in lymph node–negative primary breast cancer. *J. Clin. Oncol.*, **24**.

Galea,M.H. *et al*. (1992) The nottingham prognostic index in primary breast cancer. *Breast Cancer Res. Treat.*, **22**, 207–219.

Gentleman,R. (2005) Reproducible research: a bioinformatics case study. *Stat. Appl. Genet. Mol. Biol.*, **4**.

Gerds,T.A. and Schumacher,M. (2001) On functional misspecification of covariates in the cox regression model. *Biometrika*, **88**, 572–580.

Gerds,T.A. and Schumacher,M. (2006) Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical J.*, **6**, 1029–1040.

Goldhirsh,A. *et al*. (2003) Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer. *J. Clin.Oncol.*, **21**, 3357–3365.

Graf,E. *et al*. (1999) Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.*, **18**, 2529–2545.

Haibe-Kains,B. *et al*. (2008) Computational intelligence in clinical oncology : lessons learned from an analysis of a clinical study. In Smolinski,T.G. *et al*. (eds) *Applications of Computational Intelligence in Biomedicine and Bioinformatics: Current Trends and Open Problems* of *Studies in Computational Intelligence*. Springer, Berlin/Heidelberg, pp. 237–268.

Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.

Harrell,F.E. *et al*. (1996) Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.*, **15**, 361–387.

Heagerty,P.J. *et al*. (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, **56**, 337–344.

Hedges,L. and Olkin,I. (1987) Statistical methods for meta-analysis. *J. Am. Stat. Assoc.*, **82**, 350–351.

Kaplan,E.L. and Meier,P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53**, 457–451.

Kittler,J. *et al*. (1998) On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, **20**, 226–238.

Lewis,S. and Clarke,M. (2001). Forest plots: trying to see the wood and the trees. *Brit. Med. J.*, **322**, 1479–1480.

Loi,S. *et al*. (2007) Definition of clinically distinct molecular subtypes in estrogen receptor positive breast carcinomas through use of genomic grade. *J. Clin. Oncol.*, **25**, 1239–1246.

Michiels,S. *et al*. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–492.

Miller,L.D. *et al*. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl Acad. Sci. USA*, **102**, 13550–13555.

Olivotto,I.A. *et al*. (2005) Population-based validation of the prognostic model adjuvant! for early breast cancer. *J. Clin. Oncol.*, **23**, 2716–2725.

Park,M.Y. and Hastie,T. (2007) L1 regularization path algorithm for generalized linear models. *J. R. Stat. Soc.*, **69**, 659–677.

Pencina,M.J. and D'Agostinno,R.B. (2004) Overall C as a measure of discrimination in survival analysis: model specic population value and condence interval estimation. *Stat. Med.*, **23**, 2109–2123.

Perou,C.M., *et al*. (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

R Development Core Team (2007) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Scarff,R.W. and Torloni,H. (1968) Histological typing of breast tumors. *International histological classification of tumours*, **2**, 13–20.

Schumacher,M. *et al*. (2007) Assessment of survival prediction models based on microarray data. *Bioinformatics*, **23**, 1768–1774.

Simon,R. (2005) Roadmap for developing and validating therapeutically relevant genomic classifiers. *J. Clin. Oncol.*, **23**, 7332–7341.

Sotiriou,C. and Piccart,M.J. (2007) Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nat. Cancer Rev.*, **7**, 545–553.

Sotiriou,C. *et al*. (2003) Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl Acad. Sci.*, **100**, 10393–10398.

Sotiriou,C. *et al*. (2006a) Comprehensive molecular analysis of several prognostic signatures using molecular indices related to hallmarks of breast cancer: proliferation index appears to be the most significant component of all signatures. In Lippman.M.E. (ed.) *Breast Cancer Research and Treatment*. Vol. 100. Springer, Netherlands, p. S86.

Sotiriou,C. *et al*. (2006b) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl Cancer Inst.*, **98**, 262–272.

Sotiriou,C., *et al*. (2007) Biological mechanisms that trigger breast cancer (bc) tumor progression are molecular subtype dependent. ASCO Annual Meeting Proceedings. *J. Clin. Oncol.*, **25**, 10581.

Sweets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Therneau,T.M. and Grambsch,P.M. (2000) *Modeling Survival Data: Extending the Cox Model*. In Gail,M. *et al*. (eds) Statistics for Biology and Health Series. Springer, New York.

Thomassen,M. *et al*. (2007) Comparison of gene sets for expression profiling: prediction of metastasis from low-malignant breast cancer. *Clin. Cancer Res.*, **13**, 5355–5360.

van de Vijver,M.J. *et al*. (2002) A gene expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

van Houwelingen,H. *et al*. (2006) Cross-validated cox regression on microarray gene expression data. *Stat. Med.*, **25**, 3201–3216.

van't Veer,L.J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Varma,S. and Simon,R. (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, **7**, 1471–2105.

Wang,Y. *et al*. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.

Wilcoxon,F. (1945) Individual comparisons by ranking methods. *Biometrics. Bull.*, **1**, 80–83.

Yu,J. *et al*. (2007) Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer*, **7**, 182.