Model selection for prognostic time-to-event gene signature discovery with applications in early breast cancer data

Miika Ahdesmäki[1], Lee Lancashire[1], Vitali Proutski[1], Claire Wilson[1], Timothy Davison[1], D. Paul Harkin[1,2], and Richard Kennedy[1,2]

[1]Almac Diagnostics, 19 Seagoe Industrial Estate, BT63 5QD Craigavon, United Kingdom
[2]Queen's University of Belfast, Centre for Cancer Research and Cell Biology, BT9 7BL Belfast, United Kingdom

August 15, 2013

Abstract

Model selection between competing models is a key consideration in the discovery of prognostic multigene signatures. The use of appropriate statistical performance measures as well verification of biological significance of the signatures is imperative to maximise the chance of external validation of the generated signatures. Current approaches in time-to-event studies often use only a single measure of performance in model selection, such as logrank test p-values, ordichotomise the follow- up times at some phase of the study to facilitate signature discovery. In this study we improve the prognostic signature discovery process through the application of the multivariate partial Cox model combined with the concordance index, hazard ratio of predictions, independence from available clinical covariates and biological enrichment as measures of signature performance. The proposed framework was applied to discover prognostic multigene signatures from early breast cancer data. The partial Cox model combined with the

multiple performance measures were used in both guiding the selection of the optimal panel of prognostic genes and prediction of risk within cross validation without dichotomising the follow- up times at any stage. The signatures were successfully externally cross validated in independent breast cancer datasets, yielding a hazard ratio of 2.55 [1.44, 4.51] for the top ranking signature.

1 Introduction

In cancer medicine it is increasingly appreciated that tumours arising from the same anatomical site in different patients can represent distinct diseases at a molecular level. Advances in mRNA, miRNA and DNA analysis have allowed the classification of tumours at a molecular level and have the promise of guiding personalised treatment strategies. The largest impact is likely to be in the discovery of prognostic assays, which predict outcome in the absence of a specific treatment, and predictive assays that predict the outcome following a specified therapy. Gene expression microarrays have been at the forefront of complex analysis of tumour biology as they are able to capture the relative expression of tens of thousands of genes simultaneously. In addition, they represent a mature diagnostic platform with demonstrated clinical applicability and suitable performance for in-vitro-diagnostic regulatory approval (Van't Veer, Dai, van de Vijver, He, Hart, Mao, Peterse, van der Kooy, Marton, Witteveen, Schreiber, Kerkhoven, Roberts, Linsley, Bernards, and Friend (2002), Pillai, Deeter, Rigl, Nystrom, Miller, Buturovic, and Henner (2011)).

In prognostic studies, where follow-up times are monitored instead of a binary outcome, the time to event data is typically analysed using Cox proportional hazards regression. An issue with this approach when applied to high throughput genomic data is multidimensionality where the number of genes analysed greatly exceeds the number of samples. Although modifications for analysing high dimensional data have been proposed for the Cox model (Boulesteix and Strimmer (2006), Li and Gui (2004), Gui and Li (2005), Witten and Tibshirani (2010)) and survival analysis for random forests (Ishwaran, Kogalur, Blackstone, and Lauer (2008)), time-to-event models for prediction have not been widely adopted in prognostic biomarker signature research. In fact, many prognostic signature discovery studies do not utilise time-to-event analysis algorithms at all (see e.g. Schmidt, Böhm, von Törne, Steiner, Puhl, Pilch, Lehr, Hengstler, Kölbl, and Gehrmann (2008)) or use Cox regression for ranking the genes but dichotomise

the follow-up times to estimate binary performance measures such as area under the ROC curve (see e.g. Wang, Klijn, Zhang, Sieuwerts, Look, Yang, Talantov, Timmermans, Meijer-van Gelder, Yu, Jatkoe, Berns, Atkins, and Foekens (2005)). In one exception, Kammers and co-authors used Lasso penalised Cox regression to analyse two microarray data sets and reported the prognostic index and Brier scores for the predictions (Kammers, Lang, Hengstler, Schmidt, and Rahnenführer (2011)).

The main contribution of this article is to introduce a completely continuous framework with performance measures for model selection that do not dichotomise the follow-up times at any stage. To this end, we use partial Cox regression combined with the concordance index (Harrell's c-index (Harrell, Jr. (2010), Raykar, Steck, Krishnapuram, Dehing-Oberije, and Lambin (2007))) and hazard ratio based performance evaluation of the risk scores and follow-up times for the discovery of the optimal panel of prognostic genes. The c-index has been recommended as a general measure of the predictive power of prognostic biomarkers (Newson (2006)). It estimates the probability of concordance between predicted and observed time to event, with 0.5 for random predictions and 1 for predictions matching the order of observed event times (Harrell, Jr. (2010)). For binary responses it is equivalent (Newson (2006)) to the area under the receiving operator characteristic curve (AUC), a frequently used measure in binary classification problems. In addition, when selecting for the optimal signature among the generated signatures, we analyse the results for biological relevance and independence from technical and clinical covariates thereby ensuring clinical applicability. We present semi-automatic signature generation methods that follow the general guidelines of the Microarray Quality Control (MAQC) Consortium and have been designed with clinical utility in mind. We use this methodology to discover multigene signatures that predict risk of distant metastasis in early breast cancer and successfully validate in independent datasets.

2 Methods

2.1 Data sets

Three datasets with an endpoint defined by breast cancer distant metastasis were downloaded. These datasets included the GSE11121[1] Mainz study (Schmidt et al.

[1] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse11121

(2008), also analysed in Kammers et al. (2011)), the GSE2034[2] Rotterdam study (Wang et al. (2005)) and the GSE7390[3] TRANSBIG breast cancer study (Desmedt, Piette, Loi, Wang, Lallemand, Haibe-Kains, Viale, Delorenzi, Zhang, d'Assignies, Bergh, Lidereau, Ellis, Harris, Klijn, Foekens, Cardoso, Piccart, Buyse, Sotiriou, and the TRANSBIG Consortium (2007)). In the TRANSBIG data, all events after 10 years were censored, as in the original publication. The effect of this additional censoring is studied in more detail in the results section. Each of these three data sets were used once as training sets and the two remaining ones as external validation sets, composing a 3-by-3 table of results (Table 2). The samples were all profiled on the Affymetrix HG-U133A platform, containing 247965 probes that can be summarised further to 22283 probesets. These probesets roughly map to one or more human transcripts and represent a snapshot of expressed transcripts in the samples.

A summary of the data sets is given in Table 1. Note that the number of events for the TRANSBIG data was 62 in the original database but after censoring samples with more than 10 years of follow-up (as suggested in Desmedt et al. (2007)) this was reduced to 50. In the available clinical covariates considered for confounding effects, missing values were replaced by median of the present values. Multilevel nominal covariates were represented by $N-1$ binary indicators, where N is the number of levels in a given covariate. Covariates with underrepresented levels were discarded. In the clinical data for the TRANSBIG study, lymphocytic infiltration and angioinvasion contained too many missing values and were excluded. All of the breast cancer samples in these datasets were lymph node negative.

2.2 Pre-processing and exploratory analysis

To enable a multisample, multivariable analysis of the microarray data sets, preprocessing of the data including background correction, normalisation and probeset summarisation was first considered to ensure comparability of the gene expression levels. One option would be to combine all the chosen data sets and pre-process them together. However, due to the chosen pre-processing being multichip (see below), this approach would not be feasible if the signature was taken to clinic due to new samples coming in one at a time. Using validation samples together with the model training samples in multichip pre-processing would

[2]http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse2034

[3]http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse7390

Table 1: Summary of the data sets used in this paper. ER+ stands for estrogen receptor positive indicator. The data set identifiers are further explained under subsection Data sets.

