universität wien

DISSERTATION

Titel der Dissertation

„Survival Prediction with Microarray Data“

Verfasser

Mag. Peter Wohlmuth

angestrebter akademischer Grad

Doktor der Sozial- und Wirtschaftswissenschaften
(Dr. rer. soc. oec.)

Wien, im Juli 2013

Studienkennzahl lt. Studienblatt:	A 084 136
Dissertationsgebiet lt. Studienblatt:	Doktorat der Sozial- und Wirtschaftswissenschaften Statistik
Betreuerin:	Ao. Univ.-Prof. Mag. Dr. Martina Mittlböck

Acknowledgement

I would like to thank the supervisor of my thesis Prof. Mittlböck for the support, useful hints for my work and the freedom to follow personal research interests.

Furthermore, I appreciate Prof. Futschik to appraise my thesis.

I want to thank my wife Barbara for understanding what science means to me.

And many thanks to my parents and my siblings, who were very thoughtful in the final steps of my work.

Furthermore, I would like to thank Kathrin for reviewing linguistic expressions.

Table of contents

Chapter 1

1 Introduction

Survival prediction from microarray data has to deal with a high number of correlated variables. This entails the risk of overfitted models with low prediction performance. Appropriate techniques for model selection and model validation result in reliable prognostic models.

This research deals with several topics regarding survival prediction from gene expression data. Survival models, techniques for survival prediction and model tuning are examined with regard to the stability and accuracy of the predictions.

Survival prediction models are analyzed from different perspectives:

- The composition of the population under study: a homogeneous population of patients that is susceptible to relapses and a mixed population of susceptible and insusceptible patients.

- The signal strength of the explanatory variables: high and low-signal data.

- An initial selection of genes: preselected versus non preselected data.

Model techniques and strategies for model validation are evaluated in the first part of this work. Ten popular model approaches, resampling techniques and tuning criteria are examined in the text. The research questions are investigated on real and generated datasets.

The second part of this work is dedicated to survival models, the Cox proportional hazards model (Cox, 1972) and mixture cure models (Boag, 1949, Berkson and Gage, 1952), for mixed frail and cured patients. A new strategy for survival prediction based on cure models is presented and the stability and performance of the models are examined.

This work is arranged as follows:

Chapter 1 outlines the key topics of this thesis.

Chapter 2 exposes preliminaries for survival prediction based on gene expression data. The Cox proportional hazards model and the log partial likelihood method for parameter estimation are presented. A short genetic view on cell growth is given and the acquisition of gene expression data from tissue samples is outlined. Statistical problems regarding survival prediction from microarray data are discussed.

Chapter 3 outlines the course of the survival prediction procedure. It is used to fit and validate the survival models and to assess the survival predictions.

In **chapter 4** model approaches are depicted. The model techniques are reviewed and benefits and weaknesses of the methods are mentioned. Criteria to assess the performance of risk predictions are presented.

Chapter 5 contains the study hypotheses, the results and discussions. The datasets and the computer algorithms are described.

In **chapter 6** cure models and a new technique for survival prediction from mixture cure models is introduced. The performance of the Cox proportional hazards model and cure models are compared for datasets of mixed frail and cured subjects.

Chapter 7 is related to summaries, conclusions and new developments in survival prediction from gene expression data.

1.1 Comments on notation

The text uses standardized notations. The main notation rules are:

Random variables are denoted by uppercase letters. X represents the predictors and T the survival times, (T, Δ) is a tuple of survival time and event status and Y is a continuous response.

Observed values are characterized by lowercase letters y_i, (t_i, δ_i) and x_i, for each patient i with $i=1, \dots, N$. The number of subjects is given by N. The x_i are vectors, the y_i are scalars, a set of measurements is denoted by (y_i, x_i) and (t_i, δ_i, x_i), respectively.

Survival predictions at time t are denoted by $\hat{S}(t)$, continuous outcome predictions by $\hat{y}_i = \hat{f}(x_i)$ obtained from a function f of x_i.

1.2 Abbreviations

ACM	Accelerated failure time mixture cure model
AFT	Accelerated failure time model
AIC	Akaike information criterion
AUC	Area under the ROC curve
BIC	Bayesian information criterion
BioC	Bioconductor: Open software development for computational biology and bioinformatics
BRCA1	Gene associated with breast cancer
BRCA2	Gene associated with breast cancer
BSC	Brier score
C	Concordance index
CCM	Cox proportional hazards mixture cure model
COX	Cox proportional hazards model
CPU	Central processing unit
CRAN	Comprehensive R archive network
CV	Cross-validation
CVPL	Cross-validated log partial likelihood
Cy3	Cyanine dye
Cy5	Cyanine dye
DEV	Deviance
DNA	Deoxyribonucleic acid
DS	Dataset
EM	Expectation maximization
FN	False negatives
FP	False positives
FSS	Forward stepwise selection
GAMLSS	Model class of generalized additive models
GBS	Gradient boosting algorithm
GEO	Gene Expression Omnibus database
IAUC	Integrated area under the ROC curve
IBSC	Integrated Brier score
IDI	Integrated discrimination improvement
IPCW	Inverse probability of censoring weights
IPEC	Integrated prediction error curves
IQR	Inter quartile range
KIT	Gene associated with stromal tumors
KM	Kaplan-Meier estimates
LAS	Lasso approach
LOOCV	Leave-one-out cross-validation
MA	Model approach
mRNA	Messenger RNA
MSE	Mean squared error
NCBI	National Center for Biotechnology Information
NET	Elastic net approach
NRI	Net reclassification index
p53	Gene associated with various tumors

PCA	Principal components analysis
PCR	Principal components regression
PEC	Prediction error curves
PERFORM	Performance of survival models
PH	Proportional hazards
PL	Predictive partial log-likelihood
PLM	Penalized likelihood methods
PLOS	Public Library of Science
PLS	Partial least squares
PSPMCM	Parametric and semiparametric mixture cure models with covariates (SAS macro)
R	Language and environment for statistical computing and graphics, R Foundation for Statistical Computing
R2	Explained variation
RAM	Random access memory
RID	Ridge regression
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
rRNA	Ribosomal RNA
RSF	Random survival forest
RSS	Residual sum of squares
SAS	Statistical Analysis System, SAS/BASE®, SAS/STAT® software, SAS Institute Inc., Cary, NC, USA.
SCAD	Smoothly clipped absolute deviation
SE	Sensitivity
SP	Specificity
SPC	Supervised principal components regression
STATA	Data analysis and statistical software, StataCorp. 2011. Stata® Statistical Software: Release 12. College Station, TX: StataCorp LP.
SVD	Single value decomposition
SVM	Support vector machines
TC	Tuning criterion
tRNA	Transfer RNA
UPV	Univariate selection
URL	Uniform resource locator

Datasets

AML	Acute myeloid leukemia (Bullinger et al., 2004)
CGE	Generated high-dimensional cure dataset
DLBCL	Diffuse large B-cell lymphoma (Rosenwald et al., 2002)
GEN	Generated right-censored and high-dimensional dataset
I3V	Generated cure data (three variables related to survival and cure)
I6V	Generated cure data (three variables related to survival and other three variables related to cure)
NKI	Netherlands Cancer Institute breast cancer data (Van't Veer et al., 2002)
VDX	Erasmus Medical Center breast cancer data (Wang et al., 2005)

1.3 Glossary

Figures

Figure 6-3: Prediction performance of the survival models CCM, ACM and COX measured by IAUC, IBSC and R2 (boxes). The values are obtained from 50 random splits of the VDX data. Models with high performance are characterized by high IAUC and R2 as well as low IBSC values.

Figure 6-4: Product limit estimates and confidence intervals of the VDX data. Vertical lines represent censored observations.

Figure 6-5: Prediction performance of the survival models CCM, ACM and COX measured by IAUC, IBSC and R2 (boxes). The values are obtained from 50 random splits of the CGE data. Models with high performance are characterized by high IAUC and R2 as well as low IBSC values.

Figure 6-6: Product limit estimates and confidence intervals of the generated cure data. Vertical lines represent censored observations.

Figure 6-7: Prediction performance of the survival models CCM, ACM and COX measured by IAUC, IBSC and R2 (boxes). The values are obtained from 50 random splits of the I3V data. Models with high performance are characterized by high IAUC and R2 as well as low IBSC values.

Figure 6-8: Product limit estimates and confidence intervals of the generated cure data (I3V). Vertical lines represent censored observations.

Figure 6-9: Prediction performance of the survival models CCM, ACM and COX measured by IAUC, IBSC and R2 (boxes). The values are obtained from 50 random splits of the I6V data. Models with high performance are characterized by high IAUC and R2 as well as low IBSC values.

Figure 6-10: Product limit estimates and confidence intervals of the generated cure data (I6V). Vertical lines represent censored observations.

Tables

Table 5-1: R and Bioconductor (Bio C) packages used to prepare and analyze the data. Single packages included in base R are not listed.

Table 5-2: Survival prediction procedure to fit and assess survival prediction models. The main steps are: data splitting, model building (based on the resampling techniques 5-, 10- and 20-fold CV and the validation criteria CVPL versus IBSC) and model evaluation (with IAUC, IBSC, DEV and R2).

Table 5-3: Median performance of nine model approaches (rows) obtained from 50 random splits of the generated data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 10-fold CV and either the CVPL (four columns on the left) or IBSC tuning criterion (four columns on the right). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table 5-4: Median performance of ten model approaches (rows) obtained from 50 random splits of the generated data including 500 features. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 10-fold CV and either the CVPL (four columns on the left) or IBSC tuning criterion (four columns on the right). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table 5-5: Median performance of the resampling methods 5-, 10- and 20-fold cross-validation obtained from 50 random splits of the generated data. Survival predictions are made based on ten model approaches (rows). The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The resampling technique with the highest performance is shown in bold numbers and is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table 5-6: Median performance of the tuning criteria CVPL and IBSC obtained from 50 random splits of the generated data. Survival predictions are made based on ten model approaches (rows). The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The tuning criterion with the highest performance is shown in bold numbers and is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table 5-7: High-signal (SIGNAL: sum of the high signal variables F1-F10) and noise variables (NOISE) selected by ten model approaches (rows) obtained from 50 random splits of the generated data. The survival models are validated by 10-fold cross-validation and the CVPL tuning criterion.

Table 5-8: Median performance of eight model approaches (rows) obtained from 50 random splits of the AML data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 10-fold CV and either the CVPL (four columns on the left) or IBSC tuning criterion (four columns on the right). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table 5-9: Median performance of ten model approaches (rows) obtained from 50 random splits of the AML data including 500 features. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 10-fold CV and either the CVPL (four columns on the left) or IBSC tuning criterion (four columns on the right). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table 5-10: Median performance of the resampling methods 5-, 10- and 20-fold cross-validation obtained from 50 random splits of the AML data. Survival predictions are made based on ten model approaches (rows). The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The resampling technique with the highest performance is shown in bold numbers and is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table 5-11: Median performance of the tuning criteria CVPL and IBSC obtained from 50 random splits of the AML data. Survival predictions are made based on ten model approaches (rows). The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The tuning criterion with the highest performance is shown in bold numbers and is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table 5-12: Genes included in 50 survival models using the AML data. The frequency matrix presents in how many cases the genes (rows) are included in the survival models. The analysis is based on nine model approaches (columns). The survival models are validated by 10-fold CV and the CVPL tuning criterion. The table is sorted in decreasing order of the frequency (TOTAL). Presented is the number of model techniques (NAPP) from which a gene was selected at least once. Frequencies above 4 are shown in bold numbers.

Table 5-13: Median performance of eight model approaches (rows) obtained from 50 random splits of the DLBCL data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 10-fold CV and either the CVPL (four columns on the left) or IBSC tuning criterion (four columns on the right). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table 5-14: Median performance of ten model approaches (rows) obtained from 50 random splits of the DLBCL data including 500 features. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 10-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table 5-15: Median performance of the resampling methods 5-, 10- and 20-fold cross-validation obtained from 50 random splits of the DLBCL data. Survival predictions are made based on ten model approaches (rows). The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The resampling technique with the highest performance is shown in bold numbers and is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table 5-16: Median performance of the tuning criteria CVPL and IBSC obtained from 50 random splits of the DLBCL data. Survival predictions are made based on ten model approaches (rows). The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The tuning criterion with the highest performance is shown in bold numbers and is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table 5-17: Genes included in 50 survival models using the DLBCL data. The frequency matrix presents in how many cases the genes (rows) are included in the survival models. The analysis is based on nine model approaches (columns). The survival models are validated by 10-fold CV and the CVPL tuning criterion. The table is sorted in decreasing order of frequency (TOTAL). Presented is the number of model techniques (NAPP) from which a gene was selected at least once. Frequencies above 4 are shown in bold numbers.

Table 6-1a: Parameter estimates and p-values obtained from the new model approach and the Cox mixture cure model (reference model). 5000 survival models, one for each gene, are fitted using the VDX data (details see chapter 6.4): The same gene is included in the latency and incidence part of each model. The covariables (and the intercept) of the incidence (left column) and the covariables of the latency part (right column) are compared. Analysis of the covariables: 1. (row 3-4) Number of variables in the two model approaches, of which both of them are either significant or not significant ($p < 0.05$ and < 0.10 respectively). 2. (row 5-8) Differences (median plus 25 % and 75 % percentiles) of the parameter estimates and the p-values (all genes and genes with $p < 0.10$) in absolute numbers. Analysis of the intercept: 3. (row 10-13) Analogous to 2.

Table 6-1b: General information about the algorithm to compare the new model approach and the Cox mixture cure model. Information about the dataset and running time of the procedure.

Table 6-2: Performance of the survival models CCM, ACM and COX obtained from 50 random splits of the NKI data. The prediction accuracy is measured by IAUC, IBSC and R2. The models are assessed by the median as well as the first and third quartile of the performance values. High prediction performance is characterized by high IAUC and R2 as well as low IBSC values.

Table 6-3: Performance of the survival models CCM, ACM and COX (single columns) obtained from 50 random splits of the NKI data. The prediction accuracy is measured by median values of $AUC(t)$, $BSC(t)$ and $R2(t)$ (global columns) at time $t=1,2,\ldots,12$. The highest performance is characterized by the highest AUC and R2 as well as the lowest BSC values. They are shown in bold numbers.

Table 6-4: Number of principal components (median plus 25 % and 75 % percentiles) in the Cox (COX) and the incidence and latency parts of mixture cure models (CCM, ACM). The running time (median as well as 25 % and 75 % percentiles) of the algorithms for one prognosis based on cure and Cox models is shown in the last row of the table.

Table 6-5: Participation of single genes in survival prediction obtained from 50 random splits of the NKI data. The frequency matrix presents in how many cases the genes (rows) are involved in the survival prognosis. The analysis is based on COX models and mixture cure models (ACM and CCM with latency and incidence parts). The table is sorted in decreasing order of frequency (TOT). TOT.LAT and TOT.INC show in how many cases each gene is related to survival and to cure in total.

Table 6-6: Performance of the survival models CCM, ACM and COX obtained from 50 random splits of the VDX data. The prediction accuracy is measured by IAUC, IBSC and R2. The models are assessed by the median plus the first and third quartile of the performance values. High prediction performance is characterized by high IAUC and R2 as well as low IBSC values.

Table 6-7: Performance of the survival models CCM, ACM and COX (single columns) obtained from 50 random splits of the VDX data. The prediction accuracy is measured by median values of $AUC(t)$, $BSC(t)$ and $R2(t)$ (global columns) at time $t=1,2,\ldots,14$. The highest performance is characterized by the highest AUC and R2 as well as the lowest BSC values. They are shown in bold numbers.

Table 6-8: Number of principal components (median as well as 25 % and 75 % percentiles) in the Cox (COX) and the incidence and latency parts of mixture cure models (CCM, ACM). The running time (median plus 25 % and 75 % percentiles) of the algorithms for one prognosis based on cure and Cox models is shown in the last row of the table.

Table 6-9: Participation of single genes in survival prediction obtained from 50 random splits of the VDX data. The frequency matrix presents in how many cases the genes (rows) are involved in the survival prognosis. The analysis is based on COX models and mixture cure models (ACM and CCM with latency and incidence parts). The table is sorted in decreasing order of frequency (TOT). TOT.LAT and TOT.INC show in how many cases each gene is related to survival and to cure in total.

Table 6-10: Performance of the survival models CCM, ACM and COX obtained from 50 random splits of the CGE data. The prediction accuracy is measured by IAUC, IBSC and R2. The models are assessed by the median plus the first and third quartile of the performance values. High prediction performance is characterized by high IAUC and R2 as well as low IBSC values.

Table 6-11: Performance of the survival models CCM, ACM and COX (single columns) obtained from 50 random splits of the CGE data. The prediction accuracy is measured by median values of $AUC(t)$, $BSC(t)$ and $R2(t)$ (global columns) at time $t=1,2,\ldots,7$. The highest performance is characterized by the highest AUC and R2 as well as the lowest BSC values. They are shown in bold numbers.

Table 6-12: Number of variables (median plus 25 % and 75 % percentiles) in the Cox (COX) and the incidence and latency parts of mixture cure models (CCM, ACM). The running time (median as well as 25 % and 75 % percentiles) of the algorithms for one prognosis based on cure and Cox models is shown in the last row of the table. CORRECT represents the number of correct variables.

Table 6-13: Participation of single genes in survival prediction obtained from 50 random splits of the CGE data. The frequency matrix presents in how many cases the genes (rows) are involved in the prediction model. The analysis is based on COX models and mixture cure models (ACM and CCM with latency and incidence parts). VARS 1, 2 and 3 (rows) represent the variables associated with survival and cure. OTHERS describe the number of noise variables in the survival models. TOT.LAT and TOT.INC show in how many cases each variable is related to survival and to cure in total. The two columns on the right represent the true parameter values of the variables related to survival and cure.

Table 6-14: Performance of the survival models CCM, ACM and COX obtained from 50 random splits of the I3V data. The prediction accuracy is measured by IAUC, IBSC and R2. The models are assessed by the median plus the first and third quartile of the performance values. High prediction performance is characterized by high IAUC and R2 as well as low IBSC values.

Table 6-15: Performance of the survival models CCM, ACM and COX (single columns) obtained from 50 random splits of the I3V data. The prediction accuracy is measured by median values of $AUC(t)$, $BSC(t)$ and $R2(t)$ (global columns) at time $t=1,2,\ldots,9$. The highest performance is characterized by the highest AUC and R2 as well as the lowest BSC values. They are shown in bold numbers.

Table 6-16: Number of variables (median plus 25 % and 75 % percentiles) in the Cox (COX) and the incidence and latency parts of mixture cure models (CCM, ACM). The running time (median as well as 25 % and 75 % percentiles) of the algorithms for one prognosis based on cure and Cox models is shown in the last row of the table. CORRECT represents the number of correct variables.

Table 6-17: Participation of single genes in survival prediction obtained from 50 random splits of the I3V data. The frequency matrix presents in how many cases the genes (rows) are involved in the prediction model. The analysis is based on COX models and mixture cure models (ACM and CCM with latency and incidence parts). VARS 1, 2 and 3 (rows) represent the variables associated with survival and cure. OTHERS describe the number of noise variables in the survival models. TOT.LAT and TOT.INC show in how many cases each variable is related to survival and to cure in total. The two columns on the right represent the true parameter values of the variables related to survival and cure.

Table 6-18: Performance of the survival models CCM, ACM and COX obtained from 50 random splits of the I6V data. The prediction accuracy is measured by IAUC, IBSC and R2. The models are assessed by the median plus the first and third quartile of the performance values. High prediction performance is characterized by high IAUC and R2 as well as low IBSC values.

Table 6-19: Performance of the survival models CCM, ACM and COX (single columns) obtained from 50 random splits of the I6V data. The prediction accuracy is measured by median values of $AUC(t)$, $BSC(t)$ and $R2(t)$ (global columns) at time $t=1,2,\ldots,9$. The highest performance is characterized by the highest AUC and R2 as well as the lowest BSC values. They are shown in bold numbers.

Table 6-20: Number of variables (median plus 25 % and 75 % percentiles) in the Cox (COX) and the incidence and latency parts of mixture cure models (CCM, ACM). The running time (median as well as 25 % and 75 % percentiles) of the algorithms for one prognosis based on cure and Cox models is shown in the last row of the table. CORR(CM) and CORR(COX) represent the number of correct variables for the mixture cure models and the Cox regression model.

Table 6-21: Participation of single genes in survival prediction obtained from 50 random splits of the I6V data. The frequency matrix presents in how many cases the genes (rows) are involved in the prediction model. The analysis is based on COX models and mixture cure models (ACM and CCM with latency and incidence parts). VARS 1, 2 and 3 (rows) represent the variables associated with survival. VARS 4, 5 and 6 are the variables associated with cure. OTHERS describe the number of noise variables in the survival models. TOT.LAT and TOT.INC show in how many cases each variable is related to survival and to cure in total. The two columns on the right represent the true parameter values of the variables related to survival and cure.

Table A-1: Median performance of nine model approaches (rows) obtained from 50 random splits of the generated data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 5-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table A-2: Median performance of nine model approaches (rows) obtained from 50 random splits of the generated data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 20-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table A-3: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of nine model approaches (rows) obtained from 50 random data splits of the GEN data. The survival models are validated by 10-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

Table A-4: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of nine model approaches (rows) obtained from 50 random data splits of the GEN data. The survival models are validated by 5-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

Table A-5: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of nine model approaches (rows) obtained from 50 random data splits of the GEN data. The survival models are validated by 20-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

Table A-6: Median performance of eight model approaches (rows) obtained from 50 random splits of the AML data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 5-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table A-7: Median performance of eight model approaches (rows) obtained from 50 random splits of the AML data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 20-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table A-8: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of eight model approaches (rows) obtained from 50 random data splits of the AML data. The survival models are validated by 5-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

Table A-9: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of eight model approaches (rows) obtained from 50 random data splits of the AML data. The survival models are validated by 10-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

Table A-10: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of eight model approaches (rows) obtained from 50 random data splits of the AML data. The survival models are validated by 20-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

Table A-11: Median performance of eight model approaches (rows) obtained from 50 random splits of the DLBCL data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 5-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table A-12: Median performance of eight model approaches (rows) obtained from 50 random splits of the DLBCL data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 20-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Table A-13: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of eight model approaches (rows) obtained from 50 random data splits of the DLBCL data. The survival models are validated by 5-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

Table A-14: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of eight model approaches (rows) obtained from 50 random data splits of the DLBCL data. The survival models are validated by 10-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

Table A-15: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of eight model approaches (rows) obtained from 50 random data splits of the DLBCL data. The survival models are validated by 20-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

Chapter 2

2 Preliminaries

Objectives

The following chapter is dedicated to preliminaries for survival prediction from gene expression data. In chapter 2.1 the survival function, the hazard function and the Cox proportional hazards model are presented. In chapter 2.2 the main steps in microarray experiments are outlined. Chapter 2.3 is dedicated to statistical issues when survival is predicted from microarray data.