Data set	Samples	Events	Median follow-up time	Clinical covariates
GSE2034 HG-U133A Rotterdam	286	107	86 mo	ER+ Brain relapse
GSE7390 HG-U133A TRANSBIG	198	50	144.1 mo	Hospital Age Size Surgery Histology Grade ER+
GSE11121 HG-U133A Mainz cohort	200	46	90.5 mo	Grade Size

also introduce a positive bias similar to using validation samples as part of a training set when building a predictive model. All samples were independently background corrected per array using the per-sample background correction algorithm in Robust Multichip Average (RMA, Bolstad, Irizarry, Åstrand, and Speed (2003)), the widely accepted pre-processing tool for gene expression microarray data. Each of the training sets were quantile normalised and median polish summarised using RMA. The quantiles from training set normalisation and probe affinities from summarisation were then applied to the corresponding external validation or cross-validation test sets one sample at a time (Reference RMA, Katz, Irizarry, Lin, Tripputi, and Porter (2006)). Affymetrix control probesets were discarded after probeset summarisation.

The concept of classification difficulty estimation was introduced in Popovici, Chen, Gallas, Hatzis, Shi, Samuelson, Nikolsky, Tsyganova, Ishkin, Tatiana Nikolskaya and, Valero, Booser, Delorenzi, Hortobagyi, Shi, Symmans, and Pusztai (2010) for the purpose of exploring associations between variables and a binary endpoint. In brief, for each probeset a squared t-score is evaluated and the cumulative sum of the ordered squares is plotted to evaluate whether or not there is a strong association between the endpoint values and the data. For time-to-event analysis, we instead plot the cumulative sum of the ordered absolute univariate

Cox coefficients. The cumulative sum is additionally compared to a negative control distribution obtained by permuting the follow-up time values of the samples along with their event indicators. In general, in both classification and Cox model difficulty estimation the resulting curve is monotonously increasing approaching a plateau after the most informative variables have been added (see Figure 2 for an example). For the Cox model difficulty estimation the RMA pre-processed probesets were filtered by 50% based on low variance and intensity (see Hackstadt and Hess (2009) for motivation) as within the actual signature discovery, see details below.

2.3 Prognostic signature generation

In evaluating the generalisation performance and optimal length of the signatures, each of the training data sets was split into 5 folds of cross-validation repeated 10 times, with stratification such that the proportion of events and distributions of follow-up times were approximately equal between each fold of the cross-validation training sets and the full training set. Each cross-validation training set was normalised (quantile normalisation, storing the training set quantiles) and summarised (median polish, storing the estimated probe affinities) and the obtained pre-processing models were applied to the cross-validation test sets. Analogously, the ref-RMA pre-processing models from the full training sets were eventually used to pre-process the validation sets sample by sample. It is emphasised that cross-validation was used to guide in the selection of the optimal signature length only and not for the training of any final signature per se. The final signatures were trained by repeating the feature selection process in the full training sets (see Simon (2012) for motivation).

Within each cross-validation training set, the probesets were filtered by 50% based on variance and intensity (Hackstadt and Hess (2009)). Average rank of high variance and high intensity was used in the filter, retaining the highest ranked variables. To further lessen the computational burden within feature selection and to reduce the number of null variables, the probesets were further filtered down to approximately 1000 probesets using a univariate Cox filter, based on the Cox proportional hazards (PH) model coefficients from univariate analyses of the probesets. This second filtering step is by no means obligatory and the presented feature selection method is not limited to starting with 1000 features should sufficient computational resources be available. Both of these hard thresholds (50% followed by down to approximately 1000 covariates) were chosen instead of for example controlling for false discovery rate (FDR) in the filtering to ensure a con-

sistent starting size of probesets for the feature selection for all folds and repeats of cross-validation. For the remaining probesets, an iterative backwards elimination feature selection procedure was applied using 3-component partial Cox regression (Li and Gui (2004)), where the partial Cox coefficients of the probesets were used for ranking (see Figure 1). During each iteration, 10% of the lowest ranking probesets were discarded after predicting the cross-validation test set risk scores (see Eq. 1 for the definition of the risk score). No information leakage was allowed from the cross-validation test set, i.e. the probeset ranking was purely based on the coefficients from the cross-validation training set. This procedure was repeated until five probesets remained. The c-index (Harrell, Jr. (2010), see Eq. 4) of the continuous cross-validation test set risk score predictions was evaluated as the main performance measure. Additionally, univariate hazard ratios of the cross-validation test set risk scores, dichotomised at risk score value 0 (Eq. 2, i.e. low risk < 0, high risk > 0, Li and Gui (2004)), were evaluated. The process is summarised in Figure 1.

When deciding on the signature length, the signature lengths that maximise the HR and c-index in cross-validation may not necessarily be the same. To aid in model selection, independence to clinical covariates was also evaluated within cross-validation. A generalised Cox PH likelihood ratio test (Eq. 5) was used to this end in which a full model with dichotomised predictions and clinical covariates is compared to a reduced model with the clinical covariates only in predicting time to event.

In addition, signatures with the minimal number of probesets possible whilst maintaining performance were favoured. We suggest that the feature length selection step requires human guidance as different objectives could influence the selection of the feature length, such as desire to migrate to another platform, error interval width and so on. Additionally, in the case where the selection is somewhat arbitrary, due to the similarity of the different performance measures, any feature length with comparable performance could be selected, and the one chosen and reported on here is only as an example. As a compromise, maximum signature length was set to 600 (Kennedy, Bylesjo, Kerr, Davison, Black, Kay, Holt, Proutski, Ahdesmaki, Farztdinov, Goffard, Hey, McDyer, Mulligan, Mussen, O'Brien, Gavin, Oliver, Walker, Mulligan, Wilson, Winter, O'Donoghue, Mulcahy, O'Sullivan, Sheahan, Hyland, Dhir, Bathe, Winqvist, Manne, Shanmugam, Ramaswamy, Leon, Jr, McDermott, Wilson, Longley, Marshall, Cummins, Sargent, Johnston, and Harkin (2011) used 634 probesets) and the signature length below 600 that maximised the c-index was chosen unless other criteria showed poor results for the same length.