Survival analysis plays an important role in clinical research. It is used to examine the therapeutic success of a drug, to evaluate the number of patients that will overcome a certain disease or to assess the risk of relapses. Survival data usually contain censored subjects as the event of interest, the progression of a disease or death for instance, is not observed during the period under study. Hence the survival of a patient is specified by a mixed variable of the event status and the follow up or failure time.

Survival models are used to determine the effect of risk factors on survival and to apply survival predictions on new data. The linear and the logistic regression model that are applied to data with continuous and binary outcomes are not appropriate for the analysis of time to event data. The product-limit estimator or Kaplan-Meier estimator (Kaplan and Meier, 1958) is the simplest survival prediction model and Kaplan-Meier curves are regularly used to display survival differences between risk groups. The Cox proportional hazards model (Cox, 1972) is the most popular survival regression model to evaluate the influence of clinical variables on the hazard.

This work is dedicated to survival prediction models from high-dimensional data. The survival models are built on

1) a homogeneous sample of patients that is susceptible (part 1 of this work) and
2) a group of patients that is susceptible and insusceptible to relapses (part 2 of this thesis).

Censoring can have different meanings in the two samples.

In the first situation censoring may be due to a short follow-up interval or early study termination and is considered as partly missing outcome information. The Cox proportional hazards model is appropriate to analyze these data, if the proportionality assumptions are met.

In the second case systematic censoring appears due to patients who are insusceptible to relapses. Survival analysis for a mixed population of frail and cured patients can be performed by cure models, the bounded cumulative hazards model (Yakovlev and Tsodikov, 1996) and the mixture cure model (Boag, 1949, Berkson and Gage, 1952). Perperoglou (2006) and Perperoglou et al. (2007) introduced statistical models applied to long-term survivors. Time-varying effects models and reduced rank hazard regression models were presented and benchmarked.

The next section focuses on survival and hazard functions and the Cox proportional hazards model. The second part of this chapter deals with molecular processes in human cells and the acquisition of genetic data from tumor tissues. A further section is dedicated to model building based on microarray data.

2.1 Survival models and parameter estimates

This section refers to statistical methods for the analysis of right-censored survival data.

The survival time describes the time duration to a well defined event. This can be the progression of a disease, a relapse or death. In clinical investigations the survival time is recorded from the study inclusion, the time of diagnosis or the surgical intervention and the follow-up ends if the trial expires, by withdrawal of the informed consent or if the study endpoint is reached.

Survival analysis deals with censored time to event data and positively skewed survival times that can only take positive values (Collett, 2003). Survival times can be described by the survival function $S(t)$ and the hazard function $h(t)$ for time $t \geq 0$. Accelerated failure time models (Lawless, 1982), Cox proportional hazards models (Cox, 1972) or parametric models like the exponential or Weibull survival models can be used to model censored survival data. The survival function, the hazard function and the Cox regression model are described in the next section. The following remarks are based on Collett (2003).

2.1.1 Survival and hazard function

T is a non-negative random variable that represents the failure time. The values of T follow a probability density function $q(t)$. The distribution function represents the probability of death until time t . It is given by:

$$Q(t)=P(T \leq t)=\int_{0}^{t} q(u)\,du\,.$$

The survival function $S(t)$ at time t specifies the probability that a subject survives until t :

$$S(t)=P(T>t)=1-Q(t)\,,$$

where $S(t)$ is a decreasing function, $S(0)=1$ and $S(t)\rightarrow 0$ if $t \rightarrow \infty$.

Connections between lifetime and survival function

The survival function can be represented by a sum of event density functions $q(t)$:

$$S(t)=P(T>t)=\int_{t}^{\infty} q(u)\,du=1-Q(t)\,,$$

where the survival density function is given by:

$$s(t)=S'(t)=\frac{d\,S(t)}{d\,t}=\frac{d\int_{t}^{\infty} q(u)\,du}{d\,t}=\frac{d[1-Q(t)]}{d\,t}=-q(t)\,.$$

Hazard function

The hazard function $h(t)$ specifies the probability that a subject dies in the time period $[t, t+dt]$ when it is alive at t , related to the length of the interval dt (Collett, 2003):

$$h(t)=\frac{P(t\leq T\leq t+dt|T\geq t)}{dt}.$$

The hazard function only takes positive values $h(t)\geq 0$ and sums to infinity for $t=[0,\infty)$.

The cumulative hazard function $H(t)$ at time t can be obtained by the integral over the hazard function from time 0 to t . It is given by:

$$H(t)=\int_0^t h(u)du.$$

Relationship between the survival and hazard function

The hazard function $h(t)$ can be written as a functional of the probability density function $q(t)$ and the survival function $S(t)$:

$$h(t)dt=P(t\leq T\leq t+dt|T\geq t)=q(t)dt/S(t)=-S'(t)dt/S(t).$$

The connection between the cumulative hazard function $H(t)$ and the survival function $S(t)$ can be represented by:

$H(t)=-\log S(t)$ or $S(t)=\exp(-H(t))$.

Estimation of the survival function

The survival function of time to event data can be estimated by the Kaplan-Meier or product limit estimator (Kaplan and Meier, 1958) for instance.

$t_i<t_j$ with $i<j$ are r unique and sorted event times, n_i patients are at risk at t_i- immediately before the event and d_i patients die at t_i . The estimated survival probability in the time interval $[t_i-, t_i]$ is given by $(n_i-d_i)/n_i$.

The unconditional survival probability is represented by the Kaplan-Meier estimator of the survival function:

$$\hat{S}(t)=\prod_{t<t_i}(n_i-d_i)/n_i,$$

where $\hat{S}(t)=1$ for $t<t_i$ (Collett, 2003).

Confidence intervals of Kaplan-Meier estimates

A straight-forward approach to estimate the $(1-\alpha)$ - confidence interval of the survival function $\hat{S}(t)$ would be:

$\hat{S}(t) \pm z_{\alpha/2} se(\hat{S}(t))$,

where $z_{\alpha/2}$ represents a quartile of the normal distribution and $se(\hat{S}(t))$ the standard error of the survival estimates. As symmetric intervals may exceed the limits of $[0,1]$ another technique has to be performed: $\hat{S}(t)$ are mapped to an interval $(-\infty, \infty)$, the confidence intervals are estimated and the values are remapped to [0,1]. Popular mapping functions are logistic and log-log functions, $\log(S(t)/(1-S(t)))$ and $\log(-\log(S(t)))$ respectively (Collett, 2003).

The latter can be used to obtain the variance of $\log(-\log(S(t)))$ by:

$$var[\log(-\log(\hat{S}(t)))] \approx (\log \hat{S}(t))^{-2} \sum_{j=1}^{N} \frac{d_j}{n_j(n_j - d_j)} ,$$

where $se(\log(-\log(\hat{S}(t)))) = \sqrt{var(\log(-\log(\hat{S}(t))))}$.

Estimating the hazard and the cumulative hazard function

The hazard function $\hat{h}(t)$ (Collett, 2003) from t_i to t_{i+1} can be obtained by the Kaplan-Meier method:

$$\hat{h}(t) = \frac{d_i}{n_i \tau_i} \text{ for } t_i \leqslant t < t_{i+1} .$$

The $\tau_i = t_{i+1} - t_i$ represents the length of the i -th time period. The number of deaths at t_i are denoted by d_i and the number of patients at risk by n_i .

The Kaplan-Meier estimate of the cumulative hazard function $\hat{H}(t)$ can be obtained from the formula $H(t) = -\log(S(t))$:

$$\hat{H}(t) = -\sum_{i=1}^{r} \log \frac{(n_i - d_i)}{n_i} .$$

2.1.2 Logrank test

Survival differences between groups of patients can be examined with the logrank test. The test statistic is based on differences between the observed and expected number of deaths in time intervals that are defined by ordered and unique death times $t_i < t_j$ with $i < j$. The number of deaths in interval i is denoted by d_i , the number of deaths in group 1 and 2 by d_{i1} , d_{i2} and the number of patients at risk by $n_i = n_{i1} + n_{i2}$.

Mantel (1966) introduced a statistical approach to compare the survival in two groups: For a fixed number of deaths d_i and patients at risk n_i in time interval i , a change in the number of deaths in the first group d_{i1} at time interval i influences the number of patients at risk in group 1 n_{i1} , deaths in group 2 d_{i2} and the number of patients at risk in group 2 n_{i2} .

d_{i1} follows a hypergeometric distribution and the expected number of deaths in group 1 at t_i is given by:

$$e_{i1}=n_{i1}d_i/n_i \ .$$

Differences between the observed and expected number of deaths in time interval i are given by $d_{i1}-e_{i1}$. The sum of the differences across all time intervals leads to the statistic U_L :

$$U_L=\sum_{i=1}^{r}(d_{i1}-e_{i1}) \ .$$

For independent death times, the variance of U_L corresponds to the sum of the variances of d_i in the time interval i . The variance of d_{i1} is given by the variance of hypergeometric distributed variables:

$$v_{i1}=\frac{n_{i1}n_{i2}d_i(n_i-d_i)}{n_i^2(n_i-1)} \ .$$

The variance of U_L is obtained by $Var(U_L)=\sum_{i=1}^{r}v_{i1}=V_L$.

The statistics $U_L/\sqrt{V_L}\sim N(0,1)$ and $U_L^2/V_L\sim\chi_1^2$ express deviations from the null hypotheses of equal survival in the two groups.

2.1.3 Cox proportional hazards model

The Cox proportional hazards model (Cox, 1972) is a frequently used regression model to associate explanatory variables with survival.

Definition of the Cox regression model (Collett, 2003):

For each subject i we observe the right censored survival data (t_i,δ_i,x_i) , where $t_i=min(d_i,c_i)$ is the survival time d_i or censoring time c_i and δ_i the event indicator that takes the value 1 if an event occurs and 0 if the subject is censored. The x_i are the p -dimensional covariates for each patient.

The Cox proportional hazards model links the hazard function to the predictors x_i . The hazard at time t can be described by the baseline hazard $h_0(t)$, which remains undefined and the relative risk $\exp(x_i'\beta)$ with parameter vector $\beta=(\beta_1,\beta_2,...,\beta_p)'$.

The Cox regression model is then given by:

$$h(t|x_i)=h_0(t)\exp(x_i{}'\beta) .$$

The hazard ratios of two subjects x_i and x_j are proportional and are not related to time:

$$\frac{\exp(x_i{}'\beta)}{\exp(x_j{}'\beta)}=e^{x_i{}'\beta-x_j{}'\beta} .$$

Parameter estimation in Cox regression models

In Cox proportional hazards models the parameter values are usually estimated by the partial likelihood function (Cox, 1972). The log partial likelihood method can efficiently estimate the model parameters β , whilst the baseline hazard $h_0(t)$ is obtained from β .

The partial likelihood function $L(\beta)$ of the parameter values β expresses the relation between the hazard of a subject *i* and the cumulative hazard for the patients at risk at the event time t_i . $L(\beta)$ is given by:

$$L(\beta)=\prod_{i=1}^{N}\frac{\delta_i\exp(x_i{}'\beta)}{\sum_{j=1}^{N} I(t_i\leq t_j)\exp(x_j{}'\beta)} .$$

The logarithm of the partial likelihood function $l(\beta)$ is specified by:

$$l(\beta)=\sum_{i=1}^{N}\delta_i(x_i{}'\beta-\log(\sum_{j=i}^{N} I(t_i\leq t_j)\exp(x_j{}'\beta))) ,$$

where I is an indicator function that becomes 1 if the condition is true and 0 otherwise. The first and second derivation of $l(\beta)$ with regard to β are used to determine parameter estimates and confidence intervals of the model coefficients. Parameter estimation by the log partial likelihood is valid in the case that no ties appear, else Efron or Breslow approximations can be used (Collett, 2003).

2.2 Microarray gene expression data

This section refers to the acquisition of microarray data from tissue samples of tumors. A succeeding subchapter describes statistical issues when survival prediction is based on gene expression data.

In cancer research the survival prediction of patients was based solely on clinical and pathological data for many years. Since the human genome is decrypted and microarray technology is evolving rapidly, molecular data from human cells can be obtained and the expression levels of thousands of genes can be determined at once. Hence genetic information can be gained and sufficient data can be provided to be used in survival prognosis.

Gene expressions represent the activity level of genes. The molecular information of the cells is stored in deoxyribonucleic acid (DNA). The eukaryotic cell consists of a nucleus that is surrounded by cytoplasm and bounded by a cell membrane. The nucleus of a cell includes 23 pairs of chromosomes. The chromosomes are an organized structure of DNA. Eukaryotic cells, between 20 and 100 trillion in the human body, fulfill important tasks in the organism like cell renewal or energy production.

Cell renewal

Human cells are built by division from mitosis or meiosis. Whilst meiosis is associated with reproduction, cell renewal in the human body is related to mitosis. Mitosis creates new cells and shares the DNA between the new daughter cells. Hence the replicated cell includes the genetic information of the mother cell. Mitosis consists of several phases: the inter-, meta-, ana- and telophase.

Cancer genesis

The so called “proto-oncogenes” control cell division and monitor the DNA for errors. If replication dysfunctions occur, the division is interrupted, the cells are repaired or cell death, the apoptosis, is induced.

Damaged DNA can cause gene mutations and can lead to uninhibited cell growth. Tumor genesis and progression are affected by tumor suppressor and proto-oncogenes. Gastrointestinal stromal tumors for instance can be evoked by the gene KIT. Tumor suppressor genes are responsible for delayed growth, DNA repair and cell death. Well-known genes are p53, and the breast cancer genes BRCA1 and BRCA2 (American Cancer Society Cancer Information Database, 2011).

Mutations are dysfunctions in cell DNA that can activate and inactivate genes. The activation of proto-oncogenes and inactivation of tumor suppressor genes may be associated with the genesis of cancer cells.

Gene expressions

The DNA of a cell consists of nucleotides that contain the four bases Adenin (A), Guanin (G), Thymin (T) and Cytosin (C) as well as Desoxyribose-5'-phosphat. Polymers are linked nucleotides that form the double helix structure of the DNA. The main function of the DNA is to synthesize proteins that control important cell functions. The DNA is replicated, transcribed to messenger RNA (mRNA) and translated to proteins (Figure 2-1).

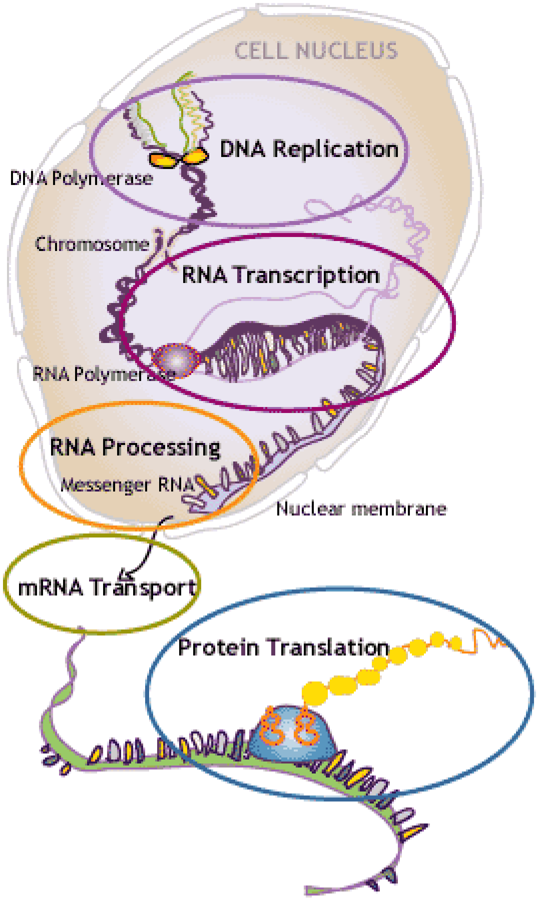

Figure 2-1: Process to express proteins from DNA (*http://nobelprize.org/educational/medicine/dna/pics/intro.gif*, Copyright © Nobel Media AB, retrieved: July 10, 2012. I obtained the owner's explicit consent to use this image in my work.).

The main phases to express proteins from genes are:

1) In the replication phase the two polynucleotide strands in double helix structure are unfold and broken by proteins. RNA is attached to the DNA by an enzyme called DNA polymerase, which copies the DNA.

2) In the transcription phase the RNA polymerase process transcripts DNA to messenger RNA (mRNA), therefore the mRNA is a copy of the gene with only one strand.

3) In the translation phase the mRNA is used as a template to install sequences of amino acids in proteins via base pairs of the mRNA.

Gene expressions represent the production of proteins from genetic information by transcription and translation. The genome that contains the hereditary information is transcribed to the transcriptome (RNA) and translated to the proteome, which is the whole set of proteins that is expressed by genes. The proteome permits conclusions with regard to the activity level of genes. Measurements on the proteome are costly compared to the transcriptome RNA (mRNA, tRNA, rRNA).

Gene expression data of tumor tissues are acquired by gene sequential methods. The gene intensities can be obtained via microarray chips.

2.2.1 Microarrays

Microarray technology is applied in the diagnostics and therapy of diseases (Debouck and Goodfellow, 1999) to analyze expression levels of genes by comparative experiments of two tissue samples. Differences between the gene expression levels of a reference and a collected sample of patients can be a trigger or an effect of a disease. Some clinical trials examine the effects of pharmaceuticals between samples that are gained before and after therapy.

Two main forms of microarrays do exist - the spotted cDNA arrays and the oligonucleid arrays. They are manufactured in different ways: The cDNA - arrays were developed by the Stanford University, the oligonucleid - arrays were designed by Affymetrix.

Microarrays are glass slides or silicon chips. In both systems a fixation matrix is used to measure the expressions of thousands of genes. Jain (2001) outlined the process steps to obtain gene expression data from tissue samples (Figure 2-2, *http://www.accessexcellence.org/RC/VL/GG/microArray.php,* Copyright © 1994-2009 by Access Excellence @ the National Health Museum, retrieved: March 20, 2013):

1) DNA sequences are attached on microarrays.
2) mRNA samples are extracted.
3) mRNA hybridizes on microarrays.
4) Fluorescence intensities are scanned.
5) Gene intensities are processed for data analysis.

Each gene is represented by DNA sequences with hundreds of base pairs. Genes that were chosen from public databases, were fixed on pre-defined positions of the matrix by ultra-violet light. mRNA from tumor samples is isolated and transcribed to cDNA via transcriptase. In the cDNA-technology, the cDNA is marked by Fluorophoren Cy3 (green color) and Cy5 (red color). The marked samples are applied on the array.

The genes hybridize with fluorescent target DNA, which means that complementary DNA-strands bind to fixed DNA via hydrogen bonds. After not bounded target DNA is washed off, the target DNA from the tumor cells is identified by the position on the fixation matrix. The amount of DNA is related to the intensity of fluorescence. The temperature, washing and undesired connections in hybridization are critical issues in microarray experiments.

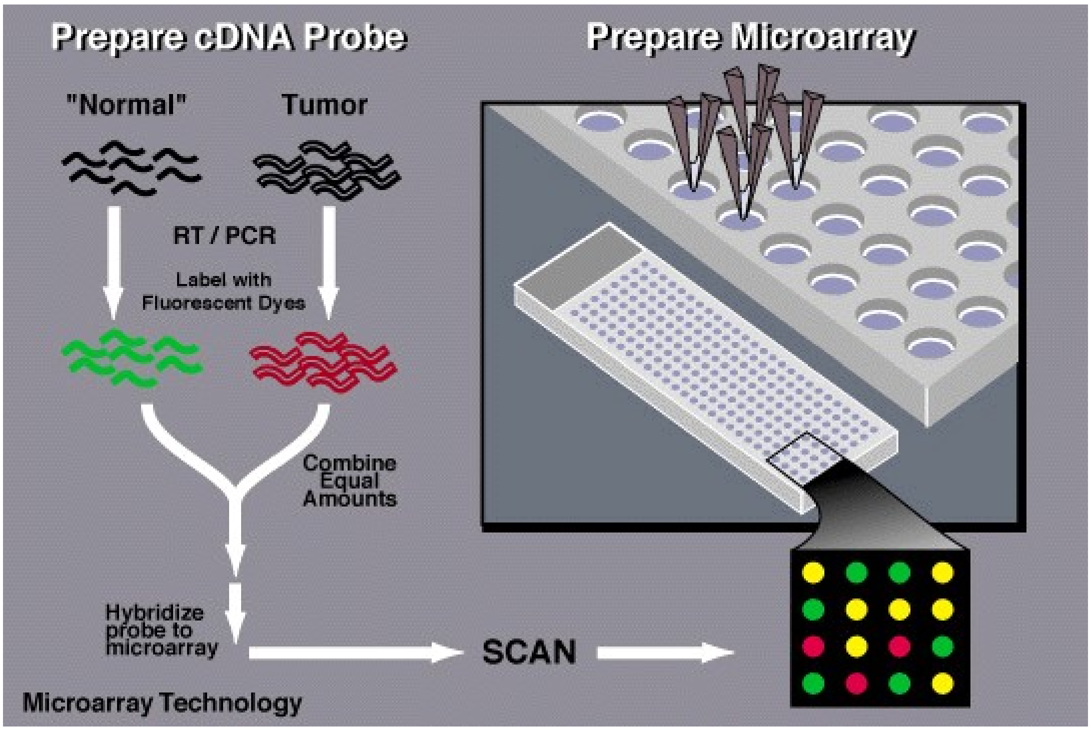

Figure 2-2: Main procedures to acquire gene expression data from tissue samples (*http://www.accessexcellence.org/RC/VL/GG/microArray.php*, Copyright © 1994-2009 by Access Excellence @ the National Health Museum, retrieved: March 20, 2013. I obtained the owner's explicit consent to use this image in my work.).

Scanning

The fluorescence intensities on a microarray chip are scanned by a laser device and saved to picture image files (Figure 2-2). The laser scanner produces different color-intensities on each spot of the array. In the cDNA technology two pictures, one with intensities of Cy3 and one with Cy5, are taken. In the quantification process the color intensities of each spot are translated into real data and the background intensities are removed.

2.2.2 Preprocessing gene expression data

The acquisition of gene expression data from tissue samples contains many "sources of variation" (Gentleman et al., 2005). Various preprocessing steps can balance these issues:

1) The first step is image analysis. The image files of the fluorescence intensities are transferred to "probe level data". Fore- and background pixels of each spot are identified and excluded. The spot intensities are estimated and the pixels are used to create a single number.

2) When microarray data are not stored in one source, they are imported from different databases and are merged for analysis.

3) Microarray experiments are accompanied by optical noise and unspecific binding and the quantities from different arrays can be measured on various scales. Background adjustment and normalization are performed to remedy these issues.

4) When gene expression measurements are represented by multiple samples, the data are summarized to obtain a single value for each gene.

5) Quality assessment is the final preprocessing task to eliminate implausible measurements. Some subjects or gene expressions have to be excluded from data analysis.

2.3 Issues in survival prediction from gene expression data

The Gene Expression Omnibus database (GEO, *http://www.ncbi.nlm.nih.gov/geo/*), the BRB-ArrayTools Data Archive from the National Cancer Institute (*http://linus.nci.nih.gov/~brb/DataArchive_New.html*) and supplementary sides of online journals like PLOS (*http://www.plos.org/*) provide a publicly available access to gene expression data of cancer patients.

A few data sources provide survival status data of patients; others contain clinical data and survival times. Some databases offer raw data or even prepared datasets. Clinical and genetic data are usually stored in separate files and are processed before data analysis. Several issues for data preprocessing and survival prediction have to be considered:

1) Quality assessment of the gene expression data:

- Data gaps and missing data.
- Outliers and implausible data.
- Meta data information.

2) Statistical issues for survival prognosis:

- The number of covariables is much higher than the number of subjects.
- The features exhibit a high level of correlation.
- The survival time may not be observed for every patient.
- There is a high risk of overfitted models.

Gene expression data are partly incomplete, include outliers and contain metadata information. Before the microarray data are linked to clinical data or survival times, the data have to be cleaned: the metadata are removed from the dataset, single subjects and genes that exhibit huge data gaps are excluded from the analysis and outliers or implausible data are deleted. Missing values are added by imputation algorithms like the nearest neighbor implementation by Hastie et al. (1999), which is described in chapter 5.

In microarray studies gene expression data are gained on a few hundred patients and a few thousand genes, also known as the $N \ll p$ paradigm. In survival analysis standard parameter techniques like the partial likelihood technique for Cox proportional hazards models (Cox, 1972) cannot be applied. Alternatives are penalized parameter estimates or partial likelihood estimates on a decreased variable space.

When inference procedures are applied to gene expression data, the cumulative type-1-error has to be taken into account. Variable selection based on univariate Score p-values for instance is affected by the inflation of the alpha-error.

Microarray data exhibit a high level of correlation. Multicollinearity can lead to inaccurate predictions and fragile models. Huang and Harrington (2002) demonstrated that high correlated predictors in multiple survival models can even lead to biased parameter estimates and unstable models in low dimensional datasets of only a few covariables.

Overfitting is an inherent problem in regression models from high-dimensional data. “Overoptimized” models contain a high number of variables that perfectly match the data but exhibit low prediction performance on new data (Bovelstad et al., 2009).

“Noise accumulation” problems are apparent in regression models based on microarray data (Fan and Fan, 2008). Hastie et al. (2009) claimed that noise can hide relevant relationships between in- and output variables and can impair model predictions.

Finally there are no established criteria to determine the optimal complexity level and to assess the accuracy of survival models from high-dimensional data. The Akaike or Bayes Information Criteria (AIC, BIC) are popular metrics for low-dimensional data. Schumacher et al. (2007) pointed out that AIC and BIC are not appropriate in case of high-dimensional data.

Survival prediction based on microarray data has to consider the risk of overfitting, correlated data and noise in the survival model. The following components are needed to cope with these issues (based on Bovelstad et al., 2007, and Boulesteix et al., 2008):

1) **Model approaches** that account for correlated and high-dimensional data.

2) **Model tuning** which is applied by resampling techniques and appropriate criteria to validate the survival models.

3) Accurate **performance metrics** for survival predictions from high-dimensional data.

Model approaches and performance criteria are outlined in chapter 4. Model tuning is described in the next chapter.

Chapter 3

3 Model tuning

Objectives

Chapter 3.1 is dedicated to the bias-variance interrelation that affects model building based on high-dimensional data. In chapter 3.2 resampling techniques and in chapter 3.3 tuning criteria are presented. The survival prediction procedure to validate, fit and assess prognostic models is outlined in chapter 3.4.

When survival models are fitted to gene expression data the model size has to be adequately controlled to avoid overfitting and to ensure accurate predictions on new data. Model tuning generates prediction models that exhibit small prediction losses, a low prediction bias and low variance.

The prediction loss can be expressed by a loss function L that presents the difference between predicted and true response values measured on data that are not involved in model building. Hastie et al. (2009) specified the performance of a linear regression model by the loss function $L(Y, \hat{f}(X))$, where Y represents the response variable, X are the covariates and $\hat{f}(X)$ are the estimated values of the response variable based on X via the prediction model f. The loss function $L(Y, \hat{f}(X))$ measures the difference between the output variable Y and the prediction estimates $\hat{f}(X)$.

The loss of a linear regression model can be evaluated by a least squares fit:

$$RSS = 1/N \sum_{i=1}^{N} (Y_i - \hat{f}(X_i))^2 ,$$

where RSS is the squared sum of the residuals for the subjects $i = 1, \ldots, N$.

The prediction loss of a survival model can be measured by the Brier Score:

$$BSC(t) = 1/N \sum_{i=1}^{N} (\delta_i(t) - \hat{S}_i(t))^2 ,$$

where $\delta_i(t)$ represents the survival status for subject i at time t and $\hat{S}_i(t)$ the survival estimates.

The Brier score (Graf et al., 1999) or the cross-validated log partial likelihood (Verweij and van Houwelingen, 1993) for instance can be used to estimate the prediction loss of survival models. Both metrics are presented in chapter 3.3.

3.1 The bias-variance dilemma

The prediction error of a statistical model is determined by the model bias and the variance of the model estimates. If the complexity of the model is low, a high prediction bias but low variance may occur; if the complexity is high, a low bias but high variance are apparent. The connection between the complexity and the prediction performance of a model is known as the "bias-variance dilemma" (Hastie et al., 2009). A closer look based on the remarks of Hastie et al. (2009) is given in the sequel.

Estimation of the loss function

Prediction models are developed from the training data τ . The test error represents the prediction error of a model that is computed based on data that are not used to fit the model. It is given by:

$$Err_\tau = E[L(Y, \hat{f}(X))|\tau] .$$

The **expected test error** is an average test error measured from random training and test samples. It is represented by:

$$Err = E[L(Y, \hat{f}(X))] = E[Err_\tau] .$$

When the same data are used to fit and to assess the models, the prediction error is obtained by the **training error**:

$$err = 1/N \sum_{i=1}^{N} L(y_i, \hat{f}(x_i)) ,$$

x_i and y_i are the observed covariate and response values.

Bias-variance decomposition

Hastie et al. (2009) demonstrated the bias-variance trade-off regarding a linear model:

$Y = f(X) + \epsilon$ with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$,

where the random variables Y , X and ϵ are the response, the covariates and the error term and f is a linear combination of X .

The number of explanatory variables is tuned. The expected prediction error at a new data point x_0 is given by the squared error loss:

$$Err(x_0) = E[(Y - \hat{f}(x_0))^2 | X = x_0] = \sigma_\epsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 .$$

The expected test error consists of an irreducible error term σ_ϵ^2, the variance of y_0 around $E(f(x_0))$, a squared bias term that represents the deviation of the mean model estimates from the true mean and a term that expresses the variance of the model estimates.

Optimal model size

If the model size increases, the squared bias term of the expected test error falls, but the variance term goes up. As a result the test error decreases until a certain complexity level has been reached and rises thereafter. The turning point represents the optimal complexity level of the model. On the other hand the expected training error steadily decreases for increasing model size.

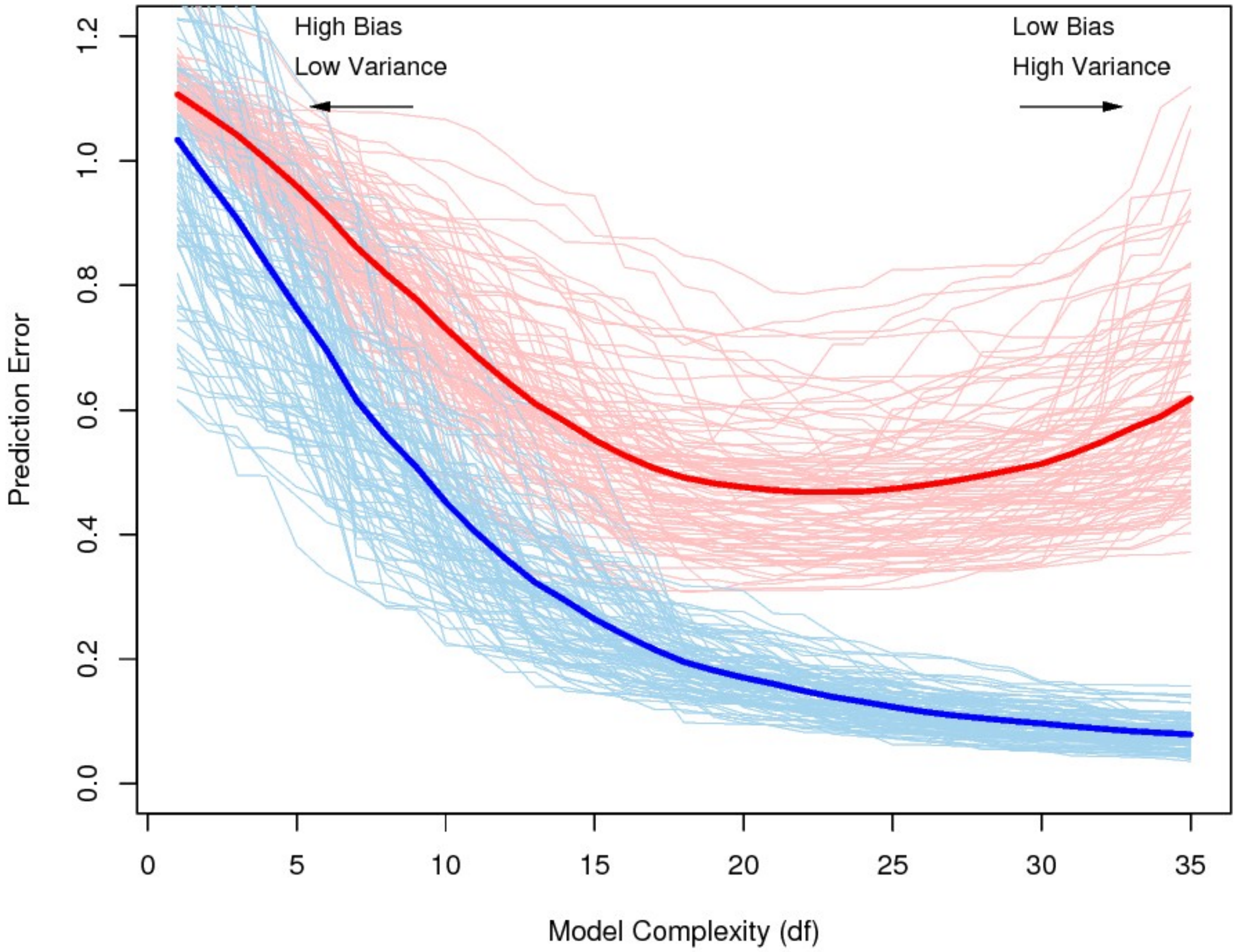

Figure 3-1: (Hastie et al., 2009) Prediction error of training (blue lines) and test data (red lines) for an increasing model size. The light curves represent the test error Err_{τ} and the training error err , the solid lines the expected test and expected training error. Results are obtained from linear models. The lasso is applied to 100 training sets of sample size 50 each. (I obtained the owner's explicit consent to use this image in my work.).

Figure 3-1 illustrates the bias-variance trade-off. The model complexity is shown on the horizontal and the prediction error of the regression models on the vertical axis. Weak blue lines represent training errors calculated from multiple samples and solid colors the average training error. The test errors are shown by weak and solid red lines.

The expected training and test error curves present different courses. The training loss is a strictly decreasing function and the test loss exhibits a global minimum. Hence the training error cannot substitute the test error and the model size has to be selected on the basis of independent data by model tuning. As gene expression data usually exhibit a low number of subjects, data splitting can cause a substantial sample bias. Resampling methods provide repeated model validation and accurate survival predictions.

3.2 Resampling techniques

Resampling techniques are regularly applied in statistics to perform significance tests, to construct confidence intervals or to determine the size of prediction models. Tuned models exhibit minimal prediction error regarding the validation data. Resampling techniques like cross-validation and bootstrapping are used to randomly draw training and validation samples and to estimate the error loss of a statistical model.

The resampling methods can be distinguished according to the following issues:

1) The subjects that are used for model building are drawn with or without replacement. This means a repeated or single inclusion of subjects in the learning sample.

2) Model validation is based on a systematic re-use of randomly split data or on randomly drawn data in every validation step. The latter causes differently sized validation samples.

3.2.1 Cross-validation

Cross-validation (Stone, 1974) is regularly used to select the size of regression models. It works on disjunctive datasets that are partly used to develop and partly exploited to tune the models.

k-fold cross-validation

Cross-validation randomly divides the data into k parts. $k-1$ parts are used to fit and one part to assess the models. The prediction error of the model is obtained by altering the data parts in a systematic manner. Hence every data sample is used k times.

The test error of a prediction model f can be expressed by a function CV . Given the complexity parameter λ the values of CV are calculated by:

$$CV(\hat{f},\lambda)=N^{-1}\sum_{i=1}^{N} L(y_i,\hat{f}_{\kappa(i)}(x_i,\lambda)) .$$

The function $\kappa(i)$ assigns a subject i to one of k data samples. $f_{-\kappa(i)}$ are the model predictions, where the subjects in $\kappa(i)$ are not involved in model development.

The number of data samples in cross-validation

Prediction models from microarray data are usually fitted using 5-fold, 10-fold or leaving-one-out cross-validation (LOOCV). Resampling techniques significantly influence the accuracy of the predictions. The optimal number of validation samples depends on the sample size (Hastie et al., 2009).

Leaving-one-out cross-validation leads to a low biased prediction error, since the training data already cover the whole data sample. As only a single subject is used to validate the predictions, high variance and a high computer effort for parameter estimation have to be considered (Hastie et al., 2009). 5- and 10-fold cross-validation exhibit a lower variance and an increased bias that can take a considerable degree for less than 50 subjects. The bias is negligible for more than 200 subjects (Hastie et al., 2009).

Verweij and van Houwelingen (1993) introduced the cross-validated log partial likelihood method as a leaving-one-out cross-validation technique for Cox regression models. Schumacher et al. (2007) used 5-fold cross-validation for model tuning based on the DLBCL data (Rosenwald et al., 2002) including 240 patients. Bovelstad et al. (2007) applied 10-fold cross-validation to the Netherlands breast-cancer data (van Houwelingen et al., 2006, based on van de Vijver et al., 2002) including N = 295 and the Norway-Stanford data (Sorlie et al., 2003) including N = 115 patients.

Subramanian and Simon (2011) examined resampling techniques to tune survival models from high-dimensional data. The authors investigated the ability of resampling techniques (the sample splitting, leaving-one-out cross-validation, 5- and 10-fold cross-validation method) to provide accurate survival models. Univariate selection, the supervised principal components regression and the lasso method were used to fit the Cox regression models. They found out that 5- and 10-fold cross-validation "provides a good balance between bias and variability".

3.2.2 Bootstrapping

The bootstrap method (Efron, 1979) is a frequently used resampling technique that can be used to estimate the variance or the confidence intervals of estimated sample means. Statistics are derived from random samples of the data that are drawn with returning.

The bootstrap method can be used to tune models from high-dimensional data. b random bootstrap samples are drawn from the data and a prediction model f is developed in each sample. The error loss of the bootstrap sample b is given by the difference between the model prediction $\hat{f}^b(x_i)$ and the response y_i applied to the objects that are not used to fit the model.

The expected prediction error is an average error loss across the bootstrap samples. It is given by:

$$\hat{Err}_B = B^{-1} \sum_{b=1}^{B} 1/|x_{b^-}| \sum_{i:x_i \in x_{b^-}} L(y_i, \hat{f}^b(x_i)) .$$

B represents the number of bootstrap samples, $|x_{b^-}|$ the subjects that are not assigned to the bootstrap sample b and $\hat{f}^b$ the estimates of the prediction model built in the sample b .

Prediction error estimates based on bootstrap samples exhibit a training-size-bias (Hastie et al., 2009). It can be balanced by the .632 estimator (Efron, 1983), a combination of the expected training error $\hat{err}$ and the bootstrap error $\hat{Err}_B$:

$$\hat{Err}^{0.632} = 0.368 * \hat{err} + 0.632 * \hat{Err}_B .$$

The factor 0.632 corresponds to the number of unique subjects in each bootstrap sample.

0.632 improvement

The .632 estimator is rigid for the compensation of the test-size-bias and can be improved by a flexible weighting scheme of the expected training and bootstrap error (Efron, 1983).

The weights are related to the relative overfitting rate $\hat{R}$ that depends on the no-information error rate γ , an error rate of a model when in- and output are independent. The no-information error rate γ can be estimated by:

$$\hat{\gamma}=1/N^2\sum_{i=1}^{N}\sum_{j=1}^{N}L(y_i,\hat{f}(x_j)) .$$

The relative overfitting rate is given by:

$$\hat{R}=\frac{\hat{Err}_B-\hat{err}}{\hat{\gamma}-\hat{err}} .$$

$\hat{R}$ can take values from 0 for no overfitting to 1 where overfitting equals $\hat{\gamma}-\hat{err}$.

The .632+ estimator is obtained by:

$\hat{Err}^{.632+}=(1-\hat{w})*\hat{err}+\hat{w}*\hat{Err}_B$ and $\hat{w}=\frac{0.632}{1-0.368\hat{R}}$.

The weight $\hat{w}$ can take values from 0.632 if $\hat{R}=0$ to 1 if $\hat{R}=1$ and $\hat{Err}^{.632+}$ from $\hat{Err}^{.632}$ to $\hat{Err}_B$.

3.3 Tuning criteria

Model tuning affects the accuracy of prediction models from microarray data. It induces survival models with high performance and avoids overfitting. Model tuning is carried out by a tuning parameter λ which is selected by a tuning criterion $TC(\lambda)$. $TC(\lambda)$ measures the error loss of a regression model of size λ.

The tuning parameter λ is obtained by:

$\hat{\lambda} = argmax_{\lambda_i} TC(\lambda_i)$ for $i = 1,2,\ldots,l$,

where $\hat{\lambda}$ is the optimal value of the tuning criteria $TC(\lambda_i)$ and λ_i is a sequence of parameter values.

The cross-validated log partial likelihood (CVPL; Verweij and van Houwelingen, 1993) and the integrated Brier score (IBSC; Graf et al., 1999) are used to tune survival models in this work. Porzelius et al. (2010) used derived criteria, the prediction error curves and the predicted partial log-likelihood to tune and benchmark survival models.

3.3.1 Cross-validated log partial likelihood

The cross-validated log partial likelihood (CVPL), introduced by Verweij and van Houwelingen (1993), is commonly used to tune survival models based on microarray data. It is a function of the tuning parameter λ:

$$CVPL(\lambda) = N^{-1} \sum_{i=1}^{N} [l(\hat{f}^{-i}(\lambda)) - l^{-i}(\hat{f}^{-i}(\lambda))].$$

The function $\hat{f}(\lambda) = x_i{}'\hat{\beta}$ represents the risk score of a survival model f and l the log-likelihood function of f. $\hat{f}^{-i}$ and l^{-i} express that the model is fitted without the i-th observation ($i = 1,\ldots,N$) and the likelihood is calculated without the i-th observation. The first term $l(\hat{f}^{-i})$ denotes that the model is developed without subject i, but the likelihood is obtained from all subjects.

The predictive partial log-likelihood (PL), Porzelius et al. (2010) for instance, generalizes the CVPL approach of Verweij and van Houwelingen (1993). PL can be applied to cross-validation and bootstrap samples. It is given by:

$$l(\hat{\beta}) = \sum_{b=1}^{B} \sum_{i:x_i \in x_b} \delta_i (x_i{}'\hat{\beta}_{\text{-b}} - \log(\sum_{j:x_j \in x_b} I(T_j \geq t_i) \exp(x_j{}'\hat{\beta}_{\text{-b}}))),$$

where x_b represents the b-th data sample. $\hat{\beta}_{\text{-b}}$ are the parameters that are estimated on all data but the b-th sample. The observations $i = 1,\ldots,N$ are assigned to B equally sized and disjunctive samples and $b = 1,\ldots,B$ with $B \leq N$.

3.3.2 Brier score

The Brier score (Graf et al., 1999) can be used to specify the mean squared prediction error of survival estimates. It is used to tune and to assess survival models in this work.

Using survival data without censoring the Brier score is defined by:

$$BSC(t)=N^{-1}\sum_{i=1}^{N}(I(T_i>t)-\hat{S}(t|x_i))^2 .$$

The observed event status at time-point t is given by $I(T_i>t)$ and T_i is the time duration under study. $\hat{S}(t|x_i)$ are the survival estimates given the variables x_i .

The Brier score is not fully obtainable when censored data appear. Graf et al. (1999) suggested to apply a weighting scheme to handle the information lacks. Contributions to the Brier score are considered in three constellations:

1) A subject gets censored before time point t and the Brier score $BSC(t)$ is obtained at time t . The unknown contributions to the Brier score are compensated by upweighting the following two situations.

2) The event occurs before t , the event status at t is $I(T_i>t)=0$. The Brier Score yields to: $(0-\hat{S}(t|x_i))^2$. These cases are weighted by the factor $1/\hat{G}(T_i)$.

3) The event status is evaluated before T_i . Then $I(T_i>t)=1$ and the contribution to the Brier Score is $(1-\hat{S}(t|x_i))^2$. These situations are multiplied by $1/\hat{G}(t)$.

The function G represents the censoring distribution of the data and can be calculated by Kaplan-Meier estimates.