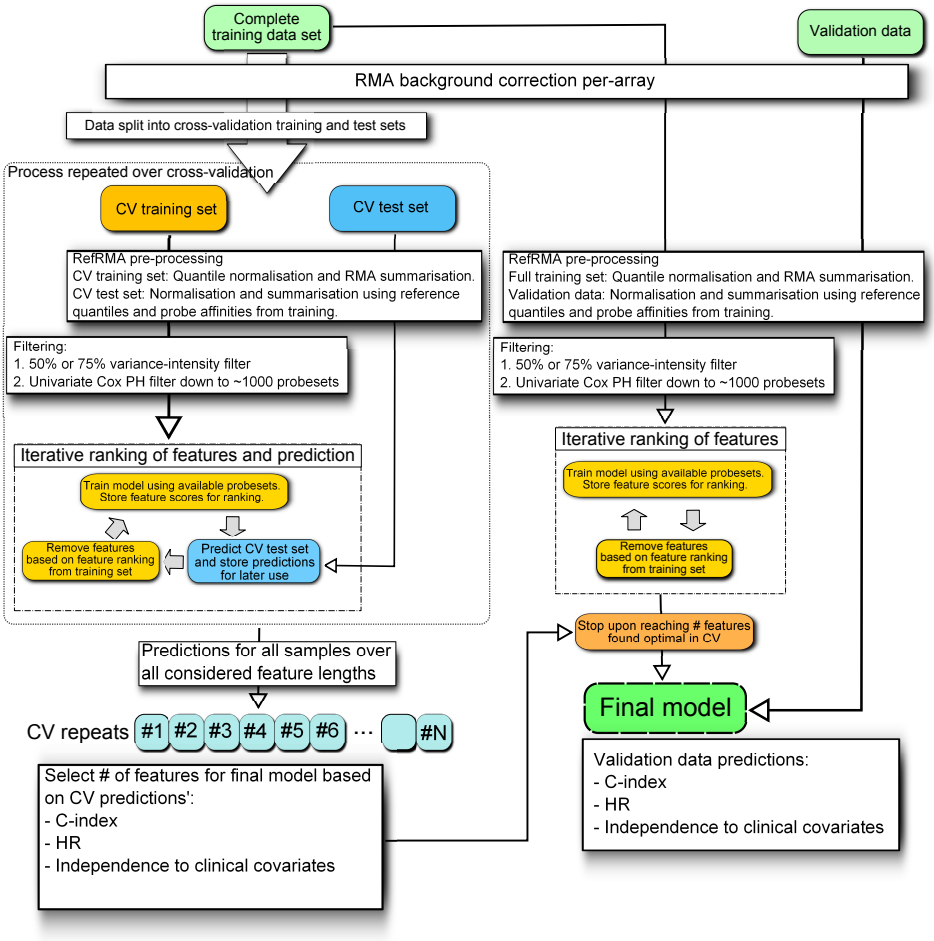


Figure 1: Flowchart of the signature generation and evaluation process.

To sum up the model selection criteria and to aid the more inexperienced researchers in comparing competing models, we list here our recommendations for selecting between models, in order of importance:

1. Select the signature that maximises the (average) concordance index within cross-validation and whose error bars or confidence interval does not overlap the 50% limit.

2. Select the signature that maximises the hazard ratio and whose error bars or confidence interval does not overlap one.

3. If clinical and / or technical confounder information is available, select the signature whose p-value in a Cox model likelihood ratio test of independence to the confounders is the lowest.

4. If none of the above criteria can distinguish between the highest ranking signatures, observe if biological processes related to the disease under study are more enriched in one signature versus another.

Our recommendations are based on our extensive experience in working with predictive classifier models where the analysis of the models is driven by AUC, whose extension into the time-to-event case the c-index is. However it is often required to select a signature score dichotomisation threshold for binary classifier models afterwards to enable estimation of sensitivity, specificity, NPV and / or PPV. In these cases the serial process is AUC based model selection followed by threshold selection driven by clinical utility (e.g. a fixed specificity). In the time-to-event space this is akin to using the c-index as a primary measure for model selection followed by HR estimation. In both domains a likelihood ratio test can be used to assess independence from clinical and / or technical covariates (logistic regression in binary classification and Cox model in time-to-event).

The partial Cox regression algorithm (Li and Gui (2004)) was chosen for the biomarker discovery analyses because it is theoretically based on the same idea as partial least squares (PLS) regression, an established method in high dimensional analyses (Boulesteix and Strimmer (2006)). It is also analogous to principal components analysis in that the first few latent components explain most of the information in the data. Additionally, partial Cox requires no parameter tuning in nested cross-validation and therefore the computational complexity is modest compared to the Lasso Cox approach taken by the authors in Kammers et al. (2011).

Due to the public unavailability of the original implementation, Partial Cox regression will be made available in R from the package `PartialCox`. The c-index and many other useful functions are available in the package `Hmisc` .

2.4 Definitions of important parameters and statistics

The risk score used in this publication is defined as the linear combination of the signature probeset values multiplied by their corresponding partial Cox model coefficients, first subtracted by the training set probeset mean values (see Li and

Gui (2004) for the Cox model and risk score definitions):

$$y^{new} = (\mathbf{x}^{new} - \overline{\mathbf{x}}^{train})'\hat{\beta}, \tag{1}$$

where y^{new} is the continuous risk score for the new sample, $\mathbf{x}^{new}$ is the vector of individual probeset values for the new sample, $\overline{\mathbf{x}}^{train}$ is the vector of probeset mean values from the training set and $\hat{\beta}$ are the partial Cox model coefficients from training.

The risk scores have by definition a sample mean of zero Li and Gui (2004), and dichotomisation of the risk scores for hazard ratio calculation is obtained via the indicator function,

$$I_{y^{new}>0}, \tag{2}$$

which equatos to one for positive values and zero otherwise.

The hazard ratio of predictions is given by

$$e^{b_p}, \tag{3}$$

where b_p is the Cox model coefficient from a Cox model when modelling survival time by the dichotomised signature risk scores.

The concordance index is best understood by considering the ordered follow-up times as a directed graph, as in Raykar et al. (2007). The total number of edges in the ordered follow-up time graph depends mainly on the number of events, as an edge is drawn only from events to any event or censoring having a follow-up time greater than that of the event considered. The concordance index is obtained by counting the number of edges in the graph where the predicted survival times for the samples agree with the direction of the edges, divided by the total number of edges:

$$c = \frac{1}{|\varepsilon|} \sum_{T_i \text{ uncensored}} \sum_{T_j > T_i} 1_{f(x_i)<f(x_j)}, \tag{4}$$

where ε denotes the total number of edges in the graph, T_i are the follow-up times, $f(x_i)$ are the predicted survival times and $1_{f(x_i)<f(x_j)}$ is an indicator variable that is one when $f(x_i) < f(x_j)$ is true, zero otherwise. The risk scores have an inverse relationship with predicted survival times, i.e. high risk score implies low predicted survival time. Therefore, for risk scores, $f(x_i) < f(x_j)$ in Eq. 4 needs to be replaced by $f_{risk}(x_i) > f_{risk}(x_j)$.

The likelihood ratio statistic used in evaluating the effect of clinical covariates (confounders) is given by

$$-2\left(\ln \hat{L}_{reduced} - \ln \hat{L}_{full}\right) \sim \chi^2_1 \tag{5}$$

where the Cox model log-likelihoods ($\ln \hat{L}$) and chi-square distribution values can readily be obtained from standard statistical software packages. The degrees of freedom default to one as the model order difference is always one here.

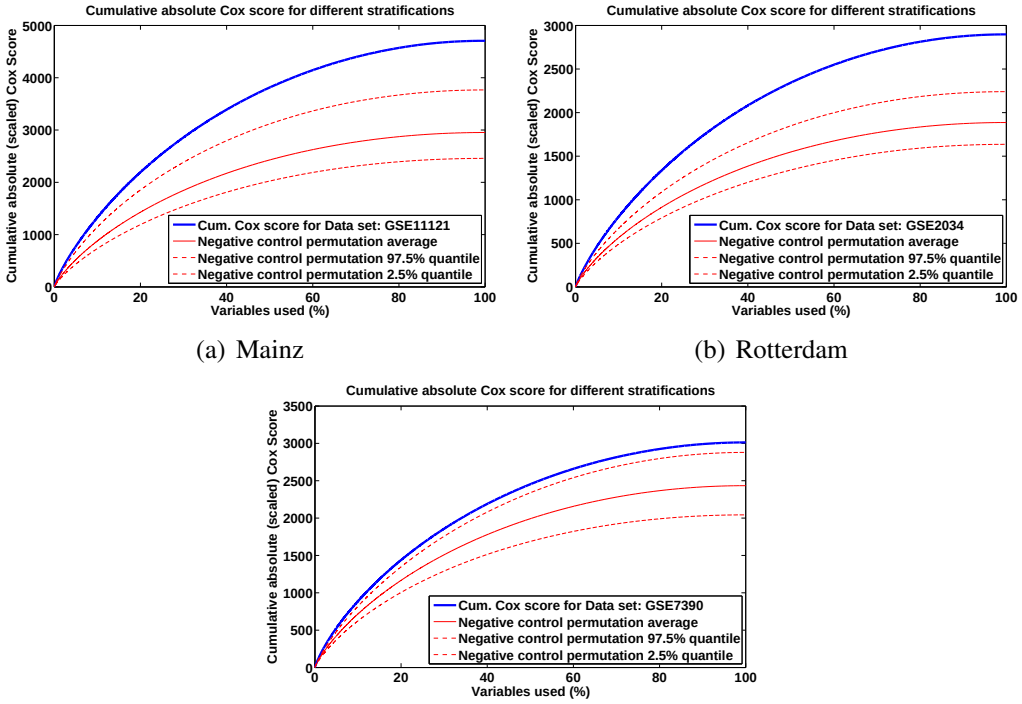
3 Results

3.1 Prognostic signature generation results

Three datasets were selected to test our methodology. The Mainz dataset (Schmidt et al. (2008)) consists of microarray data from 200 lymph node negative, ER positive (ER+ve) and negative (ER-ve) patients who did not receive systemic adjuvant therapy. The number of metastatic recurrences recorded was 46, 18 of which were beyond five years. The Rotterdam dataset (Wang et al. (2005)) consists of microarray data from 286 lymph node–negative ER+ve and ER-ve patients who did not receive adjuvant systemic therapy. Ninety three metastatic recurrences were reported, although no data is available after five years. Finally the TRANSBIG study (Desmedt et al. (2007)) consists of microarray data from 198 ER+ve and ER-ve patients who also did not receive adjuvant systemic therapy after surgery. Fifty patients developed metastatic recurrence before ten years of which 14 occurred after five years. Importantly, these datasets are entirely independent and therefore ideal for testing our methodology for identifying and validating clinically meaningful prognostic signatures. Although microarray data is used here solely, the proposed methodology is applicable to other high dimensional data equally well.

Analysis of associations between predictors (probesets in case of Affymetrix microarrays) and time to distant metastasis for each of the three training sets revealed that the association between the data and survival time (time to distant metastasis) is stronger than expected by chance. This is shown by the thick solid curve being above the 97.5% quantile (upper dashed line) of the negative control distribution (Figure 2).

Signature generation and evaluation was then performed (summarised in Figure 1). The test set prediction results from all the ten cross-validation repeats are shown in Figures 3 (c-index), 4 (HR) and 5 (minus $\log_{10}$ p-values of independence to clinical covariates) for the different training data sets. A decreasing trend in performance can be observed towards the shorter signature lengths, with the minus log p-values (Figure 5) following the same trends as the HR (Figure 4). There was clearly no steep drop in performance at any signature length, therefore leading to an amount of redundancy in selecting the optimal feature length. As



(a) Mainz (b) Rotterdam

(c) TRANSBIG

Figure 2: Cox model generation difficulty estimate plots. X-axis shows the percentage probesets included in the cumulative sum. Permutation based negative control distribution quantiles are plotted in dashed lines.

different studies have different objectives, e.g. it might be required to optimise the signature for maximum hazard ratio, to minimise error bars or limit the number of genes to less than 20 to enable migration to qPCR, all this information must be considered in a human decision guided fashion. Here emphasis was placed such that signature lengths that maximise the c-index below 600 variables are highlighted for the results from the Mainz and Rotterdam datasets, whereas for the TRANSBIG results the highlighted signature length (252) maximises HR and gives a good compromise on c-index.

Final prognostic signatures for each data set were generated by repeating the iterative feature selection process on the respective full data set. Final partial Cox models were trained at the chosen feature lengths (Mainz: 531, Rotterdam: 386, TRANSBIG: 252). For average performance of signatures of these lengths in cross-validation see the diagonal entries of Table 2. The overlap of the final

signature probesets for the three breast cancer signatures is depicted in a Venn diagram (Figure 6). The small observed overlap in the probesets comprising the three signatures shows that there is clearly some redundancy in which probesets to choose to predict the same endpoint, as the signatures largely validate across the data sets (see Prognostic signature validation results below).

Table 2: Performance measures for the breast cancer distant metastasis cross-validation test (diagonal) and validation set (off-diagonal) predictions. 'C' denotes concordance index, 'HR' hazard ratio, 'CI' the 95% confidence interval (validation sets only), 'HR p-val' the p-value for testing HR=1 (validation sets only) and 'PH p-val' the p-value for testing the proportional hazards assumption with a low value implying a violation (validation sets only). The values on the diagonal represent average performance in cross-validation.

Data	Model	Mainz 531PS	Rotterdam 386PS	TRANSBIG 252PS
Mainz	C [CI]:	0.671	0.629 [0.540, 0.708]	0.613 [0.517, 0.703]
	HR [CI]:	2.75	2.04 [1.13, 3.68]	1.66 [0.86, 3.20]
	HR p-val:		0.018	0.134
	PH p-val		0.027	0.005
Rotterdam	C [CI]:	0.617 [0.562, 0.667]	0.667	0.610 [0.555, 0.659]
	HR [CI]:	2.09 [1.37, 3.19]	2.56	1.71 [1.00, 2.91]
	HR p-val:	0.0006		0.049
	PH p-val	0.084		0.112
TRANSBIG	C [CI]:	0.648 [0.576, 0.719]	0.654 [0.581, 0.723]	0.669
	HR [CI]:	2.55 [1.20, 5.43]	2.55 [1.44, 4.51]	3.85
	HR p-val:	0.015	0.001	
	PH p-val	0.031	0.305	

Indeed the most frequently observed Gene Ontology biological processes shown to be statistically significant ($p < 0.01$) in all three signatures were cell cycle processes, RNA splicing and metabolic processes, all of which may have implications in cancer development. The results for the functional analysis using the KEGG database revealed a number of pathways of interest including the erythropoietin (EPO) signalling pathway which has been shown to influence numerous cellular functions including proliferation, apoptosis, and drug resistance, all of which could possibly contribute towards decreased survival (Hedley, Allan, and Xenocostas (2011)). Genes involved in MAP kinase signalling were also significantly enriched in these signatures. Abnormalities in MAP kinase have been shown to

affect most cellular processes required by tumours in order to survive, and thus play a critical role in the development and progression of cancer (Dhillon, Hagan, Rath, and Kolch (2007)).

3.2 Prognostic signature validation results

The final models were applied to the independent validation data sets, yielding c-index values and uncorrected hazard ratios as shown in the off-diagonal entries of Table 2. Bootstrap quantile-based 95% confidence intervals were calculated for the c-index. The confidence intervals for HR were estimated using a 1.96 standard error interval around the $\ln(HR)$. In terms of the HR confidence intervals not containing the value 1, the signatures from the Mainz and Rotterdam datasets validate externally. None of the c-index confidence intervals contain the value 0.5 and therefore all of the signatures validate based on c-index alone.

Correction of the hazard ratios for clinical covariates was not feasible due to different sets of covariates in the three data sets. However, an assessment of the clinical covariate confounding for the validation set predictions is summarised in Table 3 in terms of Cox model likelihood ratio test p-values. The p-values for the external validation set predictions were 0.003 and 0.101 when using the signature derived from the Mainz data set in predicting Rotterdam and TRANSBIG, respectively. When using the signature derived from the Rotterdam data, the p-values were 0.093 and 0.002 for Mainz and TRANSBIG. The higher (0.101 and 0.093) p-values are concordant with low proportional hazards assumption check p-values that indicate a violation of the proportional hazards assumption for these combinations of signatures and validation data; see results below and the discussion section for the implications. The p-values for the predictions from the TRANSBIG dataset model were well above 0.05 and therefore independence to clinical covariates could not be shown given the samples size.