The Brier score of the survival estimates $\hat{S}$ is defined by

$$BSC^c(t)=N^{-1}\sum_{i=1}^{N}(I(T_i>t)-\hat{S}(t|x_i))^2 W(t,\hat{G},x_i) ,$$

where the censoring weights can be obtained by

$$W(t,\hat{G},x_i)=\frac{I(T_i\leq t)\delta_i}{\hat{G}(T_i|x_i)}+\frac{I(T_i>t)}{\hat{G}(t|x_i)} .$$

The survival probabilities $\hat{S}$ are obtained from survival models. In Cox proportional hazards models the survival estimates are given by:

$\hat{S}(t|x_i)=\exp(-\hat{H}_0(t)\exp(x_i'\hat{\beta}))$ or $\hat{S}(t|x_i)=\hat{S}_0(t)^{\exp(x_i'\hat{\beta})}$, where $\hat{H}_0(t)$ and $\hat{S}_0(t)$ are the baseline cumulative hazard and the baseline cumulative survival function, which represent the survival $S(t)$ if all x_i are 0. $\hat{H}_0(t)$ and $\hat{S}_0(t)$ are linked by: $\hat{H}_0(t)=-\log(\hat{S}_0(t))$.

Collett (2003) published a technique based on Kalbfleisch and Prentice (1973) to estimate $\hat{H}_0(t)$ and $\hat{S}_0(t)$:

$$\hat{h}_0(t_j)=1-\hat{\xi}_j \text{ and } \hat{H}_0(t)=-\log(\hat{S}_0(t))=-\sum_{j=1}^{r}\log(\hat{\xi}_j) ,$$

where $h_0(t)$ presents the baseline hazard at time t , $j=1,\ldots, r$ and $t_1<t_2<\ldots<t_r$ are r death times.

The ξ_j solve the formula:

$$\sum_{l\in D(t_j)}\frac{\exp(\hat{\beta}' x_l)}{1-\xi^{\exp(\hat{\beta}' x_l)}}=\sum_{l\in R(t_j)}\exp(\hat{\beta}' x_l) ,$$

where $D(t_j)$ and $R(t_j)$ are the subjects with events and patients at risk at time t_j (Collett, 2003). $\hat{\beta}$ represents the parameter estimates in the Cox regression model. Solutions for $\hat{\xi}_j$ are not shown here.

The Brier score is a function of the time t. The integrated Brier score (IBSC) is an integrated version of the Brier Score over time t :

$$IBSC=\int_0^{t_{ub}} BC^c(t,\hat{S})\,dt ,$$

where t_{ub} with $t_{ub}>0$ represents the upper bound of the time interval for which the Brier score has been calculated. This can be the last event time for instance.

3.4 Steps to develop and assess survival models from microarray data

A survival prediction procedure is used in this work to tune, fit and evaluate models based on high-dimensional data. The main reasons are:

1) The complexity of the survival models can be adequately controlled and overfitted models can be avoided.

2) Automatic procedures ensure objective comparisons between different model techniques.

3) The results of this work can be transparently examined and reviewed.

Many publications about model approaches for gene expression data exhibit the superiority of one method over a few others. Model comparisons are performed using one or two datasets and model approaches are applied with already known tuning parameters (van Wieringen et al., 2009). The results of different publications can hardly be compared since different model techniques, resampling strategies and algorithms have been used. The models are validated on the basis of external or internal data. Different performance criteria are used.

In this work the survival prediction models based on microarray data (e.g. Bovelstad et al., 2007) are assessed in the following way:

- Random training, validation and test samples are drawn to fit the models and to evaluate the predictions.

- Resampling techniques are used to tune the models (various tuning criteria and a different number of validation samples are used).

- The procedures are applied to three datasets, ten model approaches are implemented and all procedures are executed repeatedly to avoid random results.

Course of the survival prediction procedure

Figure 3-1 summarizes the steps of the prediction procedure that is used in this work (based on Bovelstad et al., 2007):

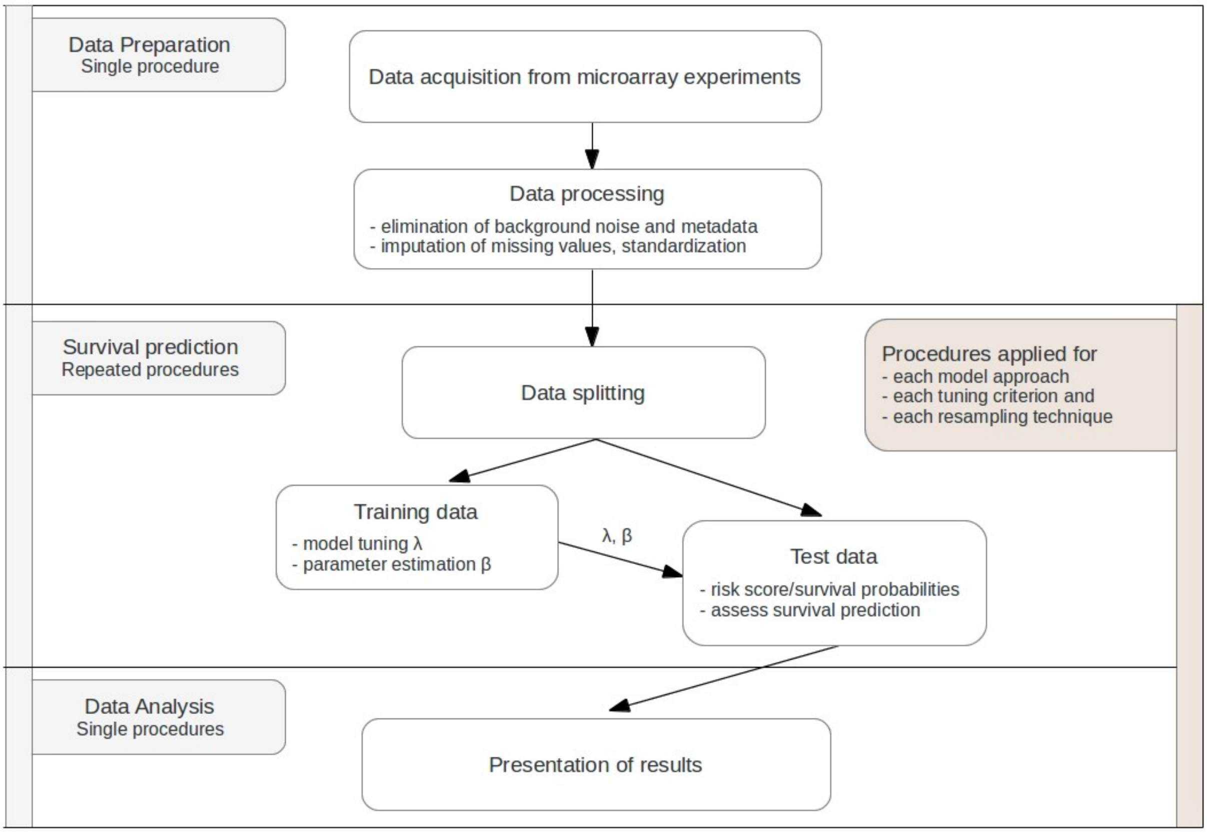

Figure 3-2: Course of the survival prediction procedure, which is used in this work.

The important process steps are:

1) **Random samples of training and test data**
The survival models are internally evaluated. The data are randomly split at a ratio of 2 to 1 into a training sample that is used to tune the survival models and to estimate the model parameters, and a test sample that is required to assess the survival predictions.

2) **Complexity selection**
The model size is selected on the basis of learning samples by k -fold cross validation. The optimal tuning parameter λ of the model approaches, for which details are given in chapter 4, is selected by the tuning criterion $TC(\lambda)$.

3) **Parameter estimation**
The parameter values $\hat{\beta}_j$ of the survival model are estimated based on the whole training data given the tuning parameter $\hat{\lambda}$.

4) **Model assessment**
The performance of the survival predictions is assessed based on the risk score $RS_i = x_i'\hat{\beta}$, where x_i are the covariables of patient i from the test data and $\hat{\beta}$ is a parameter vector that is estimated based on the training data. Appropriate evaluation metrics are described in the next chapter. Information loss is a main disadvantage in the internal setting, because only a part of the data is used to fit the survival models.

Every process step (1-4) is repeated 50 times to obtain reliable results.

Closing remarks

The following remarks propose alternatives and additions to the above process:

1) Usually the same dataset is used to fit and evaluate the survival models. If two datasets are available "external validation" of the survival models can be taken into account.

2) The cross-validated log partial likelihood and metrics based on the Brier score, IBSC (Graf et al., 1999) or prediction error curves (PEC; Porzelius et al., 2010, for instance) are popular tuning criteria. The CVPL is restricted to parametric and semiparametric models (Porzelius et al., 2010), whilst the Brier score and prediction error curves can be used for all survival models.

3) In this thesis 5-fold cross-validation (Schumacher et al., 2007, for instance), 10-fold cross-validation (Bovelstad et al., 2007, for instance) and 20- fold cross-validation are used to tune the models. Alternatives are leave-one-out cross-validation or bootstrapping for instance (Porzelius et al., 2010).

Chapter 4

4 Model approaches

Objectives

Chapter 4 is dedicated to approaches to predict survival from microarray data. The benefits and deficiencies of the model techniques are discussed in the text. Performance metrics are introduced in the last section of chapter 4. They are used to assess the accuracy of the survival predictions.

Model approaches allow robust survival predictions from correlated and high-dimensional data. In the last twenty years a high number of techniques was introduced. Almost all approaches were applied to linear models and were later extended to censored survival data. The model approaches can be classified taking different considerations into account:

- **Underlying statistical method**. The most considerable categories are subset selection and shrinkage-based techniques, ensemble- and variance-based methods. Subset selection approaches identify single or a set of genes that are related to survival. Shrinkage techniques apply penalized parameter estimation. Ensemble techniques provide a combination of base learners via single regression models. Variance-based methods project the data to a lower feature space so that the partial log-likelihood method can be used to estimate the parameter values of the Cox model.

- **Feature selection and feature extraction methods** (van Wieringen et al., 2009). Feature extraction techniques predict survival from aggregated data, while feature selection approaches work on individual genes. Models based on feature selection include untransformed data and are very easy to interpret, but extraction techniques usually have higher performance (van Wieringen et al., 2009).

- **Supervised and unsupervised approaches**. Supervised techniques lower the feature space with respect to the outcome variable, whilst unsupervised methods operate independently of the survival response. Unsupervised approaches are time saving but show poorer performance than supervised methods (Bovelstad et al., 2007, Witten and Tibshirani, 2010).

- **Uni- and multivariate approaches** (van Wieringen et al., 2009). Univariate methods do not consider relationships between the features, whilst multivariate techniques take the influences of other variables into account. Therefore multivariate procedures are time consuming.

In chapter 4.1 model techniques for high-dimensional data are introduced. In chapter 4.2 performance criteria are described.

4.1 Model fitting techniques

The most considerable techniques to predict survival from gene expression data are presented in this section. Subset selection, derived-direction, penalized and ensemble techniques are described in the sequel.

4.1.1 Subset selection strategies

Subset selection approaches identify single or a set of genes that are associated with survival and include the features in a multiple regression model. In the high dimensional setting path-searching algorithms can be used to select survival models, since it is costly and complex to detect optimal gene subsets. One deficiency of these procedures is that only local optima can be reached (Witten and Tibshirani, 2010).

The tuning parameter λ in subset selection approaches controls the number of genes that is included in the prediction model. Univariate selection and forward stepwise selection can be used to predict survival from gene expression data, whilst the backward and the stepwise selection method cannot be applied.

4.1.1.1 Univariate selection

Univariate selection (Bovelstad et al., 2007, for instance) identifies genes with a strong one-dimensional effect on survival and incorporates the features in a multiple survival model.

The algorithm works in the following way:

1) Univariate Cox regression model $h_j(t)=h_0(t)*\exp(\beta_j x_j)$ are applied for each gene x_j with $j=1,\ldots,p$ and the Score test is used to prove the hypothesis: $H_0:\beta_j=0$ versus the alternative $H_A:\beta_j\neq 0$.

2) The features are sorted by the p-values of the Score test in ascending order.

3) The λ genes with the lowest p-values are included in a multiple Cox model, where the tuning parameter λ represents the model size.

If the partial log-likelihood method is used to estimate the parameters in Cox regression models, λ has to be lower than the sample size N .

Univariate selection is a supervised feature selection approach that does not transform the feature space. It leads to sparse models, but does not take correlations between genes into account. The final model is easy to interpret and can be applied with low computational costs.

4.1.1.2 Forward stepwise selection

Forward stepwise selection (Bovelstad et al., 2007, for instance) is a feature selection technique that develops the prediction models sequentially. In every model step the most significant gene given the model with previously included genes is incorporated in the survival model.

The forward stepwise selection works as follows:

1) The first step is analogous to univariate selection. Score p-values of univariate Cox models $h_j(t)=h_0(t)*\exp(\beta_j x_j)$ are determined for each gene x_j with $j=1,\ldots,p$ .The gene with the lowest p-value of the Score test is selected first.

2) In every further step: One gene is included in the model that obtains the lowest p-value of the Score test based on the existing model. The model size is controlled by the tuning parameter λ .

The survival model is then given by:

$$h_i(t)=h_0(t)*\exp(\sum_{m=1}^{\lambda}\beta_m x_m) ,$$

where x_m represents the feature that enters the model in the m -th step. The complexity parameter λ specifies the number of features that are included in the survival model.

Forward stepwise selection is a supervised, feature selection and multivariate approach. It leads to sparse models and the results are easy to interpret, but it is a path-searching technique that only obtains locally optimized models (Witten and Tibshirani, 2010).

4.1.2 Methods using derived input directions

Model approaches based on derived input directions (Hastie et al., 2009) summarize the data by variance- or similarity-based techniques like factor analysis, principal components analysis or cluster methods. The data are aggregated using linear combinations, cluster medoids or variance components of the original data. Popular model approaches are the principal and supervised principal components regression (Bair and Tibshirani, 2004) and the partial least squares method (Park et al, 2002).

The model techniques transform the original variables X_j , $j=1,...,p$, to linear components Z_m , $m=1,...,M$ and $M<p$ that are used as explanatory variables in the survival models. The variance- and covariance based methods create uncorrelated components, but the transformed data can hardly be interpreted.

4.1.2.1 Principal components regression

Principal component analysis (PCA; Pearson, 1901) is a statistical method that is used to transform correlated data to uncorrelated, orthogonal variables. The principal components are linear combinations of the data and obtain the highest possible variance under the constraint that the components are orthogonal.

Principal components regression (PCR) is a model selection approach for high-dimensional and correlated data. The variables are mapped from a p -dimensional input space to a lower q -dimensional subspace ($q\leq N$, $q\ll p$). The first principal components are used for survival prediction.

The principal components are obtained by single value decomposition (SVD) of mean corrected data X . The single value decomposition (SVD) of the $N\times p$ data matrix X is given by:

$$X=UDV' ,$$

where the $N\times p$ matrix U and the $p\times p$ matrix V are orthogonal matrices that represent the column and the row space of X (Hastie et al., 2009). The diagonal elements of D are the non-negative eigenvalues of X .

The covariance of X is given by $S=X'X/N$, where $X'X$ can decompose to:

$$X'X=VD^2V' .$$

The columns of the diagonal matrix V , the v_j , are the eigenvectors or principal components directions of X . The first principal component z_1 is defined as:

$$z_1=Xv_1 ,$$

where z_1 has maximal variance $Var(z_1)=d_1^2/N$. Further principal components are uncorrelated to each other and exhibit the highest variance of the data.

Principal components regression models

The first λ principal components z_j , $j=1,...,\lambda$, are included in the survival prediction model, whereas the number of components λ is tuned. The Cox proportional hazards models is then given by:

$h(t)=h_0 \exp(z_1\beta_1+...+z_\lambda\beta_\lambda)$,

where $h(t)$ represents the hazard at time t , h_0 the baseline hazard and $\beta_1,...,\beta_\lambda$ the coefficients of the survival model.

Principal components analysis is scale sensitive and the data have to be standardized beforehand. Principal components regression is a feature extraction, unsupervised and multivariate approach. The benefits of principal components regression are that the data reduction is variance conserving, the components are uncorrelated and the approaches cause low computational costs. The deficiencies are that the impact of individual genes on survival can only be interpreted with difficulty and the components are not related to survival. Therefore this model techniques may have poor model performance.

4.1.2.2 Supervised principal components regression

The supervised principal components regression (SPC) was introduced by Bair and Tibshirani (2004). The SPC approach preselects genes that are related to survival. Furthermore, it performs principal components analysis using the significant genes and carries out survival prediction from the principal components.

Univariate Score tests are applied for every feature x_j with $j=1,...,p$. The univariate Cox regression models are given by:

$h(t)=h_0 \exp(x_j\gamma_j)$ for $j=1,..,p$ and the model parameters γ_j .

The genes are sorted by significance and principal components analysis is performed on the λ_1 most significant genes. The first λ_2 components z_l are incorporated in a Cox regression model that is given by:

$h(t)=h_0 \exp(z_1\beta_1+...+z_{\lambda_2}\beta_{\lambda_2})$,

where $z_l=f(x_k)$ and $k=1,..,\lambda_1$. The x_k are sorted by the p-values of the Score test in ascending order. Model tuning is applied to a grid of values for λ_1 and λ_2 .

Supervised principal components regression is a hybrid method with subset selection and variance-based elements. On the one hand it is a univariate, supervised, feature selection approach and the other hand a feature extraction, unsupervised and multivariate technique. The SPC approach has average performance (Witten and Tibshirani, 2010, Bovelstad et al., 2007, van Wieringen et al., 2009). One deficiency of the supervised principal components regression is that two tuning parameters have to be estimated, which produces high computational costs.

4.1.2.3 Partial least squares method

Wold (1966) introduced the partial least squares (PLS) technique to fit linear models; Nguyen and Rocke (2002) and Park et al. (2002) to develop survival models. The partial least squares approach creates supervised linear combinations of the original variables (Nguyen and Rocke, 2002) that exhibit maximal covariance with the survival response.

Li and Gui (2004) outlined an algorithm to apply the partial least squares method to survival data. The Z_i represents the i -th partial least squares component that is obtained via the residual matrix V_{ij} from Cox regression models and $j=1,\ldots,p$ numbers the covariates consecutively.

The algorithm works as follows:

1) First a matrix $V_{1\mathrm{j}}$ of mean corrected features of x_j is calculated.

2) For each feature x_j the parameter γ_j is estimated from a one-dimensional Cox models that is given by: $h(t)=h_0(t)\exp(\gamma_{1\mathrm{j}} V_{1\mathrm{j}})$.

3) The first component is obtained by: $Z_1=\sum_{j=1}^{p} w_{1\mathrm{j}}\hat{\gamma}_{1\mathrm{j}} V_{1\mathrm{j}}$, where the sum of weights $w_{1\mathrm{j}}$ in the first sequence is set to 1.

4) The first step summarized information of the X by maximizing the covariance with the outcome variable. For all further steps $i+1$, the component Z_{i+1} is obtained by the residuals $V_{(i+1)j}$, when V_{ij} is regressed on Z_i : $V_{(i+1)j}=V_{ij}\frac{V_{ij}'Z_i}{Z_i'Z_i}Z_i$.

5) The parameter estimates $\hat{\gamma}_{(i+1)j}$ are obtained by a Cox regression model for each feature j : $h(t)=h_0(t)\exp(\gamma_1 Z_1+\ldots+\gamma_i Z_i+\gamma_{(i+1)j} V_{(i+1)j})$.

6) Then the Z_{i+1} is given by: $Z_{i+1}=\sum_{j=1}^{p} w_{(i+1)j}\hat{\gamma}_{(i+1)j} V_{(i+1)j}$. Li and Gui (2004) suggested to determine the weights w_{ij} by $w_{ij}=Var(V_{ij})$.

The last three steps are repeated to generate further components. Survival prediction is applied by the λ first PLS components $Z_1,\ldots,Z_\lambda$.

Further partial least squares algorithms were introduced by Bastien (2004) and Bastien et al. (2005). Partial least squares is a supervised, multivariate and features extraction method. The role of single features on survival cannot be obtained easily. Li and Gui (2004) introduced an approach to recalculate the parameter estimates of X_j from the components Z_j .

4.1.3 Penalized likelihood methods

Standard regression models, the linear model or the Cox proportional hazards model for instance, induce biased estimates using correlated data. Penalized likelihood methods (PLM) perform constraint parameter estimation on the data and exhibit a low prediction error (Huang and Harrington, 2002). Hastie et al. (2009) remarked that shrinkage methods can obtain "interpretable" and accurate models.

The penalized likelihood methods for the Cox proportional hazards model are obtained by the log partial likelihood method (Cox, 1972) and constraints regarding the parameter values β. The log partial likelihood function is given by:

$$l(\beta)=\sum_{i=1}^{N}\delta_i(x_i{}'\beta-\log(\sum_{j=1}^{N}I(t_j\leq t_i)\exp(x_j{}'\beta))),$$

where β represents the parameter values, x_i the covariates, t_i the survival time and δ_i the event status of the subject i, where δ_i will be 1 if an event occurred and 0 otherwise.

The penalized log partial likelihood method with a restriction term p_λ for constrained parameter estimation is given by:

$$l_{pen}(\beta)=l(\beta)-\sum_{j=1}^{p}p_\lambda(\beta_j).$$

The complexity parameter λ represents the amount of shrinkage, where high penalties lead to low parameter values. Popular implementations of the penalty technique are the lasso (L1 shrinkage; Tibshirani, 1997) and the ridge regression (L2 penalization; Hoerl and Kennard, 1970) and hybrid forms like the elastic net.

4.1.3.1 Ridge survival regression

Hoerl and Kennard (1970) introduced the ridge regression for linear models. The authors performed a constrained estimation of the model parameters β_j, $j=1,...,p$, by $\sum\beta_j^2\leq\lambda$, to prevent biased estimates for correlated data.

Van Houwelingen et al. (2006) presented the penalized likelihood method with quadratic shrinkage terms for Cox proportional hazards models. The parameter values β are estimated by the log partial likelihood function $l(\beta)$ and the penalty function $p_\lambda(\beta_j)=\lambda\sum\beta_j^2$, where the tuning parameter λ controls the amount of shrinkage. The parameter estimates are given by:

$\hat{\beta}=argmin_\beta l(\beta)$ subject to $p_\lambda(\beta_j)$.