The proportional hazards assumption was verified for the dichotomised validation set predictions. The p-values for checking this assumption were evaluated using the correlation between the scaled Schoenfeld residuals for the dichotomised predictions and the ranking of individual follow-up times. Table 2 shows these p-values were below 0.05 (indicating a violation) when using either of the Rotterdam or TRANSBIG dataset signatures to predict the Mainz data, as well as when using the Mainz dataset signature in predicting risk of distant metastasis in the TRANSBIG data.

The Kaplan-Meier plots of the validation set predictions are shown in Figure 7 for the three breast cancer signatures.

The effect of varying the definition of censoring in the TRANSBIG data set on the resulting hazard ratios was also briefly considered. Reducing the censoring threshold from ten to five years increased the HR to 5.12 and the c-index to 0.667, compared to 2.55 and 0.617 when using the recommended censoring threshold of ten years. Using the data without censoring any events, the HR was just above 1 and the c-index was 0.602. See Figure 8 for Kaplan-Meier plots when varying the censoring threshold.

4 Discussion

In this study we have taken a novel approach to generate multigene signatures using time to event data combined with comprehensive model selection criteria. Primary emphasis in the selection of signatures was placed on the concordance index and the simple univariate hazard ratio of the cross-validation test set predictions, supported by evaluations of biological enrichment and independence from available clinical covariates.

Using the Mainz dataset for training, a prognostic signature of 531 probesets was chosen based on performance under cross-validation (see Figures 3-5). This signature validated in the Rotterdam dataset with a HR for risk of distant metastasis following surgery of $2.09[1.37, 3.19]$ (p-value = 0.0006) and in the TRANSBIG dataset with a HR of $2.55[1.20, 5.43]$ (p-value = 0.0154). Importantly, due to our methodology, the signature performance was independent from known prognostic factors such as tumour size, grade, ER status and age (p-values for predictions in Table 3, column 2, below 0.05). The performance characteristics of our signature compares favourably to a B-cell metagene that yielded a HR of 1.28 on Rotterdam data high-grade tumours and HR of 1.2 on the TRANSBIG data set enriched for younger patients (Schmidt et al. (2008)).

The Rotterdam dataset was also used to generate a signature of 386 probesets. This signature validated with a HR for metastatic recurrence of $2.04[1.13, 3.68]$, (p-value = 0.0175) and $2.55[1.44, 4.51]$, (p-value = 0.0013) in the Mainz and TRANSBIG datasets respectively. This signature was also independent from known prognostic factors in the TRANSBIG dataset (Table 3), indicating clinical utility.

The validation performance of the signatures from both the Mainz and Rotterdam datasets compare favourably with prognostic signatures that have entered clinical practice; the MammaPrint 70 gene signature (Van't Veer et al. (2002)) yielded an unadjusted HR of 2.32 in external validation (TRANSBIG data, Buyse,

Loi, van't Veer, Viale, Delorenzi, Glas, d'Assignies, Bergh, Lidereau, Ellis, Harris, Bogaerts, Therasse, Floore, Amakrane, Piette, Rutgers, Sotiriou, Cardoso, Piccart, and On behalf of the TRANSBIG Consortium (2006)) and Oncotype DX achieved an adjusted HR of 2.81 (Paik, Shak, Tang, Kim, Baker, Cronin, Baehner, Walker, Watson, Park, Hiller, Fisher, Wickerham, Bryant, and Wolmark (2004)), thereby demonstrating the potential clinical utility of the prognostic signatures generated using our approach.

A 252 probeset signature was generated from the TRANSBIG dataset. The performance for this signature, however, was inferior to the other two signatures developed here when evaluated in the Mainz and Rotterdam datasets with HR evaluated at $1.66[0.86, 3.20]$, (p-value = 0.1341) and $1.71[1.00, 2.91]$, (p-value = 0.0487) respectively. The performance was independent from grade and tumour size in the two datasets (Table 3). Although it is unclear why this signature failed to perform as well as the other two, it is interesting to note that the mean patient age was 46 in TRANSBIG versus 60 and 54 in the Mainz and Rotterdam datasets respectively. The greater representation of younger women in the TRANSBIG training set may result in over representation of specific molecular subtypes such as Basal-like tumours (Millikan, Newman, Tse, Moorman, Conway, Smith, Labbok, Geradts, Bensen, Jackson, Nyante, Livasy, Carey, Earp, and Perou (2008)) thereby reducing the ability of the signature to validate in older populations. Interestingly The Kaplan-Meier plot in Figure 7(a) demonstrates that this signature predicts the outcome in the Mainz dataset until five years after which the performance drops. This may indicate that this signature does not capture the biology representing late recurrence adequately and highlights the risks in dichotomising a population based on an arbitrary time point to develop a prognostic signature.

One of the key assumptions in the Cox model is that of proportional hazards (PH). In a regression setting this means that the survival and hazard curves for the predicted risk groups must be parallel over time, i.e. these curves must not cross. For a quantitative test, we used the Schoenfeld residuals from the Cox model to test this proportionality, where a p-value < 0.05 (relationship between residuals and time) means that the proportional hazards assumption should be rejected and we cannot rely on the ratio of the hazard functions being accurate. Therefore, when these assumptions are violated, alternative measures such as the c-index give a more accurate measure of signature performance in this setting, since it does not rely on these assumptions.

The PH assumption check (Table 2) confirms that in some cases there was a reason to doubt the assumption with respect to the predicted dichotomised risk. The PH test p-values for the dichotomised predictions from the Rotterdam and

TRANSBIG data based signatures were below 0.05 for prediction upon the Mainz patient samples. Similarly, the PH test p-value when using the Mainz data based signature in prediction of the TRANSBIG data was smaller than 0.05. In these cases the c-index is likely to give a better view of the predictive performance as it is a rank correlation based measure and does not depend on the proportional hazards assumption. This is clearly illustrated in the Kaplan-Meier plot in Figure 8(a), where the signature initially stratified the patients well (c-index greater than 0.6) but the PH assumption is violated and HR is very close to one. Consequently we advocate that hazard ratios should always be accompanied by the corresponding PH check p-values and c-index values. In all of the three breast cancer signatures the average c-index under cross-validation was approximately 0.67 and the validation set c-index values between 0.61 and 0.65, showing only a minor positive bias between cross-validation and independent validation. Factors influencing the observed minor drop from cross-validation to validation may include differences in population, laboratories, reagents and temporal differences.

5 Conclusions

In this paper we presented rigorous model selection criteria and developed a semi-automated signature generation framework that was used to discover prognostic multigene assays using time to event data. To facilitate model selection between different signature lengths and conditions, we applied four model selection criteria in parallel, namely the concordance index, hazard ratio, independence from clinical covariates and biological enrichment of the signature genes. Considering multiple criteria in parallel aided in selecting signatures that validated externally, as is evident from the results across the three breast cancer data sets. We now plan to test this methodology using multi-analyte data generated by other technologies that are entering mainstream molecular profiling such as next-generation sequencing, metabolomics arrays and protein arrays, and also work on shortening the signatures using forward feature selection methods.

References

Bolstad, B. M., R. A. Irizarry, M. Åstrand, and T. P. Speed (2003): "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, 19 (2), 185–193.

Boulesteix, A.-L. and K. Strimmer (2006): "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data," *Briefings is Bioinformatics*, 8, 32–44.

Buyse, M., S. Loi, L. van't Veer, G. Viale, M. Delorenzi, A. M. Glas, M. S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. Harris, J. Bogaerts, P. Therasse, A. Floore, M. Amakrane, F. Piette, E. Rutgers, C. Sotiriou, F. Cardoso, M. J. Piccart, and On behalf of the TRANSBIG Consortium (2006): "Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer," *J Natl Cancer Inst*, 98(17), 1183–92.

Desmedt, C., F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M. S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. L. Harris, J. G. Klijn, J. A. Foekens, F. Cardoso, M. J. Piccart, M. Buyse, C. Sotiriou, and the TRANSBIG Consortium (2007): "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series," *Clin Can Res*, 13, 3207–3214.

Dhillon, A. S., S. Hagan, O. Rath, and W. Kolch (2007): "MAP kinase signaling pathways in cancer," *Oncogene*, 26, 3279–3290.

Gui, J. and H. Li (2005): "Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data," *Bioinformatics*, 21(13), 3001–3008.

Hackstadt, A. J. and A. M. Hess (2009): "Filtering for increased power for microarray data analysis," *BMC Bioinformatics*, 10, 11.

Harrell, Jr., F. E. (2010): *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, Springer.

Hedley, B. D., A. L. Allan, and A. Xenocostas (2011): "The role of erythropoietin and erythropoiesis-stimulating agents in tumor progression," *Clin Cancer Res*, 17(20), 6373–6380.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer (2008): "Random survival forests," *Annals of Applied Statistics*, 2(3), 841–860.

Kammers, K., M. Lang, J. G. Hengstler, M. Schmidt, and J. Rahnenführer (2011): "Survival models with preclustered gene groups as covariates," *BMC Bioinformatics*, 12, 478.

Katz, S., R. A. Irizarry, X. Lin, M. Tripputi, and M. W. Porter (2006): "A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database," *BMC Bioinformatics*, 7, 464.

Kennedy, R. D., M. Bylesjo, P. Kerr, T. Davison, J. M. Black, E. W. Kay, R. J. Holt, V. Proutski, M. Ahdesmaki, V. Farztdinov, N. Goffard, P. Hey, F. McDyer, K. Mulligan, J. Mussen, E. O'Brien, Gavin, Oliver, S. M. Walker, J. M. Mulligan, C. Wilson, A. Winter, D. O'Donoghue, H. Mulcahy, J. O'Sullivan, K. Sheahan, J. Hyland, R. Dhir, O. F. Bathe, O. Winqvist, U. Manne, C. Shanmugam, S. Ramaswamy, E. J. Leon, W. I. S. Jr, U. McDermott, R. H. Wilson, D. Longley, J. Marshall, R. Cummins, D. J. Sargent, P. G. Johnston, and D. P. Harkin (2011): "Development and independent validation of a prognostic assay for stage II colon cancer using formalin-fixed paraffin-embedded tissue," *J Clin Oncol.*, 29(35).

Li, H. and J. Gui (2004): "Partial Cox regression analysis for high-dimensional microarray gene expression data," *Bioinformatics*, 20 Suppl. 1, i208–i215.

Millikan, R., B. Newman, C.-K. Tse, P. Moorman, K. Conway, L. Smith, M. Labbok, J. Geradts, J. Bensen, S. Jackson, S. Nyante, C. Livasy, L. Carey, H. S. Earp, and C. Perou (2008): "Epidemiology of basal-like breast cancer," *Breast Cancer Res Treat*, 109(1), 123–139.

Newson, R. (2006): "Confidence intervals for rank statistics: Somers' d and extensions," *The Stata Journal*, 6(3), 309–334.

Paik, S., S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park, W. Hiller, E. R. Fisher, D. L. Wickerham, J. Bryant, and N. Wolmark (2004): "A multigene assay to predict recurrence of Tamoxifen-treated, node-negative breast cancer," *N Engl J Med*, 351, 2817–2826.

Pillai, R., R. Deeter, C. Rigl, J. Nystrom, M. H. Miller, L. Buturovic, and W. Henner (2011): "Validation and reproducibility of a microarray-based gene expression test for tumor identification in formalin-fixed, paraffin-embedded specimens," *The Journal of Molecular Diagnostics*, 13(1), 48–56.

Popovici, V., W. Chen, B. D. Gallas, C. Hatzis, W. Shi, F. W. Samuelson, Y. Nikolsky, M. Tsyganova, A. Ishkin, K. R. H. Tatiana Nikolskaya and, V. Valero, D. Booser, M. Delorenzi, G. N. Hortobagyi, L. Shi, W. F. Symmans, and L. Pusztai (2010): "Effect of training-sample size and classification difficulty on the accuracy of genomic predictors," *Breast Cancer Res*, 12, (1):R5.

Raykar, V. C., H. Steck, B. Krishnapuram, C. Dehing-Oberije, and P. Lambin (2007): "On ranking in survival analysis: Bounds on the concordance index," in *NIPS*.

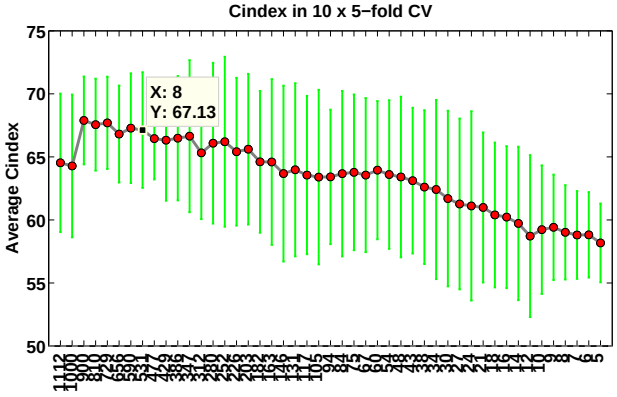
Schmidt, M., D. Böhm, C. von Törne, E. Steiner, A. Puhl, H. Pilch, H.-A. Lehr, J. G. Hengstler, H. Kölbl, and M. Gehrmann (2008): "The humoral immune system has a key prognostic impact in node-negative breast cancer," *Cancer Res*, 68, 5405.

Simon, R. (2012): "Clinical trials for predictive medicine," *Statistics in Medicine*, 31, 3031–3040.

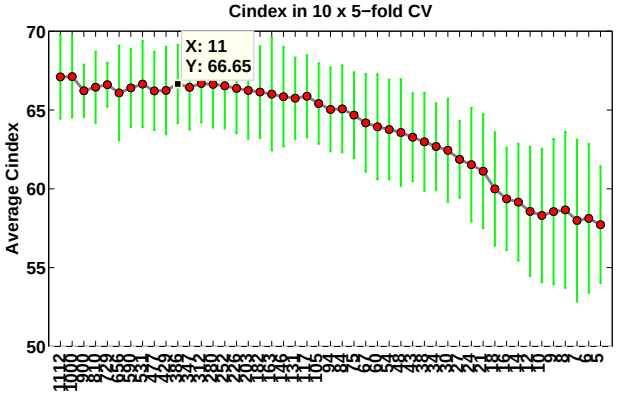
Van't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend (2002): "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, 415, 530–536.

Wang, Y., J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens (2005): "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, 365, 671–679.