The ridge regression is a multivariate, supervised feature extraction approach, where quadratic penalties shrink the parameter values towards but not exactly to zero. On high-dimensional data the ridge regression provides models with a high number of variables. The parameter values have a low level. The ridge regression leads to high performance models (Bovelstad et al., 2007). The models are stable and the parameter estimates are continuous for increasing values of the shrinkage parameter. The L2-penalized regression method was applied by van Houwelingen et al. (2006), Hastie and Tibshirani (2004) and Pawitan et al. (2004) for instance.

4.1.3.2 Lasso approach

The lasso, the least absolute shrinkage and selection operator approach, is a parameter shrinkage technique that was introduced by Tibshirani (1996) for linear models. The lasso technique performs constrained parameter estimation with respect to a shrinkage term $p_\lambda(\beta_j)=\sum|(\beta_j)|\le\lambda$. As the parameter estimates of some variables are shrunken to zero, the lasso can be used for variable selection.

In Cox proportional hazards models the penalized log partial likelihood is used to estimate the parameter values:

$\hat{\beta}=argmin_\beta l(\beta)$ subject to $\sum|(\beta_j)|\le\lambda$,

where $l(\beta)$ represents the log likelihood function, $\sum|(\beta_j)|\le\lambda$ is the shrinkage term, β are the parameter values and the shrinkage parameter λ monitors the degree of penalization.

The lasso approach is a feature selection, multivariate and supervised approach. It provides continuous estimates and sparse models (Tibshirani, 1997) that allow accurate predictions for new data.

Tibshirani (1997) introduced the lasso approach for Cox regression models. The author described the log partial likelihood function $l(\beta)$ by the one-term Taylor expansion $(z-\eta)'A(z-\eta)$, where X is the matrix of the covariates, $\eta=X\beta$ and $A=\delta^2 l/\delta\eta\eta'$ is the second derivation of l by η and $z=\eta+A^{-1}u$ with $u=\delta l/\delta\eta$. The tuning parameter λ is fixed and the parameter values are set to zero, $\hat{\beta}=0$. Iteratively η , u , A and z are estimated and $(z-\eta)'A(z-\eta)$ with $\sum|(\beta_j)|\le\lambda$ are minimized to estimate and recalculate the β until they converge.

Further developments of the lasso approach are presented by Goeman (2010), who used a gradient ascent implementation and a Newton-Raphson algorithm to estimate the parameter values and by Park and Hastie (2007) who introduced path algorithms.

4.1.3.3 Further shrinkage methods

Combinations of the lasso and the ridge regression approach or weighted lasso methods are further developments of shrinkage-based approaches. The adaptive lasso approach (Zhang and Lu, 2007) and the elastic net (Zou and Hastie, 2005, for the linear model, Simon et al., 2011, for the Cox regression model) are described below. The SCAD (smoothly clipped absolute deviation; Fan and Li, 2001) and relaxed lasso (Meinshausen, 2007) are not presented in this text.

The elastic net

Zou and Hastie (2005) introduced the elastic net for linear regression models as a mixed lasso and ridge regression approach. The lasso technique can be used to identify influential features that are related to survival. The number of selected variables is usually lower than the number of subjects. Correlated variables are seldom selected by the lasso. The ridge regression shows higher performance but can hardly be used to identify single features that are related to survival. Correlated features are incorporated in the regression model and the number of predictors can be higher than the number of subjects (Engler and Li, 2009, based on Zou and Hastie, 2005).

The elastic net approach performs parameter estimation by the log partial likelihood method, $\hat{\beta}=argmin_{\beta} l(\beta)$, and a penalty function that links absolute and quadratic parameter shrinkage terms. The penalty function is given by:

$$p_{\lambda_1,\lambda_2}=\lambda_1 |(\beta_j)|+\lambda_2 \beta_j^2 ,$$

where the shrinkage terms are restricted by the tuning parameters λ_1 and λ_2 .

Engler and Li (2009) presented the elastic net approach for Cox regression models based on the quadratic programming technique of Tibshirani (1997). Goeman (2010) used a gradient-ascent algorithm to implement the elastic net technique.

The elastic net is a mixed feature extraction and feature selection procedure. It is a multivariate and supervised approach. For highly correlated data more variables are selected than by the lasso approach. The elastic net requires high computational costs as two tuning parameters have to be estimated.

Adaptive lasso

The adaptive lasso approach for linear models was introduced by Zou (2006) and for survival models by Zhang and Lu (2007). The lasso technique shrinks the parameter estimates; the adaptive lasso applies individual weights to the parameter values of the lasso approach to compensate for the biased estimates.

The log partial likelihood function l is maximized with regard to the shrinkage term:

$$p_{\lambda}(\beta_j)=\lambda w_j |(\beta_j)| .$$

The parameters β_j and the weights w_j are calculated iteratively and λ is the tuning parameter. Zhang and Lu (2007) presented an implementation of the adaptive lasso approach for Cox proportional hazards models. The authors recommend start values $w_j=1/|(\hat{\beta}_j^{\nu})|$, where the initial parameter estimates for β_j^{ν} can be obtained by the original lasso approach. The parameter ν specifies the number of iterations.

The adaptive lasso is a supervised, multivariate and feature selection approach. It leads to sparse models. The parameter estimates are more stable than SCAD (Zou, 2006). Additional computational costs are required to estimate the tuning parameter, the parameter values and the parameter weights.

4.1.4 Survival ensemble methods

Ensemble learning is a machine-learning technique that generates strong prediction rules from many weak base learners. Single rules can be regression models that are aggregated in a parallel or sequential fashion. If ensembles are represented by a set of regression models, the aggregation is obtained from model averaging, in classification schemes counting methods can be used (Hothorn et al., 2006, for instance).

The most popular ensemble methods are boosting (Schapire, 1990) and bagging particularly with the random forest approach (Breimann, 1996 and 2001). Further techniques are stacking (Wolpert, 1992), Bayesian approaches like Bayesian model averaging (Leamer, 1978, and Kass and Raftery, 1995) and neural networks (Hanson and Salamon, 1990). In this work bagging applied to random forests and boosting are discussed.

Bagging

Breiman (1996) introduced the bootstrap aggregating approach (which is called bagging). Bagging is a parallel ensemble technique that was invented to obtain stable predictions with low variance (Bühlmann, 2004). The bagging procedure works for a set of learning data, M samples of size N', that are drawn with replacement from the training data (of size N with $N \geq N'$). Prognoses are based on each data sample and are summarized to a strong predictor. Bagging can be applied to regression models or to classification trees for instance. In this work the bagging algorithm is investigated by regression trees in high-dimensional data settings.

Boosting

The boosting approach was introduced by Schapire (1990), Freund (1995) and Freund and Schapire (1996), the gradient boosting approach by Friedman (2002). Gradient boosting is a sequential model approach with "iterative fitting of appropriately defined residuals" (Hothorn et al., 2006).

This section is dedicated to the implementation of the random forest and the gradient boosting algorithm. The boosting method is discussed in the context of survival ensembles techniques based on Hothorn et al. (2006) and van der Laan and Robins (2003).

Ensemble methods are model approaches for high-dimensional data. They can be applied to models with continuous, censored or binary outcome. Ensemble techniques are used to develop a regression model f from a set of regression models F, that the loss L between the model predictions $f(X)$ and the outcome Y is minimized. In the linear model the random variables Y and X represent the outcome and the covariates. $F_{Y,X}$ is the distribution function of X and Y. The expected loss of the model f is given by:

$$E_{Y,X} L(Y, f(X)) = min_{f \in F} \int L(Y, f(x)) d F_{Y,X} .$$

For censored data with survival time $T = min(D, C)$, $Y = \log(T)$ and the survival status $\delta = I(D \leq C)$, where C is the censoring and D the event time, the loss function can be obtained by a weighting scheme for the subjects. The inverse probability of censoring weights (van der Laan and Robbins, 2003) is given by:

$w_i = \delta_i / G(T_i | X_i)$, where G can be determined by product limit estimates of the censoring function given the covariates X . The expected prediction loss is given by:

$$E\,L(Y, f(X)|G) = N^{-1} \sum_{i=1}^{N} L(Y_i, f(X_i)) w_i \,.$$

4.1.4.1 Gradient boosting

Hothorn et al. (2006) presented a boosting approach for survival models using the gradient descent method. The boosting algorithm sequentially updates the survival predictions that are made from a prediction rule f :

$$\hat{f}^{(b)}(.) = \hat{f}^{(b-1)}(.) + v\, f(.|\nu) \text{ (4.01).}$$

The variable b describes the current sequence, v is the step size $0 < v \leq 1$ and $f(.|\nu)$ is the base learner with the parameter vector ν of a regression model. The ν are estimated by:

$$\hat{\nu} = argmin_{\nu} \sum_{i=1}^{N} w_i (U_i - f(X_i|\nu))^2 \text{ (4.02),}$$

where the w_i are the inverse probability of censoring weights that compensate for censoring and the pseudo-response variable U_i represents the residuals. U_i is obtained by:

$$U_i = \frac{-d\,L(Y_i, f^{(b)}(X_i))}{d\,f^{(b)}(X_i)} \text{ (4.03).}$$

The gradient boosting algorithm for censored data (Hothorn et al., 2006) is given by the following steps:

- The loop counter $b=0$, $U_i = Y_i$, $\hat{f}_0(.) = \hat{f}(.|\nu)$.
- The residuals U_i are updated by (4.03), the base learner $f(.|\nu)$ is obtained by weighted least squares (4.02).
- The survival predictions are recalculated $f^{(b+1)}(.) = f^{(b)}(.) + v\, f(.|\nu)$ with step size $0 < v \leq 1$.
- The loop counter is increased by 1: $b = b+1$, step 2 and 3 are repeated until $b = B$. B specifies the maximal number of iterations.

The log-survival time Y is estimated by $Y = f^{(B)}(X)$.

Gradient boosting is a supervised, multivariate approach and a feature selection technique. Boosting leads to stable survival predictions and requires a low run time.

4.1.4.2 Random survival forest

Hothorn et al. (2004) presented a random forest algorithm for censored survival data. The algorithm estimates the survival of an observation x_{new} (p-dimensional covariate vector). For each bootstrap sample a survival tree is constructed. The elements of the leafs with the new observation x_{new} are determined. The Kaplan-Meier estimates of the aggregated sample (elements of leafs with x_{new} across all bootstrap samples) represent survival predictions for x_{new} .

The random forest algorithm comprises the following main steps:

1) Splitting criteria (based on log-rank statistics of daughter leafs for instance) and stopping rules for the survival trees are determined.

2) Bootstrap samples $B^{(b)}$ with $b=1,2,\ldots$ are drawn with replacement from the data: $B^{(b)}=\{(t_i^{(b)},\delta_i^{(b)},x_i^{(b)}),i=1,2,\ldots,N\}$, where t_i represents survival times, δ_i the event status and x_i the covariate vector of observation i .

3) Within each bootstrap sample $B^{(b)}$ survival trees are constructed. Subjects that belong to the same leaf as the new observation x_{new} are determined. This set of subjects is given by: $B^{(b)}(x_{new})=\{(t_i^{(b)},\delta_i^{(b)},x_i^{(b)})\in B^{(b)}|x_i^{(b)}\in\tau(x_{new},B^{(b)})\}$, where $\tau(x_{new},B^{(b)})$ represents the leaf of the survival tree the observation x_{new} belongs to.

4) The aggregated data sample $B_A(x_{new})$ unites the datasets $B^{(b)}(x_{new})$ from 3): $B_A(x_{new})=[B^{(1)}(x_{new}),B^{(2)}(x_{new}),\ldots]$.

5) Survival estimates of x_{new} are derived from the aggregated sample $B_A(x_{new})$ using the Kaplan-Meier method: $\hat{S}_A(.|x_{new})=\hat{S}_{B_A(x_{new})}(.)$.

Random survival forests exhibit stable survival predictions, but the prediction performance was seldom compared to other model approaches (van Wieringen et al., 2009). The random forest approach is a multivariate, supervised and feature selection approach. Since the aggregated model predictions are not based on a single prediction model the influential genes on survival cannot be found easily. Another implementation of survival trees regarding ensemble methods was presented by Hothorn et al. (2006), who used the ensemble framework from van der Laan and Robins (2003).

4.1.5 Further model approaches

A large amount of statistical publications are dedicated to survival prediction from high-dimensional data. The most popular approaches are described in 4.1.1 to 4.1.4. Some further methods are listed below:

Kernel techniques

Li and Luan (2003) as well as Evers and Messow (2008) introduced support vector machines (SVM) in Cox regression models, linear kernels and quadratic shrinkage terms. Liu et al. (2005) presented a further approach that is based on SVM techniques.

Gene groups, Bayesian techniques, shrinkage-based techniques

Tibshirani (2009) introduced the univariate shrinkage approach and Hastie et al. (2000) presented the gene shaving method based on cluster methods. The supervised harvesting technique, an approach based on cluster analysis, was described by Hastie et al. (2001). Bayesian methods for model selection were introduced by Kaderali et al. (2006). Survival prediction using transformation models was presented by Xu et al. (2005). Goeman et al. (2004) described the global test approach that can be used to examine the influence of gene groups on survival time.

4.2 Assessment of survival predictions

Objectives

Chapter 4.2 is dedicated to performance criteria for survival predictions from high-dimensional data. Discrimination-based criteria are presented in chapter 4.2.1 and metrics of the prediction loss are shown in chapter 4.2.2.

The performance of model approaches is assessed from risk score predictions Z_i :

1) The parameters $\hat{\beta}_{train}$ are estimated from the training data.
2) Risk scores are given by: $Z_i = x_i{}'\hat{\beta}_{train}$, where the x_i are the covariables of the i-th patient from the test sample.

The risk score Z is not an event probability, since it is not limited to the interval [0,1] . Nevertheless Z indicates if a patient is more or less likely to experience an event.

In the statistical literature several metrics are regularly used to assess survival predictions from microarray data. These are the concordance index (Harrell et al., 1996), the Brier score (Graf et al., 1999), the cross-validated log partial likelihood (Verweij and van Houwelingen, 1993), the deviance between the log-likelihood function of the prediction model and the null model, a model without covariates, and the explained variation R2 of the survival model (Graf et al., 1999, for instance).

Bovelstad et al. (2007) assessed survival predictions with logrank tests between two risk groups that were formed by risk score predictions. Haibe-Kains et al. (2008) evaluated the performance of survival predictions by the integrated area under the ROC (receiver operating characteristic) curve and by the sensitivity and specificity of risk scores. Porzelius et al. (2010) used integrated prediction error curves (IPEC) and the predictive partial log-likelihood (PL) to assess the performance of resampling techniques. Benner et al. (2010) evaluated survival predictions based on generated data using the number of false positive (FP) and false negative (FN) predictions plus the mean squared error of the parameter estimates with regard to the true values. He also used the integrated Brier score and the explained variation R2.

Performance criteria for survival predictions can be classified roughly as discrimination, overall performance or goodness of fit and reclassification metrics.

4.2.1 Discrimination ability metrics

Discrimination based criteria exhibit the capability of predictions to correctly rank patients with respect to their survival times. Popular discrimination metrics are the p-value of the logrank test between two risk groups (Mantel, 1966, and Peto and Peto, 1972), the hazard ratio and metrics using time-dependent ROC curves. The latter are the sensitivity and specificity of a model (Heagerty et al., 2000) as well as the area under the ROC curve (Chambless and Diao, 2006).

4.2.1.1 Logrank test and the hazard ratio

The logrank test or Mantel-Cox test (Mantel, 1966, and Peto and Peto, 1972) is a test statistic that is used to compare the survival distribution of two groups. Statistical details are given in chapter 2. Bovelstad et al. (2007) for instance used the p-value of the logrank test and Haibe-Kains et al. (2008) the hazard ratio to evaluate survival predictions from high-dimensional data. The subjects were assigned to a low or high risk group according to individual risk scores. Survival differences between the groups were examined by both statistics. High values of the hazard ratio and low p-values of the logrank test indicated high prediction performance.

Bovelstad et al. (2007) applied a median split of the risk scores to form two risk groups. Patients that are above the median risk score are assigned to the high-risk group and patients that are below or equal to the median risk score are referred to the low-risk group. Haibe-Kains et al. (2008) assigned two thirds of the largest risk scores to the high and one third of the patients to the low-risk group. Subramanian and Simon (2011) claimed that the cut-off values used for group assignment are arbitrary and hazard ratios are "not a metric of predictive ability".

4.2.1.2 Time-dependent ROC curve

In medical diagnostics sensitivity and specificity are used to identify cut-off values of a continuous marker that is used to forecast a binary event (a positive or negative test result). Sensitivity specifies the proportion of correctly identified events (positive test result) and specificity the proportion of correctly identified “non-events” (negative test result). The ROC curve is a common graphical representation of sensitivity and 1-specificity for all possible cut-off values of a continuous predictor. The performance of a predictor can be investigated by the area under the ROC curve. An AUC between 0.5 to 0.7 means arbitrary or poor and 0.8 to 1.0 means high to excellent discrimination of the outcome.

Heagerty et al. (2000) introduced time-dependent ROC curves. $ROC(t)$ at time t summarizes the discrimination potential of a continuous predictor Z and can handle censored observations. The event status at t is denoted by $\delta(t)$.

Time-dependent sensitivity (SE) and specificity (SP) for the cut-off value c of the marker Z and the time t are given by:

$SE(c,t)=P(Z>c|\delta(t)=1)$ and

$SP(c,t)=P(Z\leq c|\delta(t)=0)$.

Sensitivity and specificity are estimated by:

$$\hat{P}_{KM}(Z>c|\delta(t)=1)=\frac{(1-\hat{S}_{KM}(t|Z>c))(1-\hat{F}_Z(c))}{(1-\hat{S}_{KM}(t))} \text{ and}$$

$$\hat{P}_{KM}(Z\leq c|\delta(t)=0)=\frac{\hat{S}_{KM}(t|Z\leq c)\hat{F}_Z(c)}{\hat{S}_{KM}(t)},$$

where $\hat{F}_Z(c)=1/n\sum 1(Z_i\leq c)$ represents the estimated empirical distribution function of the predictor Z. $S(t)=P(T>t)$ is the survival function and $S(t|Z>c)$ the conditional survival function for the subset $Z>c$. The survival estimates are obtained by the Kaplan-Meier method. The Kaplan-Meier estimator does not guarantee monotonicity. A nearest neighbor smoothing kernel (Heagerty et al., 2000, and Akritas, 1994), a weighted Kaplan-Meier estimator, ensures monotone ROC curves.

4.2.1.3 Area under the ROC curve

Chambless and Diao (2006) introduced a method to estimate the area under the time-dependent ROC curve $AUC(t)$ from pairwise comparisons of the event status $\delta_i(t)$, $\delta_j(t)$ at time t and the predictions Z_i, Z_j for subjects $i, j=1,2,\ldots,N$ and $i\neq j$.

The area under the time-dependent ROC curve is given by:

$$AUC(t)=P(Z_i>Z_j|\delta_i(t)=1,\delta_j(t)=0)=\frac{P(Z_i>Z_j,\delta_i(t)=1,\delta_j(t)=0)}{P(\delta_i(t)=1)P(\delta_j(t)=0)}.$$

For ordered and unique (event) times t_l with $l,k=1,\ldots,N'$, $N'\leq N$ and $k\leq l$, the survival function S and the hazard function h, $AUC(t)$ is estimated by:

$$AUC(t_l)=\sum\gamma_k h(t_k)(1-h(t_k))S(t_{k-1})^2-\sum\tau_k h(t_k)\times\frac{(1-S(t_{k-1}))S(t_{k-1})}{S(t_l)(1-S(t_l))}.$$

The variables γ_k and τ_k are given by:

$\gamma_k=P(Z_i>Z_j|\delta_i(t_k)=1,\delta_i(t_{k-1})=0,\delta_j(t_k)=0)$ and

$\tau_k=P(Z_i>Z_j|\delta_i(t_{k-1})=1,\delta_j(t_{k-1})=0,\delta_j(t_k)=1)$.

γ_k and τ_k can be estimated by counting the sizes of two sets:

$\hat{\gamma}_k=count\{i:1\leq i\leq k-1,Z_d(i)>Z_d(k)\}/(k-1)$ and

$\hat{\tau}_k=count\{j\in R_k:Z_d(k)>Z_j\}/(\Re_k-1)$,

where $Z_d(i)$ represents the score value at t_i, R_k is the risk set and $\Re_k$ the number of patients in the risk set.

Haibe-Kains et al. (2008, supplementary data) compared risk predictions regarding $AUC(t)$ at time t, the integrated area under the ROC curve (IAUC) and the specificity $SP(t)$, at certain time points t for a specificity $SE(t)$ of 90 %. The integrated area under the ROC curve is a global metric of the discrimination potential of a predictor and can be obtained by:

$$IAUC=\sum_{k=1}^{K}\frac{t_k-t_{k-1}}{t_K-t_1}AUC(t_k),$$

where t_k with $k=1,\ldots,K$ are unique and sorted observation times. $AUC(t_k)$ is the area under the ROC curve at time t_k (based on Haibe-Kains, 2009).

4.2.1.4 Concordance index

The concordance index is a rank based performance criterion. It is based on the risk estimates and survival times of pairwise compared subjects. Harrell et al. (1996) specify the C index as the proportion of concordant with respect to comparable pairs. Pairs are said to be comparable if the survival times of both subjects are known or one survival time is known that is shorter than the censoring time. Concordance is assumed if $T_i > T_j$ and $Z_i < Z_j$, and $T_i < T_j$ and $Z_i > Z_j$, where T represents the survival time and Z the risk score for patient $i, j=1,\ldots,N$.

Harrell's C index is given by:

$$C=\frac{p_c}{p_c+p_d},$$

where p_c and p_d are the probability of concordance and discordance.

Further definitions of the concordance index were provided by Gönen and Heller (2005) and Uno et al. (2011) but are not shown here.

4.2.2 Overall performance and goodness of fit metrics

Overall performance criteria are regularly used to assess the accuracy of survival predictions from high-dimensional data. The main approaches are based on:

- The comparison of the likelihoods of two comparative models.
- The prediction loss of the models as the difference between predicted survival probabilities and the true event status.
- The explained variation of the survival predictions.

The most popular metrics are likelihood ratio statistics of the difference in deviance, the Brier score (Graf et al., 1999) and the R2 criterion (Nagelkerke, 1991, for instance).