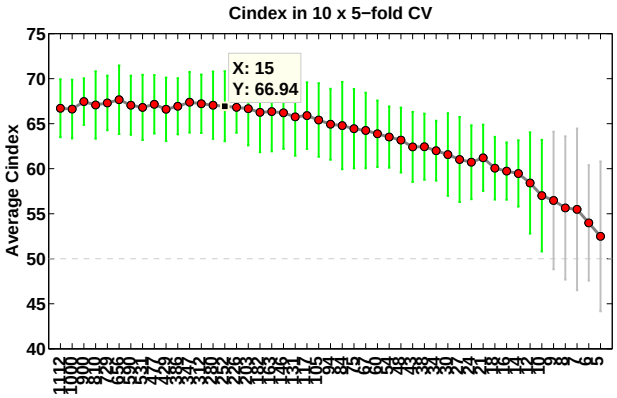
Witten, D. and R. Tibshirani (2010): "Survival analysis with high-dimensional covariates," *Statistical Methods in Medical Research*, 19, 29.



(a) Mainz

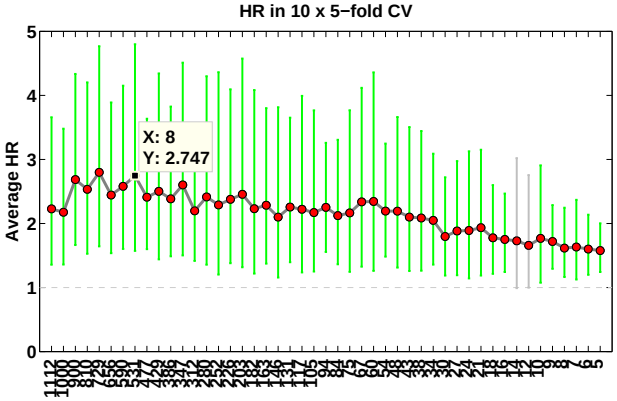


(b) Rotterdam

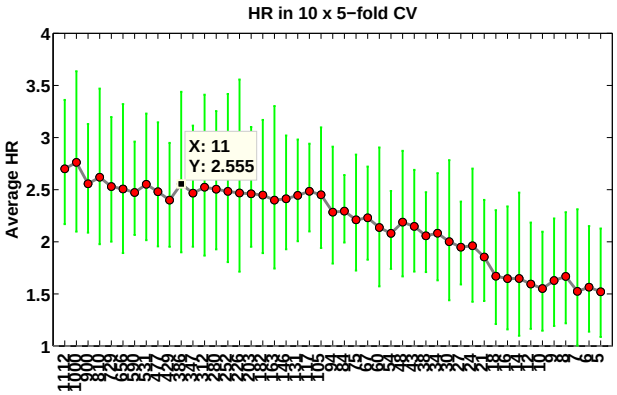


(c) TRANSBIG

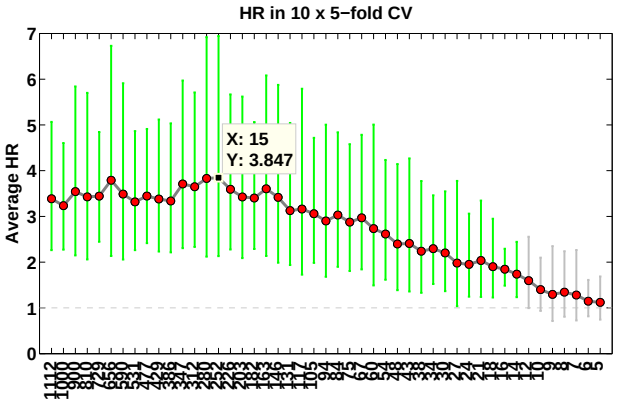
Figure 3: Concordance index summaries over considered feature lengths and cross-validation repeats. Means (circles) of the ten cross-validation repeats are plotted with the two standard deviation (2SD) based prediction intervals. Dashed gray line corresponds to concordance index of 0.5. Y-axis: c-index for the cross-validation test set predictions. X-axis: Probeset lengths evaluated.



(a) Mainz

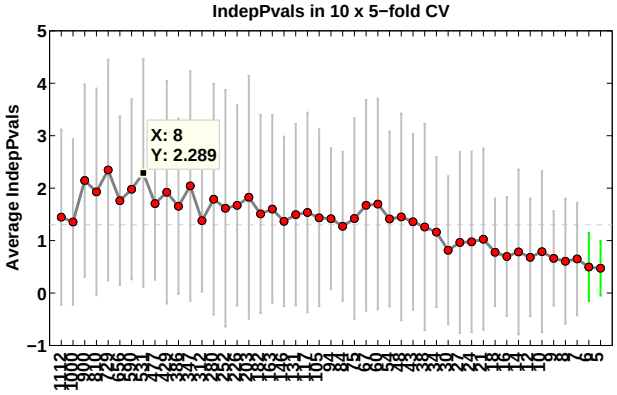


(b) Rotterdam

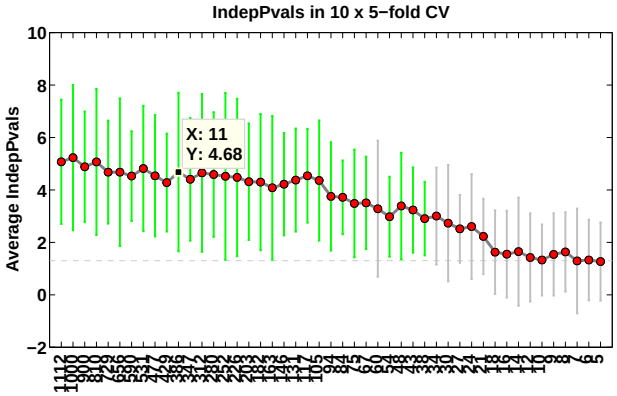


(c) TRANSBIG

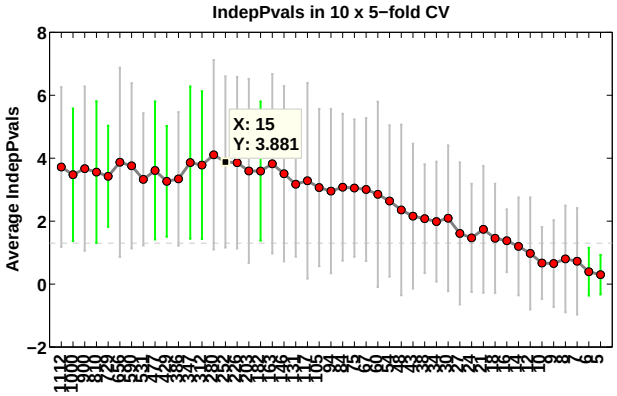
Figure 4: Hazard ratio summaries over considered feature lengths and cross-validation repeats. The means (circles) and 2SD intervals were evaluated on the natural logarithmic scale and then exponentiated back to original scale. Dashed gray line corresponds to HR of 1. Y-axis: Hazard ratios of dichotomised cross-validation test set predictions. X-axis: Probeset lengths evaluated.



(a) Mainz



(b) Rotterdam



(c) TRANSBIG

Figure 5: Minus log10 p-value summaries from likelihood ratio testing of importance of the predictions given the clinical covariates. Dashed gray line corresponds to 0.05 level on the log-scale. Y-axis: Mean minus log10 p-values of the cross-validation test set predictions. X-axis: Probeset lengths evaluated.