4.2.2.1 Deviance and likelihood ratio statistics

The deviance is a goodness of fit statistic that is used to compare two models M_0 and M_1 via the log-likelihood functions l_0 and l_1. In survival prediction models the log-likelihood function is used for parameter estimation and statistical inference regarding parameter values of regression models. The survival model M_1 can be compared to the null model M_0 for instance. M_1 includes the predictors X. $\beta=(\beta_1,\beta_2,\ldots)$ is the the model parameter vector. In this work the deviance is computed using the Cox partial log likelihood function of the survival model $l(\hat{\beta})$ and of the null model without covariates $l(0)$:

$$DEV=-2*(l(0)-l(\hat{\beta})).$$

Bovelstad et al. (2007) and van Wieringen et al. (2009) applied likelihood ratio statistics to assess the performance of survival predictions. Bovelstad et al. (2007) compared absolute values of the deviance between survival models; van Wieringen et al. (2009) used p-values of likelihood ratio statistics to assess survival predictions.

4.2.2.2 Brier score

The Brier score (Graf et al., 1999) is used to select and assess survival models in this work. It was introduced in chapter 2.

The Brier score at time t , $BSC^c(t)$, is defined as the difference between the survival probabilities $\hat{S}(t|x_i)$, derived from risk prediction models, and the true survival status $I(T_i>t)$ at time t :

$$BSC^c(t)=N^{-1}\sum_{i=1}^{N}(I(T_i>t)-\hat{S}(t|x_i))^2 W(t,\hat{G},x_i) \text{ , where } W(t,\hat{G},x_i)=\frac{I(T_i\leq t)\delta_i}{\hat{G}(T_i|x_i)}+\frac{I(T_i>t)}{\hat{G}(t|x_i)} \text{ .}$$

The function W is a reweighting scheme. It compensates for the information loss of Brier contributions, which are not available (details in chapter 2). G presents the censoring distribution of the data for $i=1,\ldots,N$ subjects.

The integrated Brier score (Graf et al., 1999) is given by:

$IBSC=\int_0^{t_{ub}} BC^c(t,\hat{S})\,dt$, where $t_{ub}>0$ represents the upper bound of the time interval of interest.

Van Wieringen et al. (2009) assessed survival predictions by the Brier score. Porzelius et al. (2010) used a derived metric, the integrated prediction error curves, to tune and evaluate survival predictions.

4.2.2.3 Predictive accuracy

Schemper and Henderson (2000) introduced a metric of “predictive accuracy and variation in survival models”. The metric is specified as the difference between the estimated survival probability of a survival model $S(t|X)$ and the observed survival status. The explained variation of a survival model is obtained from the relative differences between the predictive accuracy in the null and the prediction model with covariates X :

$$V(\tau)=\frac{D_0(\tau)-D_X(\tau)}{D_0(\tau)} \text{ .}$$

The two functions $D_0(\tau)$ and $D_X(\tau)$ represent “overall measures of marginal and predictive accuracy” (Schemper and Henderson, 2000).

They are estimated by:

$$D_0(\tau)=\frac{\int_0^\tau S(t)(1-S(t))f(t)dt}{\int_0^\tau f(t)dt} \text{ and}$$

$$D_X(\tau)=\frac{\int_0^\tau E_X[S(t|X)(1-S(t|X))]f(t)dt}{\int_0^\tau f(t)dt} .$$

$S(t)$ and $S(t|X)$ represent the survival probabilities obtained from the Kaplan-Meier method and the survival model. $f(t)$ is the density function. IPCW (inverse probability of censoring weights) like metrics (Schemper and Henderson, 2000, and Hielscher et al., 2010, for instance) are used to compensate for the information loss caused by a decreasing number of patients over time due to censoring.

4.2.2.4 Explained variation R2

The explained variation of survival predictions describes the variation (randomness) of the survival response that is considered by the prognostic model. The R2 criterion exhibits the residual variance of the predictions. In a linear model it can be represented by $R2=1-MSE/MST$, where MSE and MST are the variances of the regression model and the outcome. In survival prediction models the explained variation can be derived from the Brier score (Graf et al., 1999), the deviance (Nagelkerke, 1991) and from prediction accuracy metrics (Schemper and Henderson, 2000) for instance.

The Brier score can be translated to a metric of explained residual variation. R2 at fixed time points s is given by $R2(s)=1-BSC(s)/BSC_0(s)$. A global criterion of explained variation can be obtained from $R2=1-IBSC/IBSC_0$, where BSC_0 and $IBSC_0$ represent the Brier and integrated Brier score of a prediction model without covariates.

Nagelkerke (1991) introduced the explained variation R2 based on the deviance DEV (see details above). The variable N represents the number of observations:

$$R2_{DEV}=1-\exp(DEV/N) .$$

Schemper and Henderson (2000) introduced a metric of explained variation:

$$R2_D=\frac{1-D_X(s)}{D_0(s)} ,$$

where $D_0(s)$ and $D_X(s)$ represent criteria of predictive accuracy.

Hielscher et al. (2010) gave a comprehensive overview of R2 metrics including further approaches based on Cox and Snell (1989) and Magee (1990).

4.2.2.5 Further performance criteria

In the survival literature further approaches based on graphical inspection, inferential statistical and score criteria are used to evaluate survival predictions from high-dimensional data. Some popular techniques are:

Kaplan-Meier plots of risk group predictions. This visual method is related to discrimination based criteria, where two or more risk groups are built by risk scores or survival probabilities. High performance models show large survival differences between the risk groups.

Criteria based on the cross-validated log partial likelihood method (Verweij and van Houwelingen, 1993). The cross-validated log partial likelihood approach is the standard criterion to tune survival models from high-dimensional data. Porzelius et al. (2010) used the predictive partial log-likelihood to tune and assess survival models. Nevertheless the partial likelihood has to be applied with care as it depends on the number of patients and can hardly be interpreted.

Reclassification techniques. Popular approaches are net reclassification improvement (NRI) and integrated discrimination improvement (IDI). The NRI and IDI were introduced by Cook and Ridker (2009), a time-dependent version of NRI was presented by Chambless et al. (2011). As far as this is known reclassification techniques were never used to compare survival predictions from high-dimensional data.

Chapter 5

5 Objectives and results of the comparison study

Objectives

In chapter 5 model approaches and validation strategies for survival models based on microarray data are examined. The main objectives of this work are presented in chapter 5.1. The used datasets are described in chapter 5.2. The survival prediction procedure is outlined in chapter 5.3. Results and conclusions are shown in chapter 5.4.

Introduction

Gene expression data are related to survival times (Witten and Tibshirani, 2010) in order to:

- **Identify genes that affect survival:** Popular statistical techniques are the global test approach from Goeman et al. (2004), subset selection techniques or the lasso method for instance.

- **Obtain accurate survival predictions:** the model performance depends on tuning strategies and model approaches like subset selection, variance-based, shrinkage-based and ensemble methods.

The main topics of this thesis are the performance of model fitting techniques, the detection of predictor variables in the survival models and model tuning strategies. Ten model approaches are assessed in this work. These are the forward stepwise selection (FSS), the univariate selection (UPV), the principal components regression (PCR), the supervised principal components regression (SPC), the partial least squares regression (PLS), the ridge regression (RID), the lasso (LAS), the elastic net (NET), the random survival forest (RSF) and the gradient boosting method (GBS).

Bovelstad et al. (2007), van Wieringen et al. (2009), Witten and Tibshirani (2010) and Haibe-Kains et al. (2008) compared model approaches to fit survival models from high-dimensional data. The authors analyzed three or more datasets (except Witten and Tibshirani, 2010). The complexity of the survival model was selected by resampling techniques (Bovelstad et al., 2007, Witten and Tibshirani, 2010, and Haibe-Kains et al., 2008, partly) and the predictions were assessed for data that were not used to develop the survival model. The univariate selection, forward stepwise, principal and supervised principal components approach, the partial least squares method and penalized strategies like the lasso or ridge regression were compared in almost all of the studies (Bovelstad et al., 2007, van Wieringen et al., 2009, and Witten and Tibshirani, 2010).

The investigations revealed some consistent and some complementary results. The partial least squares and the ridge regression approach confirmed to have high performance. The lasso achieved average to low performance (Bovelstad et al., 2007, and van Wieringen et al., 2009). Based on Witten and Bovelstad the univariate selection shows high and low performance and the supervised principal components method good and average results respectively. The capacities of survival forest, boosting approaches and elastic net were seldom compared with other model approaches.

Bovelstad et al. (2007) demonstrated superiority of penalized methods over subset selection and variance-based techniques. The subset selection methods showed less accurate results, variance-based methods achieved only average results. Bovelstad showed, that for some datasets, like the Dutch (van Houwelingen et al., 2006) and Norway/Stanford breast cancer data (Sorlie et al., 2003), the partial least squares method could surpass lasso results. The principal components regression even outperformed the supervised principal components method.

Van Wieringen et al. (2009) compared ensemble methods (random forest and bagging), shrinkage-based approaches (the lasso and ridge regression), variance-based methods (partial least squares and supervised principal components regression) and the univariate selection. The authors demonstrated that ridge regression and the partial least squares approach performed best, tree-based methods and supervised principal components techniques obtained average results and univariate selection showed poor performance. The results of the lasso approach was only marginally better than results from the univariate selection.

In contrast Witten and Tibshirani (2010) found out that univariate selection, supervised principal components methods and shrinkage techniques of higher order revealed a higher performance than the lasso and the principal components regression.

Validation strategies for survival models are further issues of this work. Resampling techniques, 5-, 10- and 20-fold cross-validation, as well as tuning criteria, cross-validated log partial likelihood and the integrated Brier score, are examined.

Bovelstad et al. (2007) and Witten and Tibshirani (2010) for instance used the cross-validated log partial likelihood criterion by Verweij and van Houwelingen (1993) to validate survival models. One disadvantage of the cross-validated log partial likelihood technique is that it can not be applied to nonparametric models (Porzelius et al., 2010). Porzelius et al. (2010) tuned and assessed survival models by applying the integrated prediction error curve (IPEC), a derivation of the Brier score and by predicted partial log-likelihood, an extension of the cross-validated log partial likelihood method. The authors did not detect performance differences between the validation criteria.

Survival models based on high-dimensional data are usually tuned by 10-fold cross-validation (Bovelstad et al., 2007, for instance). The optimal number of subsamples for repeated model validation was hardly examined in the literature. Porzelius et al. (2010) could not detect relevant differences regarding model performance by comparing survival models that were tuned by leaving-one-out and 10-fold cross-validation as well as 10 and 100 bootstrap samples.

In this work the prediction performances of 5-, 10- and 20-fold cross-validation are compared. The leaving-one-out cross-validation (LOOCV) is controversially discussed in the literature and is not considered in this work. Shao (1993) claimed that model selection with the LOOCV technique in linear models is “asymptotically inconsistent” as it does not choose the model with the highest performance for increasing sample sizes. Hastie et al. (2009) (and Braga-Neto and Dougherty, 2004, as well as Xiao et al., 2007, for instance) claimed that N -fold cross-validation may lead to high variances of the tuning criteria and to high computational costs. The authors further argued that in small sample sizes ($N < 100$) 5-fold cross-validation may lead to a considerable prediction bias on the one hand but to low variance of the tuning criterion on the other hand. It needs to be mentioned that these findings were gained from continuous outcome variables.

Subramanian and Simon (2011) examined k = 2-, 5-, 10- and N -fold cross-validation as well as the sample split technique for survival models built from high-dimensional data. They applied the univariate selection, the lasso and the supervised principal components approach to high- and low-signal data of different sample sizes (N = 40, 80 and 160 randomly drawn from the data). The resampling techniques were compared in terms of the mean squared error (MSE) and the bias of AUC predictions. The "real AUC values" were derived from predictions using the whole sample or for large datasets in case of generated data. The N -fold cross-validation has a low mean squared error for high-signal data with low sample size. For increasing N (and for high-signal data) MSE differences disappear. For moderate and low-signal data a low number of validation samples exhibits a lower MSE than LOOCV or sample split. Using no-signal data (and N = 40) the differences are largest, for higher N the differences decrease. Hence a larger number of cross-validation samples for high-signal data and a lower number for low-signal data seem reasonable. Slight advantages appear for 5- and 10-fold cross-validation. Nevertheless the optimal setting depends on the signal strength, the sample size and the model approach.

Tuning settings (k = 5-, 10- and 20-fold cross-validation as well as the IPEC and CVPL criteria) are subordinate research questions in this work.

5.1 Study Objectives

Three issues are addressed in this work: The evaluation of model approaches based on high-dimensional data, the identification of predictor variables in the survival model and the assessment of tuning strategies. Tuning criteria and resampling techniques are investigated in this thesis.

Primary objectives

The performance of model fitting approaches is the primary objective of this work. Subset selection techniques, the forward stepwise selection (FSS) and the univariate selection (UPV), variance-based methods, the principal components regression (PCR) and supervised principal components regression (SPC) and the partial least squares regression (PLS), shrinkage-based approaches, the ridge regression (RID), the lasso (LAS), the elastic net (NET), and ensemble methods, random survival forest (RSF) and the gradient boosting method (GBS), are investigated in this thesis.

Two key questions will be examined:

1) Which model approach performs best?

   - Does the elastic net approach that was seldom compared to other model approaches achieve the same level of performance as the partial least squares or ridge regression?
   - Do the partial least squares, the ridge regression and the supervised principal components methods, which reveal accurate survival predictions in some publications, outperform the other model approaches?
   - Are the results homogeneous or do they present a considerable variance?

2) Which category of model approaches, the subset selection, variance-, shrinkage-based and ensemble techniques, achieves accurate predictions?

   - Do shrinkage-based methods outperform subset selection and derived direction methods as shown in several studies?
   - Is the prediction performance of ensemble techniques comparable to penalty strategies?

Secondary objectives

The secondary objectives of this thesis are dedicated to model validation strategies. The tuning criteria, the cross-validated log partial likelihood and the integrated Brier score, as well as resampling techniques, 5-, 10- and 20-fold cross-validation, are investigated.

Two research questions are examined:

1) Do models tuned by the integrated Brier score achieve the same level of performance as models tuned by cross-validated log partial likelihood?

2) Can survival models validated by 5- and 10-fold cross-validation be surpassed by prediction models tuned by 20-fold cross-validation?

Exploratory objectives

The exploratory objectives of this study examine predictor variables in survival models fitted from the model approaches.

The tertiary hypothesis regarding the generated data is:

- Do the model approaches select the correct variables and how often do they include noise?

The exploratory research question with respect to the real data is:

- Which genes are frequently selected by the model fitting techniques?

Considerations to detect influential genes on survival

The tertiary objective of this work requires the identification of predictor variables in the survival model. Using feature selection approaches like subset selection methods or the lasso approach, the genes of the prediction model can be detected easily but for gene extraction strategies possibly all genes are used to predict survival. Hence a relevance score is created for each gene that expresses how much each feature "contributes" to survival prediction.

For partial least squares, principal components regression, supervised principal components regression, the elastic net and the ridge regression the score values are obtained based on the following considerations. The gene expression data are standardized before the analysis to mean 0 and variance 1.

- The partial least squares method associates PLS components with survival. The components are regressed on the features (Li and Gui, 2004, for details) and the absolute values of the parameter estimates represent the relevance score for each gene.

- In principal components analysis the factor loadings describe the correlation of a factor and a single variable. For the PCR and SPC approach the relevance score is derived from the factor loadings of each gene. The loadings of the components are weighted by the parameter estimates of the principal components in the final survival model. The absolute values of the weighted factor loadings, summed up for each gene, constitute the relevance score for each gene.

- For the ridge regression, the elastic net and the gradient boosting approach, the absolute values of the parameter estimates are used to determine the relevance score for each gene.

For feature extraction approaches only the genes with the highest score values are analyzed to ensure an almost balanced number of features regarding selection and extraction approaches. The generated data consist of ten true effects on survival (see chapter 5.2 for details). The ten most “relevant” variables from the generated data are analyzed. For the forward stepwise selection and univariate selection survival models from microarray data include a median number of two genes. Therefore the two most “relevant” genes are analyzed. Tree models are not analyzed, because reliable algorithms to identify single genes are not available.

However the following issues have to be mentioned:

- Genes with a high impact on survival might not be included in the survival models as correlated data can mask the impact of genes on survival.

- Genes are sorted by the score values in decreasing order and only a small number of the features is analyzed. The number depends on the size of the survival models fitted from forward stepwise and univariate selection.

5.2 Datasets and software

The main objectives of this work are examined on the basis of three data sets. Two data sets are publicly available, the Adult Acute Myeloid Leukemia data (AML) presented by Bullinger et al. (2004) and the diffuse large-B-cell lymphoma data (DLBCL) published by Rosenwald et al. (2002). One further dataset is generated from a Weibull model. Small data gaps in the real data were imputed. In case of large data gaps the variables were excluded from the analysis. The gene expression data were processed and standardized to mean 0 and variance 1. Details are given below.

Missing gene expression data are imputed by the k -nearest neighbor algorithm from Hastie et al. (1999). The procedure works as follows: Suppose that an observation x_{ic} shows one or more missing values and a complete dataset x_c is available.

For every patient with incomplete data x_{ic} the algorithm runs as follows:

1) Identify the k (5 to 10 for instance) nearest neighbors (observations) with a complete dataset. The neighbors are identified by Euclidean distances between x_{ic} and x_c based on non-missing data of x_{ic} .

2) Missing data of x_{ic} are imputed by average values of the nearest neighbors.

Adult Acute Myeloid Leukemia data (AML)

Bullinger et al. (2004) tracked 116 patients suffering from acute myeloid leukemia. Gene expressions were collected from blood and bone marrow samples. The data include 6283 evaluable genes. The median follow-up was 334 days and 611 days for patients who survived. A total of 68 patients died.

Clinical and molecular data are available at the Gene Expression Omnibus Database *www.ncbi.nlm.nih.gov/geo/* by the accession number GSE425 (retrieved: April 30, 2012). The data were transferred to R and processed based on the following considerations:

Subjects with more than 1800 and genes with more than 20 % missing values were excluded from the analysis. The k -nearest neighbor approach (Hastie et al., 1999) was used to impute missing data via the R package *impute*. Finally the gene expression data were standardized to mean 0 and variance 1 since some model approaches are not scale invariant. A similar procedure was applied by van Wieringen et al. (2009).

Diffuse large-B-cell lymphoma data (DLBCL)

Rosenwald et al. (2002) published clinical and genetic data of 240 patients suffering from diffuse large-B-cell lymphoma. The data were collected retrospectively and included 7399 features. The median observation time was 2.8 years; patients that survived (43 %) were monitored for 7.3 years.

Gene expression data and clinical information were obtained by the BRB-array tools data archive developed by Prof. Simon from the National Cancer Institute. The data are available under the web link *http://linus.nci.nih.gov/~brb/DataArchive_New.html* (retrieved: April 20, 2012). No remarkable data gaps were detected and the missing values were completed by the k -nearest neighbor approach of Hastie et al. (1999) using the R package *impute*. 5000 genes with the highest variance were considered for data analysis.

Generated data

The generated dataset includes high-dimensional and high-signal survival data. It is generated (based on the considerations of Subramanian and Simon, 2011) in the following way:

1) 150 uniformly distributed survival times t were generated in an interval between 0.2 and 10.
2) 15 % of the observations were randomly censored using a binomially distributed variable.
3) 10 high-signal variables x_i with $i=1,2,\ldots,10$ were generated by $x_i=\ln(t)/\beta+\epsilon_i$, where $\beta=1.5$ and $\epsilon\sim N(0,0.5)$.
4) Additionally 990 normally distributed variables x_i, $i=11,12,\ldots,1000$, with mean 0 and variance 1 were generated that represent the features with no impact on survival.

The generated dataset consists of 150 observations and 1000 gene features. 10 genes are related to survival.

Software and additional software packages

The open-source software R *http://www.r-project.org/* and Bioconductor *http://www.bioconductor.org/*, were used to analyze the data. The calculations were performed on a terminal server of the “Medizinische Universität Wien” in Vienna, Austria, running Windows Server 2008. Algorithms to process the data and to present the results were prepared on an Linux desktop PC running R 2.14.2 and Bioconductor 2.9.

R / BioC library	Issue	Usage
impute	Data processing	Data imputation
survpack	Model building	Partial least squares method
penalized	Model building	Shrinkage-based approaches
party	Model building	Random survival forest
mboost	Model building	Gradient-boosting
survcomp	Model assessment	Brier score, AUC, deviance
Cairo	Plots	Graphics device

Table 5-1: R and Bioconductor (BioC) packages used to prepare and analyze the data. Single packages included in base R are not listed.

Base R, inter alia, the *base*, *stats*, *graphics* and *survival* package and additional packages were installed from *http://cran.at.r-project.org/* and from *http://www.bioconductor.org/*. The R package *survpack* was taken from *http://www.stats.gla.ac.uk/~levers/research.html* (retrieved: April 12, 2012). A list of R packages used to analyze the data is shown in table 5-2. Only the main packages are listed, dependencies between R packages are not shown.

5.3 Strategies for data analysis

The research questions are examined in the following way: A survival prediction procedure (based on Bovelstad et al., 2007, and described in chapter 3.4) is used to tune, fit and assess survival models based on gene expression data. The procedure optimizes the performance of model approaches (primary objective of this thesis) with respect to the tuning parameter λ , whereas

1) different datasets (high-signal and low-signal data, preselected and non-preselected data),
2) different tuning criteria (cross-validated log partial likelihood, integrated Brier score) and
3) different resampling techniques (5-, 10- and 20-fold cross-validation)

are used to fit the survival models. The prediction accuracy of each constellation (consisting of a model approach and a tuning strategy) is measured for three datasets using several performance metrics (IAUC, IBSC, DEV and R2). The main objectives of this work are evaluated from these results.

The ten model approaches are:

1) Univariate selection (UPV),
2) Forward stepwise selection (FSS),
3) Principal components regression (PCR),
4) Supervised principal components regression (SPC),
5) Partial least squares (PLS),
6) Lasso (LAS),
7) Ridge regression (RID),
8) Elastic net (NET),
9) Boosting (GBS),
10) Random survival forest (RSF).

The generated data are analyzed in chapter 5.4.1, the AML data in chapter 5.4.2 and the DLBCL data in chapter 5.4.3.

Optimization

The survival models are tuned in the following way: Assume that the performance (PERFORM) of a survival model depends on the dataset (DS), the model approaches (MA), the tuning parameter λ and the validation setting V(K,C), where K represents k-fold cross-validation and C the validation criterion. The goal is to maximize the model performance:

$$PERFORM(DS, MA(\lambda), V(K, C)) \rightarrow max$$

with regard to the tuning parameter λ . The complexity parameter has different meanings for the model approaches: Using UPV and FSS λ is associated with the number of features in the survival model, regarding PCR and PLS with the number of components, with respect to LAS and RID with the amount of shrinkage, with regard to RSF with the number of survival trees and regarding GBS with the width of the boosting steps. Using SPC the tuning parameter is related to the number of preselected features and the number of principal components in the prediction model. With respect to NET it is associated with the amount of L1 and L2 shrinkage.

Algorithm

The course of the survival prediction procedure is summarized in table 5-2. The arguments of the prediction algorithm are the dataset, the model approach and the tuning parameter:

1) **m = 1**, where m is a counter that represents the number of repetitions.

2) **While m <= 50** the following steps are performed:

3) **Random split of the data**: 2/3 of the observations of each dataset are used to tune and fit the prediction model and 1/3 of the subjects to evaluate the survival predictions.

4) **Complexity selection**: The following steps are applied to the learning data for a set of tuning parameter values λ_i (I = 1,2,...):
   I. 5-, 10- and 20-fold cross-validation are performed:
      1. k-1 samples are used to fit a model based on λ_i and the
      2. k-th sample to validate the survival predictions.
   II. The optimal model size $\hat{\lambda}$ is assessed with the metrics CVPL (high values are better) and IBSC (low values are better). This leads to 6 “performance-optimal” models (from all combinations of CVPL/IBSC and 5-fold/10-fold/20-fold cross-validation).

5) **Model fitting**: The parameter values β_{train} (for each of the 6 models) are estimated using the whole learning sample given the model size $\hat{\lambda}$.

6) The tertiary hypothesis is examined for the survival models of 5.

7) **Survival prediction**: A risk score is derived from $\hat{\beta}_{train}$ using the test data (see details above). The performance of the survival predictions is assessed with IAUC, IBSC, DEV and R2.

8) **M = m + 1** and go to 2.

Table 5-2 presents the course of the survival prediction procedure. The steps are applied 50 times.

Data part	Operation	Complexity selection based on	Model assessment based on	Results
Training data (2/3 N)	Model selection	5- / 10- / 20- fold CV and CVPL & IBSC tuning criterion		Tuned survival model
	Model assessment		Influential genes on survival	Key features in the survival model (exploratory endpoint)
Test data (1/3 N)	Model assessment		IAUC, IBSC, Deviance, R2	Findings in benchmark study (primary, secondary objectives)

Table 5-2: Survival prediction procedure to fit and assess survival prediction models. The main steps are: data splitting, model building (based on the resampling techniques 5-, 10- and 20-fold CV and the validation criteria CVPL versus IBSC) and model evaluation (with IAUC, IBSC, DEV and R2).

Analysis

The research questions are investigated in the following way:

1) **Primary objective**: The prediction accuracy of the model approaches is assessed based on IAUC, IBSC, DEV and R2. Only the results from 10-fold cross-validation are taken into account. The full dataset and 500 preselected features are used to analyze the primary objective.

2) **Secondary objectives**:

   a) **Tuning criterion**: The performance of the CVPL and IBSC tuning criterion is assessed in terms of IAUC, IBSC, DEV and R2. Results are displayed for all model approaches.

   b) **Validation samples**: The accuracy of 5-, 10- and 20-fold cross-validation is evaluated with IAUC, IBSC, DEV and R2. The results are outputted for all model techniques.

3) **Tertiary objective**: Influential genes on survival are analyzed for all model approaches. The survival models were tuned by 10-fold cross-validation and the model size was chosen by the CVPL criterion.

Presentation

Results are presented using median tables and boxplots of the performance values (primary and secondary objectives). Frequency tables are used to depict the influential genes on survival (tertiary objective).

Comments

The partial least squares and the random survival forest approach led to serious algorithmic and model stability problems using datasets with more than 1000 features. Since model comparison of only eight approaches would reduce the validity of the primary objective, another process was performed.

The full dataset and 500 preselected data are used to analyze the primary objective:

1) Full dataset

   a) Eight model approaches are evaluated using the real data.

   b) Nine model techniques are applied to the generated data.

   Subset selection techniques, the principal and supervised principal components regression, shrinkage-based approaches and the gradient boosting method are compared using the AML dataset (4724 features), the DLBCL dataset (5000 features) and the generated dataset (1000 variables). All models except the PLS and RSF approach are applied to the real and all techniques except the PLS method to the generated data.

2) 500 pre-selected features

All model approaches are compared using generated data with 10 high-signal and 490 random variables as well as 500 pre-selected variables of the real data.

The preselection procedure was performed in the following way: Univariate selection was applied to 100 random subsamples of size 0.9* N (based on the considerations of Hastie et al., 2009). Univariate Cox models were used to obtain p-values based on the Score test. The genes within each subsample were sorted by significance. 500 genes with the lowest global rank index, as a sum of single ranks from the data subsamples, were selected.

5.4 Results

The results section is structured as follows: The generated data are analyzed in the first subchapter, the Adult Acute Myeloid Leukemia data (AML; Bullinger et al., 2004) in the second section and the diffuse large-B-cell lymphoma data (DLBCL; Rosenwald et al., 2002) in the third subchapter. Each section includes the results of the primary objective using the full data as well as the pre-selected data, the secondary and the exploratory objectives.

5.4.1 Results for the generated data

The generated dataset contains 150 observations and 10 high-signal variables (see 5.2 for details). The original setting consists of 990, the reduced dataset of 490 noise variables.

5.4.1.1 Analysis of the primary objectives

The first analysis refers to the prediction performance of nine model approaches using the dataset with 1000 variables. The accuracy of the forward stepwise selection (FSS), the univariate selection (UPV), the principal components regression (PCR), supervised principal components regression (SPC), the ridge regression (RID), the lasso (LAS), the elastic net (NET), the random survival forest (RSF) and the gradient boosting method (GBS) are evaluated. The partial least squares approach (PLS) is only applied to the dataset with 500 variables.

The survival models are tuned by 10-fold cross-validation and either the cross-validated log partial likelihood (CVPL) or the integrated Brier score (IBSC). The prediction performance of the models is assessed with the integrated area under the time-dependent ROC curve (IAUC), the integrated Brier score (IBSC), the deviance of the fitted model and the null model (DEV) and the explained variation (R2) based on the Brier score (Graf et al., 1999, and Hielscher et al., 2010). High values of IAUC and R2 and low values of IBSC and DEV indicate high prediction performance.

The results of the primary objective for the full dataset are shown in table 5-3 and figure 5-1. The median values of the assessment criteria for the model fitting techniques are shown in table 5-3. The best three approaches for the performance metrics are presented in bold numbers. The four right columns represent the results for the IBSC tuning criterion and the four left columns the values for the CVPL tuning criterion. Figure 5-1 displays the results by a boxplot matrix. The four right boxes are dedicated to models tuned by IBSC and the four left boxes to models validated by CVPL.

The SPC, FSS and UPV approaches achieve the best performances. LAS, GBS and NET are almost as effective. PCR, RID and RSF present lower performances. NET and RSF revealed high variability of the performance metrics. The other approaches showed moderate variance. Only minor differences were detected between models tuned by IBSC and CVPL and between results assessed with the performance criteria IAUC, IBSC, DEV and R2.

	CVPL				**IBSC**			
	IAUC	**IBSC**	**DEV**	**R2**	**IAUC**	**IBSC**	**DEV**	**R2**
UPV	**0.820**	**0.099**	**-40.85**	**0.514**	**0.823**	**0.095**	**-43.83**	**0.529**
FSS	**0.821**	**0.093**	**-45.45**	**0.526**	**0.837**	**0.091**	**-51.78**	**0.543**
PCR	0.677	0.152	-6.08	0.134	0.698	0.143	-8.70	0.181
SPC	**0.831**	**0.092**	**-52.17**	**0.563**	**0.834**	**0.090**	**-53.16**	**0.564**
LAS	0.811	0.100	-36.26	0.483	0.808	0.100	-31.83	0.487
RID	0.679	0.149	-1.86	0.151	0.673	0.150	-0.01	0.135
NET	0.784	0.111	17.82	0.409	0.726	0.133	-5.26	0.250
RSF	0.642	0.160	6.44	0.068	0.730	0.141	-2.69	0.167
GBS	0.799	0.104	-33.75	0.462	0.817	0.098	-34.57	0.492

Table 5-3: Median performance of nine model approaches (rows) obtained from 50 random splits of the generated data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 10-fold CV and either the CVPL (four columns on the left) or IBSC tuning criterion (four columns on the right). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Results described in detail:

- The UPV, FSS, SPC, LAS, GBS and the NET approach show high median IAUC values (> 0.78 for models tuned by CVPL and > 0.72 for models selected by IBSC). The median IAUC of the PCR, RID and the RSF techniques range from 0.64 to 0.73. The variability of the IAUC values is low (inter quartile range IQR ranges from 0.03 to 0.06) for almost all approaches; only NET achieves an IQR up to 0.10 for models that were tuned by IBSC.

- UPV, FSS, SPC, LAS and GBS show low IBSC values. The median IBSC value is between 0.09 and 0.11, whilst PCR, RSF and RID lay above 0.14. NET presents lower IBSC values for models tuned by CVPL than for models validated by the IBSC criterion. The model techniques demonstrate a moderate variation of the results (IQR ranges from 0.01 to 0.02). Only the NET and the RSF method for models tuned by IBSC reveal an IQR marginally above 0.02.

- Almost all approaches show a deviance lower than 0. Only NET and RSF approaches achieve positive values for models validated by CVPL and are minimally below 0 for models selected by IBSC. The median deviance values of the RID and PCR approaches range from -9 to 0. UPV, FSS, SPC, LAS and GBS demonstrate high prediction performances (DEV values range from -53 to -32). NET shows a high IQR of about 80 for models tuned by CVPL, otherwise the IQR of the results is lower than 25.

- The median R2 values of the UPV, FSS, SPC, LAS and GBS are high (range from 0.46 to 0.56 for models validated by CVPL and between 0.49 and 0.56 for models selected by IBSC). PCR, RSF and RID values range from 0.07 to 0.15 (models tuned by CVPL) and from 0.14 and 0.18 (models selected by IBSC). NET presented average results (0.41 based on models tuned by CVPL and 0.25 based on models selected by IBSC). All model techniques except NET show a moderate variation of the R2 values (IQR range from 0.05 to 0.12). The IQR of NET lays above 0.20 for models chosen by IBSC.

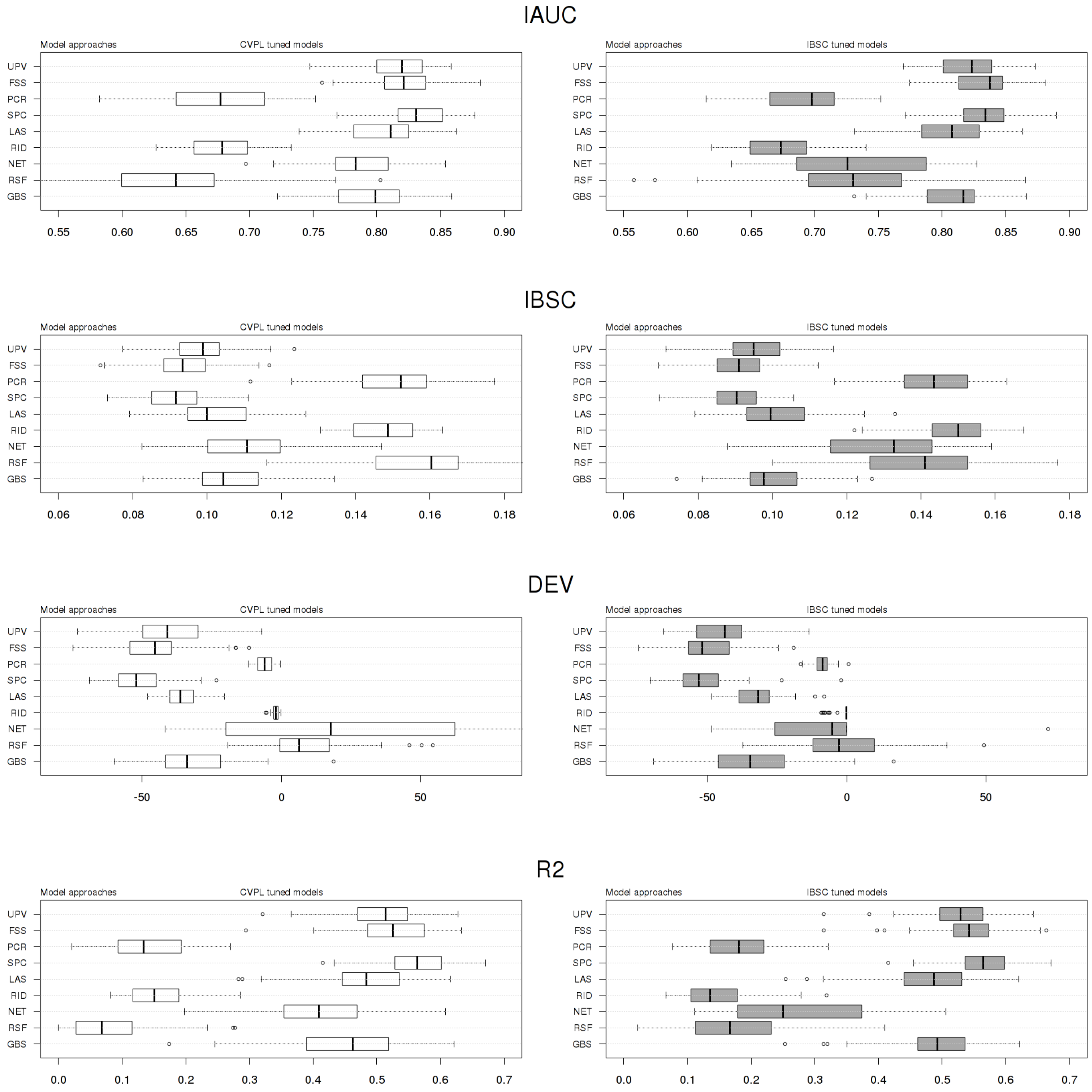

Figure 5-1: Performance of nine model approaches obtained from 50 random splits of the generated data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2. The survival models are validated by 10-fold CV and either the CVPL or IBSC tuning criterion. High prediction performance is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Analysis of the primary objective using the dataset with 500 variables

The second part of the primary analysis is related to the evaluation of ten model approaches regarding the dataset with 10 high-signal and 490 uninformative variables. Table 5-4 and Figure 5-2 present the s using the dataset with 500 variables.

	CVPL				IBSC			
	IAUC	IBSC	DEV	R2	IAUC	IBSC	DEV	R2
UPV	0.820	0.099	-40.85	0.514	0.823	0.095	-43.83	0.529
FSS	0.821	0.093	**-45.45**	**0.526**	**0.838**	0.091	**-52.15**	**0.540**
PCR	0.826	0.093	-39.48	0.464	0.831	**0.088**	-46.09	0.477
SPC	0.831	0.092	**-52.17**	**0.563**	0.834	0.090	**-53.16**	**0.564**
PLS	**0.857**	**0.075**	**-60.27**	**0.546**	**0.857**	**0.075**	**-60.27**	**0.546**
LAS	0.812	0.099	-36.95	0.487	0.813	0.097	-35.04	0.488
RID	**0.844**	**0.084**	-20.66	0.501	**0.840**	**0.087**	-0.04	0.480
NET	**0.832**	**0.089**	-21.29	0.494	0.834	0.090	-14.03	0.481
RSF	0.660	0.159	9.38	0.074	0.728	0.142	-7.01	0.187
GBS	0.799	0.104	-34.14	0.462	0.818	0.097	-34.53	0.493

Table 5-4: Median performance of ten model approaches (rows) obtained from 50 random splits of the generated data including 500 features. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 10-fold CV and either the CVPL (four columns on the left) or IBSC tuning criterion (four columns on the right). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

All model techniques except the RSF approach reveal high prediction performances. The PLS method outperforms the other model approaches. RID, NET, UPV, FSS, PCR, SPC, LAS and GBS seem to achieve the same level of performance. The results of the RSF method and the NET approach show a high variation of the deviance values. The other methods reveal a low variability of the performance values.

Results described in detail:

- All model approaches except the RSF technique present accurate predictions. The median IAUC values from all approaches except the RSF method range from 0.80 to 0.86 for survival models tuned by CVPL and from 0.81 to 0.86 for models validated by IBSC. The median values of the RSF approach are 0.66 and 0.73, the median values of PLS 0.86 and 0.86. The IQR of the IAUC values ranges from 0.03 to 0.05. Only the IQR of the RSF methods is 0.1 and 0.05.

- The median IBSC values of all techniques except RSF range from 0.10 to 0.08. The RSF approach achieves 0.16 for models chosen by CVPL and 0.14 for models selected by IBSC. The variance of the IBSC results ranges from 0.01 to 0.02 and is marginally above 0.02 for RSF models chosen by IBSC.

- The median deviance values for the RSF approach are positive for survival models validated by CVPL. UPV, FSS, PCR, SPC, PLS, LAS and GBS results range from -60 to -34 (PLS achieves the lowest median value). RID and NET present median values of about -20 for models validated by CVPL as well as 0 and -14 for models selected by IBSC. The performance of the model approaches reveal a moderate variability (IQR up to 25), only the variance of NET results was higher (IQR about 50).

- The median R2 values of the UPV, PCR, LAS, RID, NET and GBS techniques range from 0.46 to 0.51 for models tuned by CVPL and from 0.48 to 0.53 for models selected by IBSC. R2 is high for SPC (0.56/0.56), PLS (0.55/0.55) and FSS (0.53/0.54) approaches. The median values of RSF are low (0.07/0.19). The variance of R2 is below 0.10. Only RSF and GBS are marginally above 0.10.

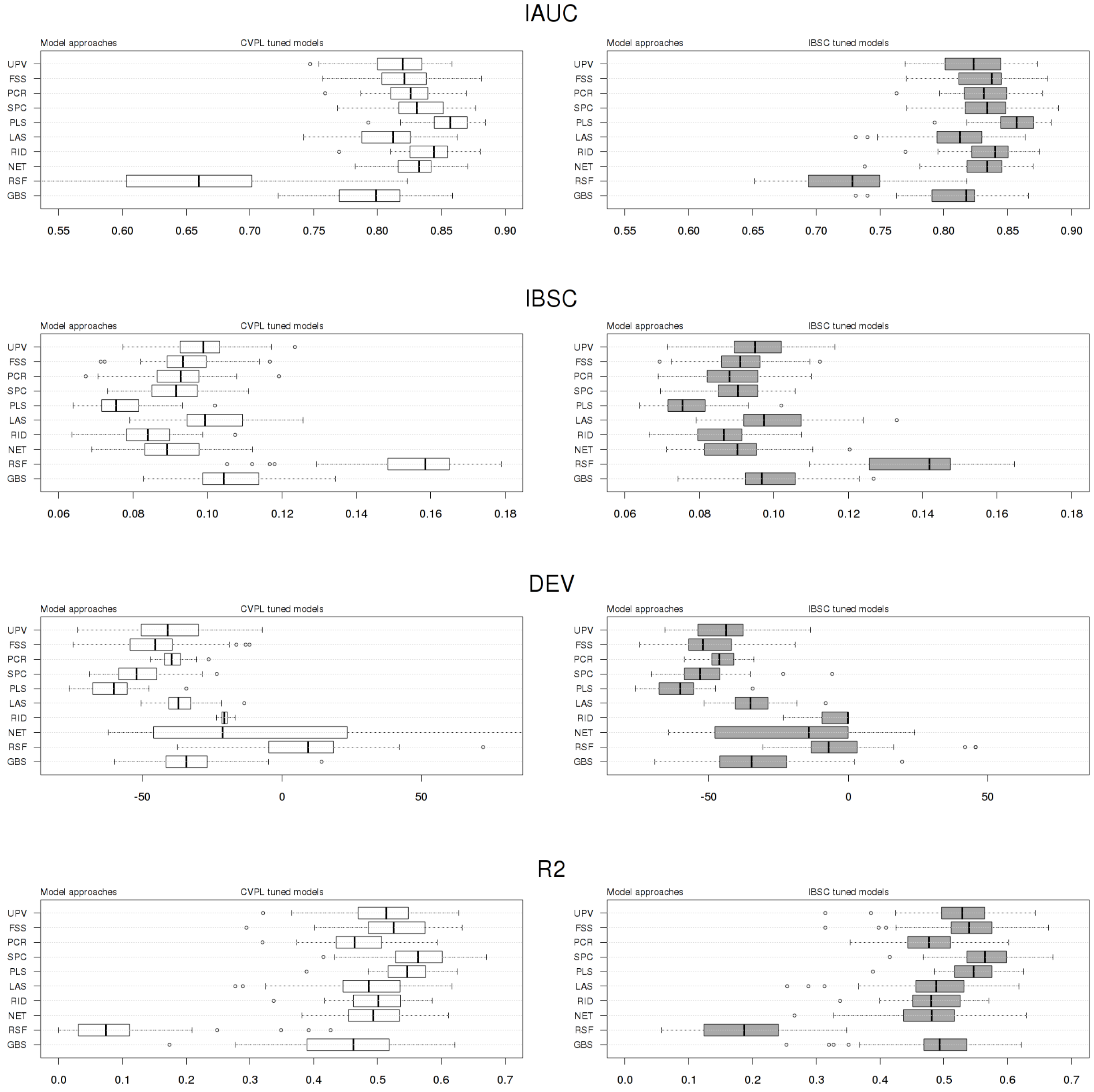

Figure 5-2: Performance of ten model approaches obtained from 50 random splits of the generated data including 500 features. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2. The survival models are validated by 10-fold CV and either the CVPL or IBSC tuning criterion. High prediction performance is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Conclusions

The subset selection methods achieve high prediction performance using the full dataset. Subset selection, variance- and shrinkage-based approaches perform best for the reduced dataset with 500 variables. Ensemble approaches reveal low prediction performance. The prediction accuracy of partial least squares surpasses the other approaches. The results present a low variance. The performance of PLS remains unknown for datasets with more than 1000 variables.

Feature selection approaches perform well for both simulated datasets. SPC has high prediction accuracy and is the most powerful approach for the dataset with 1000 variables. SPC, UPV, FSS, LAS and GBS are the best five methods for the full dataset and perform well for the data of reduced dimensionality.

PCR and RID show low performance for the full dataset and high performance using the dataset with 500 variables. The RSF method presents a considerable spread of the results and a low median performance. The NET approach has intermediate results using the full dataset and high performance for the dataset with 500 variables. The NET approach achieves a considerable variation.

Aside PLS, feature selection approaches demonstrate superiority over extraction methods. Using the dataset with 500 variables the performance of the model approaches present marginal differences. The performance criteria of the model techniques do not show a high variability. Only RSF has a low performance and NET shows a high spread of the results.