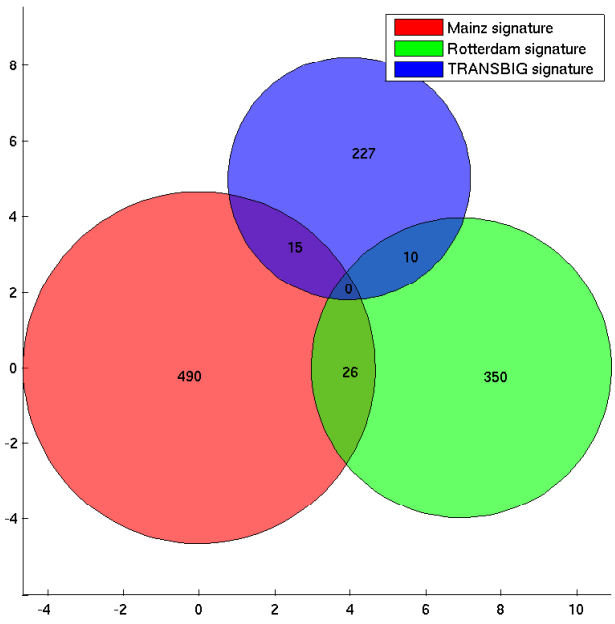
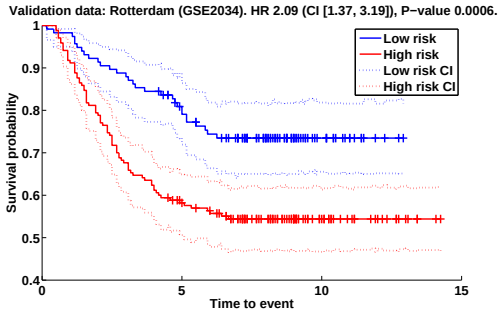


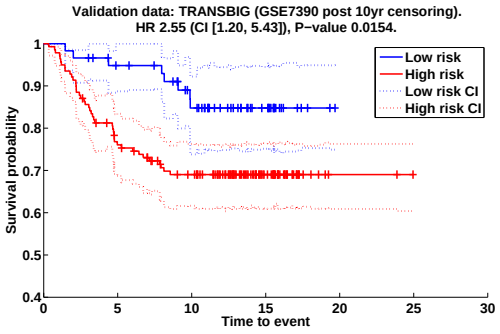
Figure 6: Venn diagram of the probeset content for the three finalised breast cancer signatures.

Table 3: Likelihood ratio test p-values for the breast cancer validation set predictions and clinical covariates (confounders). The p-values are obtained by testing the multivariate Cox model likelihood of the predictions and the clinical covariates in modelling survival time versus reduced models. For each entry in the table the corresponding covariate was dropped and the likelihood ratio p-value of the reduced model was evaluated. Low p-values imply a significant drop in the likelihood.

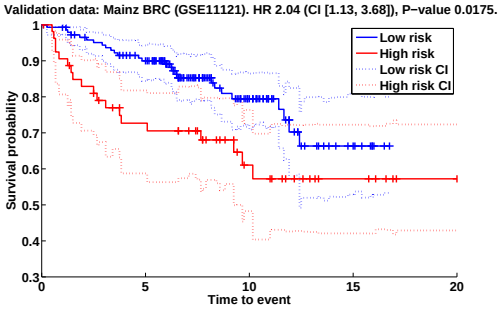
Data	Model	Mainz 531PS	Rotterdam 386PS	TRANSBIG 252PS
Mainz	Predictions:		0.093	0.307
	Grade:		0.052	0.035
	Size:		0.207	0.136
Rotterdam	Predictions:	0.003		0.230
	ER+:	0.981		0.826
	Brain relapse:	0.0002		0.0001
TRANSBIG	Predictions:	0.101	0.002	
	Hosp.Guy:	0.029	0.004	
	Hosp.Igr:	0.290	0.127	
	Hosp.Jrh:	0.145	0.126	
	Hosp.Kar:	0.114	0.018	
	Age:	0.344	0.325	
	Size:	0.031	0.011	
	Surgery:	0.960	0.735	
	Hist.1:	0.763	0.446	
	Hist.2:	0.452	0.179	
	Hist.3:	0.610	0.474	
	Grade:	0.736	0.830	
	ER+:	0.287	0.272	



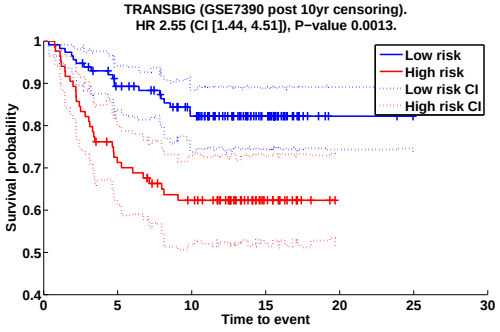
(a) Model: Mainz, Data: Rotterdam



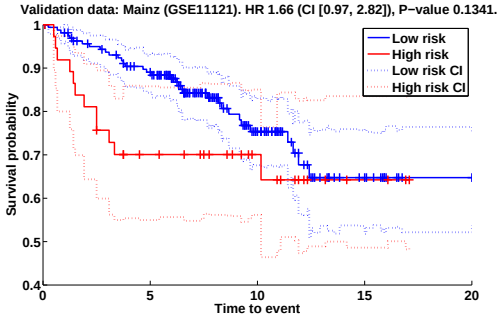
(b) Model: Mainz, Data: TRANSBIG



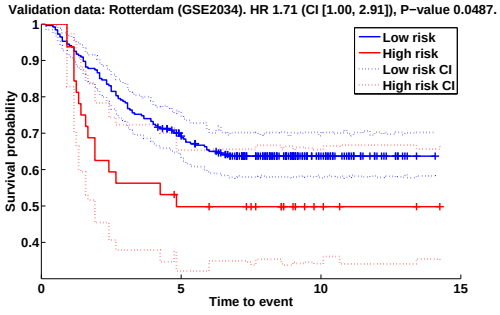
(c) Model: Rotterdam, Data: Mainz



(d) Model: Rotterdam, Data: TRANSBIG

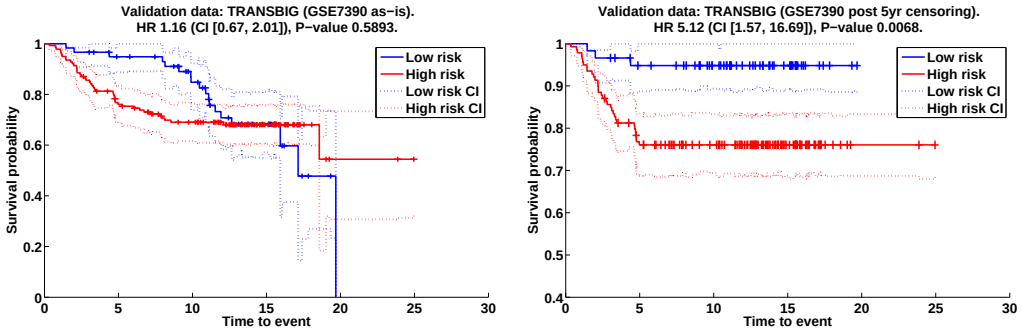


(e) Model: TRANSBIG, Data: Mainz



(f) Model: TRANSBIG, Data: Rotterdam

Figure 7: Kaplan-Meier plots of dichotomised validation set predictions. The solid curves represent the predicted low and high risk groups. Bootstrap based 2.5% and 97.5% quantiles are shown for the survival probability curves.



(a) Data: TRANSBIG as-is

(b) Data: TRANSBIG 5yr censored

Figure 8: Kaplan-Meier plot of dichotomised TRANSBIG data set predictions using the 531 probeset signature generated from the Mainz data. The solid curves represent the predicted low and high risk groups. (a) Data taken as is. (b) All samples with follow-up of more than 5 years censored.