5.4.1.2 Analysis of the secondary objectives

The secondary objectives of this work are dedicated to model validation strategies, particularly resampling techniques for model validation and tuning criteria. The performance of the tuning strategies is assessed in terms of IAUC, IBSC, the deviance and the R2 criterion.

Analysis of resampling techniques

The first analysis refers to resampling techniques, in particular 5-, 10- and 20-fold cross-validation. The results of the secondary objective are presented in table 5-5 and figure 5-3. The median values of the performance criteria for the resampling techniques are shown in table 5-5. The best resampling technique for each model approach and for each performance metric is presented in bold numbers. Figure 5-3 displays the results for each model fitting technique and for each performance value in a boxplot matrix. The resampling techniques are analyzed using the full dataset. Only the values from the PLS approach are computed using the dataset with 500 variables.

	IAUC			**IBSC**			**DEV**			**R2**		
	5-fold	**10-fold**	**20-fold**	**5-fold**	**10-fold**	**20-fold**	**5-fold**	**10-fold**	**20-fold**	**5-fold**	**10-fold**	**20-fold**
UPV	0.825	0.822	**0.833**	0.094	0.097	**0.091**	-44.08	-41.98	**-48.09**	0.537	0.524	**0.543**
FSS	0.829	0.826	**0.831**	**0.090**	0.092	0.091	-44.30	-50.64	**-51.43**	0.544	0.530	**0.548**
PCR	0.682	0.692	**0.694**	0.147	0.148	**0.146**	-6.36	**-8.05**	-7.37	0.150	0.161	**0.163**
SPC	**0.841**	0.832	**0.841**	**0.086**	0.091	0.087	-53.93	-52.79	**-55.49**	**0.578**	0.564	0.568
PLS	0.851	0.857	**0.863**	0.077	**0.075**	**0.075**	-57.22	-60.27	**-63.00**	0.533	0.546	**0.557**
LAS	0.811	0.810	**0.816**	0.100	0.100	**0.098**	-33.91	-35.02	**-35.25**	0.485	0.485	**0.492**
RID	**0.681**	0.677	**0.681**	**0.147**	0.149	0.148	-0.08	**-0.92**	-0.24	**0.153**	0.148	0.145
NET	0.761	0.770	**0.772**	0.121	0.119	**0.116**	-0.01	-0.01	**-1.51**	0.340	0.357	**0.375**
RSF	0.670	**0.687**	0.678	0.152	**0.148**	0.149	6.58	3.93	**1.75**	0.104	**0.115**	0.108
GBS	0.805	0.811	**0.816**	0.102	0.102	**0.097**	-33.79	-34.16	**-39.06**	0.476	0.483	**0.487**

Table 5-5: Median performance of the resampling methods 5-, 10- and 20-fold cross-validation obtained from 50 random splits of the generated data. Survival predictions are made based on ten model approaches (rows). The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The resampling technique with the highest performance is shown in bold numbers and is characterized by high IAUC and R2 as well as low IBSC and DEV values.

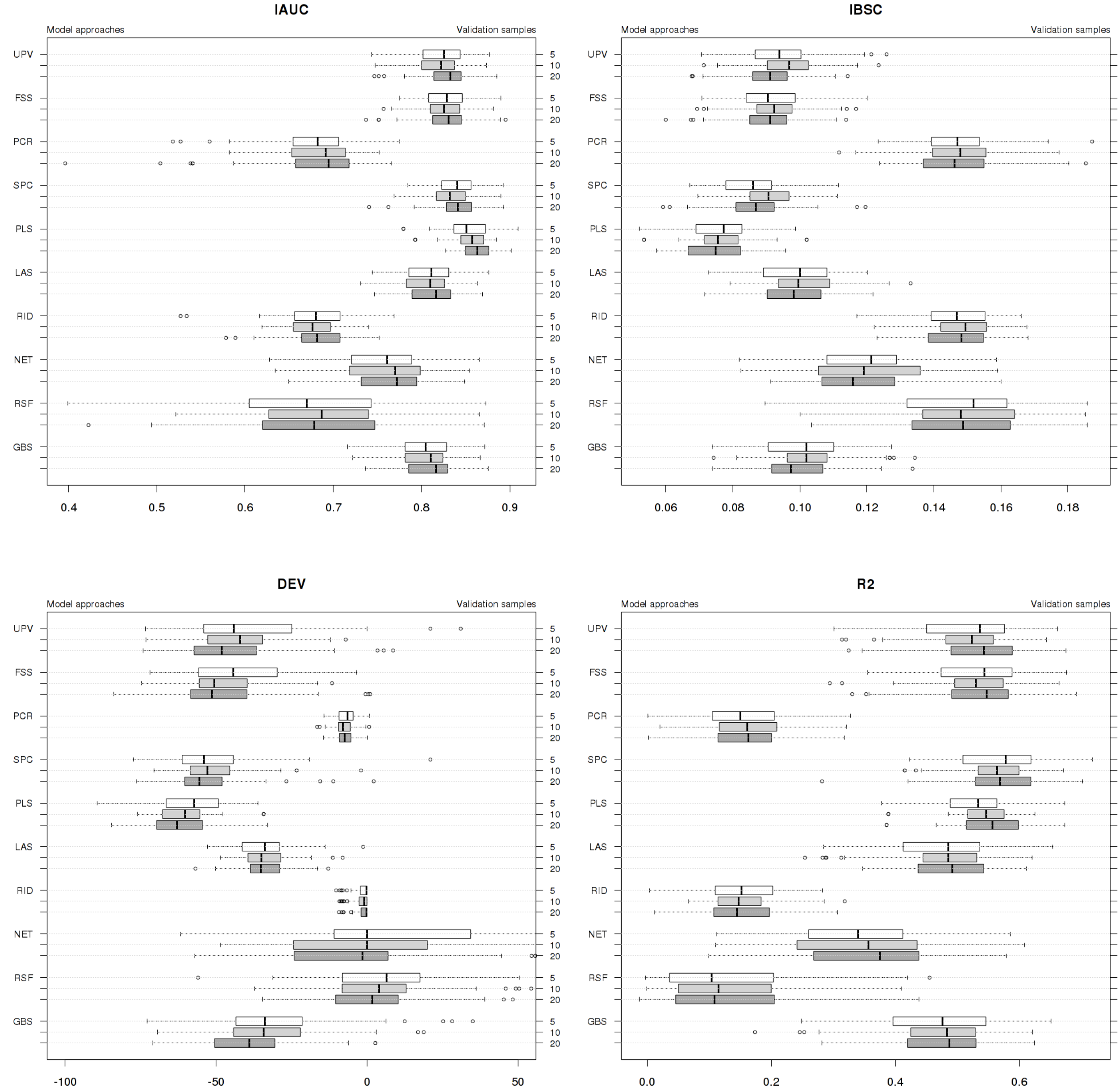

Figure 5-3: Performance of the resampling methods 5-, 10- and 20-fold cross-validation obtained from 50 random splits of the generated data. Survival predictions are made based on ten model approaches. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (boxes). High prediction performance is characterized by high IAUC and R2 as well as low IBSC and DEV values.

The performance of the 5- and 10-fold cross-validation technique is almost equal and is marginally outreached by 20-fold cross-validation. The accuracy of 20-fold cross-validation is higher for almost all model approaches and for all performance criteria. The variance between the resampling techniques is almost equal.

Results described in detail:

- **With respect to median IAUC values**: 20-fold cross-validation performs best for 9 approaches, 10-fold cross-validation for 1 model fitting technique and 5-fold CV for 2 model approaches.

- **Regarding median IBSC results:** 20-fold cross-validation achieves the highest prediction accuracy for 6 model approaches, 10-fold cross-validation for 2 and 5-fold cross-validation for 3 model techniques. The differences of the median IAUC and the IBSC values between the two best resampling techniques are lower than 0.01.

- **In terms of median DEV values:** 20-fold cross-validation shows the highest performance for 8 model approaches, 10-fold cross-validation for 2 and 5-fold cross-validation for 0 approaches. The differences of median DEV values between the two best resampling techniques are lower than 3.

- **With regard to median R2 values:** 20-fold cross-validation presents the highest prediction accuracy for 7 approaches, 10-fold cross-validation for 1 and 5-fold cross-validation for 2 model fitting techniques. The differences of median R2 values between the two best resampling techniques are lower than 0.02.

Analysis of the tuning criterion

The second research question of the secondary hypothesis examines the accuracy of the tuning criteria CVPL and IBSC. The results of the secondary objective are displayed in table 5-6 and in figure 5-4. Table 5-6 presents the median values of the performance metric for the tuning criteria. The best value of the validation metric for each model approach and for each performance metric is presented in bold numbers. Figure 5-4 displays the results for each model fitting technique and for each performance criterion in a boxplot matrix. The resampling techniques are applied to the full dataset. Only the values from the PLS approach are calculated using the dataset with 500 variables.

Only small differences appear between the performances of the CVPL and IBSC tuning criteria. Using the PLS, SPC and LAS approaches the performances of the validation criteria are almost equal. IBSC surpasses CVPL applying the UPV, FSS, PCR, RSF and GBS technique. CVPL outperforms IBSC only using the shrinkage-based approaches RID and NET. The deviance values of the NET approach shows high variance for the CVPL criterion.

	IAUC		**IBSC**		**DEV**		**R2**	
	CVPL	**IBSC**	**CVPL**	**IBSC**	**CVPL**	**IBSC**	**CVPL**	**IBSC**
UPV	0.822	**0.829**	0.096	**0.092**	-41.79	**-46.48**	0.523	**0.541**
FSS	0.824	**0.836**	0.093	**0.089**	-44.53	**-51.41**	0.532	**0.548**
PCR	0.673	**0.698**	0.151	**0.144**	-5.74	**-8.74**	0.134	**0.176**
SPC	0.838	**0.840**	0.088	**0.087**	-53.99	**-54.31**	0.565	**0.571**
PLS	**0.857**	**0.857**	**0.076**	**0.076**	**-60.60**	**-60.60**	**0.546**	**0.546**
LAS	**0.812**	**0.812**	0.100	**0.098**	**-35.74**	-32.31	0.483	**0.489**
RID	**0.681**	0.677	**0.148**	0.149	**-1.58**	-0.01	**0.152**	0.138
NET	**0.784**	0.732	**0.110**	0.128	17.23	**-6.15**	**0.407**	0.260
RSF	0.626	**0.734**	0.160	**0.139**	8.03	**-2.76**	0.050	**0.183**
GBS	0.804	**0.819**	0.104	**0.097**	-34.30	**-36.58**	0.462	**0.494**

Table 5-6: Median performance of the tuning criteria CVPL and IBSC obtained from 50 random splits of the generated data. Survival predictions are made based on ten model approaches (rows). The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The tuning criterion with the highest performance is shown in bold numbers and is characterized by high IAUC and R2 as well as low IBSC and DEV values.

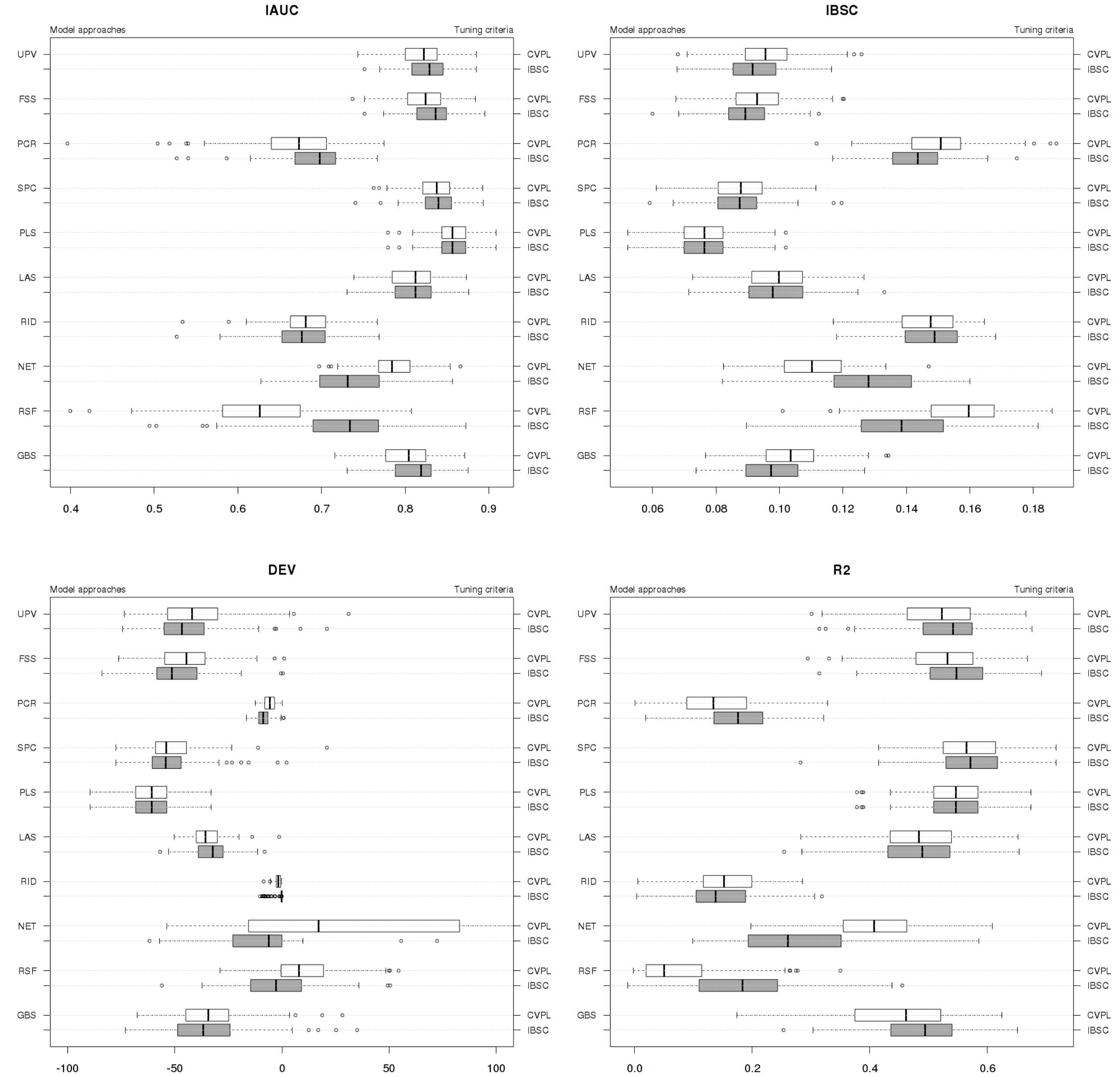

Figure 5-4: Performance of the tuning criteria CVPL and IBSC obtained from 50 random splits of the generated data. Survival predictions are made based on ten model approaches. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (boxes). High prediction performance is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Results described in detail:

- **With reference to median IAUC performance values:** The IBSC tuning criterion demonstrates superiority over CVPL for 6 approaches (differences < 0.11). CVPL is better than the IBSC validation criterion for 2 approaches (differences <= 0.05). The IQR is almost equal. Only survival predictions from the NET approach and the IBSC tuning criterion achieve a higher variation.

- **In terms of median IBSC performance results:** The IBSC tuning criterion performs best for 8 and CVPL for 3 model approaches. The differences between the validation criteria are lower than or equal to 0.02. The variance of the IBSC tuning criterion is higher applying NET and RSF approaches.

- **With regard to median DEV results:** The IBSC validation criterion achieves the highest performance for 8 approaches (differences < 24) and CVPL for 3 model techniques (differences < 4). Using NET approaches the IQR of CVPL is much higher than the IQR of IBSC. The median CVPL is higher than 0 performing RSF and NET model techniques, whilst the median IBSC is lower than 0.

- **With reference to median R2 values:** The IBSC tuning criterion has a higher accuracy than CVPL for 7 model techniques (differences < 0.14) and lower performance for 2 approaches (differences < 0.15). IBSC has a high variance using NET and RSF approaches.

5.4.1.3 Analysis of the exploratory objective

The exploratory hypothesis examines whether the model fitting techniques find the correct survival model. The number of correct variables and the number of noise variables in survival prediction models are determined. Using feature selection methods the explanatory variables in the survival models can be identified directly. Applying gene extraction techniques heuristics are used to detect variables that predict survival (details are given in chapter 5.1).

Nine model fitting techniques are used to build survival models from 50 random splits of the data. The models are tuned by 10-fold cross-validation and the CVPL criterion. The survival models are investigated using the full dataset. The PLS approach is applied to the dataset with 500 variables.

Table 5-7 presents the frequency of the correct variables (**SIGNAL**: F1 to F10) and of noise variables (**NOISE**: F11 to F500 or F1000 respectively) included in 50 survival models. Performing feature extraction strategies, only the 10 covariates with the highest relevance score are considered. This means a total of 500 variables. The UPV approach includes 441, FSS 413 and GBS 330 variables in 50 survival models.

The model approaches detect some variables frequently: F7 (408/450, 90.67 %), F4 (90.00 %), F6 (87.11 %), F3 (83.33 %), F10 (76.44 %), F9 (76.22 %), F2 (71.33 %). Other variables are selected in a maximum of two of three models: F8 (67.78 %), F1 (65.11 %) and F5 (53.78 %).

VARIABLE	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	SIGNAL	NOISE
UPV	38	37	45	50	26	48	50	39	49	42	424	17
FSS	50	50	50	50	50	45	50	29	17	15	406	7
PCR	8	2	8	9	5	13	8	12	7	10	82	418
SPC	50	48	50	50	35	50	50	46	50	50	479	21
PLS	50	50	50	50	43	50	50	50	50	48	491	9
LAS	23	36	43	50	15	47	50	22	38	42	366	134
RID	42	40	50	50	27	49	50	38	50	49	445	55
NET	32	43	46	50	30	49	50	42	45	49	436	64
GBS	0	15	33	46	11	41	50	27	37	39	299	31

Table 5-7: High-signal (SIGNAL: sum of the high signal variables F1-F10) and noise variables (NOISE) selected by ten model approaches (rows) obtained from 50 random splits of the generated data. The survival models are validated by 10-fold cross-validation and the CVPL tuning criterion.

The model approaches are assessed by the ratio of correct variables and incorrect variables in the survival models. A high rate of high-signal variables is detected using FSS (proportion of correct variables: 98.31 % / absolute number of informative variables: 406 / absolute number of non-informative variables: 7), PLS (98.20 % / 491 / 9), UPV (96.15 % / 424 / 17), SPC (95.80 % / 479 / 21), GBS (90.61 % / 299 / 31), RID (89.00 % / 445 / 55) and NET (87.20 % / 436 / 64) approaches. LAS (73.20 % / 366 / 134) achieves a lower proportion of relevant variables and PCR shows (16.40 % / 82 / 418) poor results.

Results described in detail:

- Survival models built from UPV approaches include two high-signal features in all 50 survival models. Each of the correct variables is included in more than 25 models. Models from FSS approaches select the right variables in at least 15 models and 6 correct variables are included in all 50 survival models. FSS incorporates the lowest number of noise variables (7).

- The prediction models fitted from variance-based approaches achieve the following results: The correct variables are included in 2 to 13 models (PCR), in 35 to 50 models (SPC) and at least 43 models (PLS), whilst 418 (PCR), 21 (SPC) and 9 (PLS) noise variables are incorporated in the survival models. 0 (PLS), 7 (SPC) and 8 (PLS) correct variables are included in all 50 survival models. PCR reveals poor results, which may be due to the fact that PCR is an unsupervised model approach and outcome and predictor variables are associated by chance.

- Survival models from shrinkage approaches include the correct variables at least 15 (LAS), 27 (RID) and 30 times (NET). LAS and NET incorporate 2 variables in all models and RID even 4 variables. 134 (LAS), 55 (RID) and 64 (NET) random variables are selected.

- Using GBS approaches one correct variable is never included in the survival models. The other informative variables are selected at least 11 times. One variable is included in all survival models and 31 noise variables are selected.

- In accordance with the ratio of correct and incorrect variables in the survival models, the FSS, UPV, SPC and PLS model fitting techniques achieve high prediction performance and NET, LAS, PCR average or low performance. A high number of correct variables are included in survival models fitted from GBS and RID approaches. But the models reveal a medium or low prediction performance.

5.4.2 Results for the AML data

The AML dataset (Adult Acute Myeloid Leukemia data; Bullinger et al., 2004) consist of 103 patients and 4724 gene expressions. The primary objective is examined using either the full dataset or 500 pre-selected features. Secondary and tertiary research questions are investigated using the full AML dataset. The PLS and RSF approaches are only applied to the AML dataset with 500 variables.

5.4.2.1 Analysis of the primary objectives

The first analysis refers to the prediction accuracy of eight model techniques UPV, FSS, PCR, SPC, LAS, RID, NET and GBS using the full AML dataset. The prediction models are validated by either the cross-validated log partial likelihood (CVPL) or the integrated Brier score (IBSC) and by 10-fold cross-validation. The integrated area under the ROC curve (IAUC), the integrated Brier score (IBSC), the deviance (DEV) and the explained variation (R2) based on the Brier score are used to evaluate the performance of the survival models.

The results are displayed in table 5-8 and figure 5-5. Median values of the performance criteria for the model approaches are presented in table 5-8. The best three techniques for the performance metrics are shown in bold numbers. The four right columns represent the results for the IBSC tuning criterion and the four left columns the values for the CVPL tuning criterion. Figure 5-5 presents the results by a boxplot matrix. The four right boxes are dedicated to models tuned by IBSC and the four left boxes to models validated by CVPL.

	CVPL				**IBSC**			
	IAUC	**IBSC**	**DEV**	**R2**	**IAUC**	**IBSC**	**DEV**	**R2**
UPV	0.575	0.204	6.41	0.021	0.592	0.193	4.93	0.046
FSS	0.536	0.208	5.47	0.011	0.572	0.202	4.97	0.024
PCR	0.573	0.204	**-0.42**	0.021	0.608	**0.186**	**-2.08**	0.057
SPC	**0.590**	0.196	1.32	**0.042**	**0.627**	**0.182**	**-1.43**	**0.075**
LAS	**0.616**	**0.191**	**-2.55**	**0.060**	**0.634**	0.193	**-1.69**	0.058
RID	**0.591**	**0.189**	**-2.63**	**0.056**	0.607	0.188	-1.03	**0.059**
NET	**0.590**	**0.195**	69.77	0.036	0.604	0.191	0.69	0.058
GBS	0.580	**0.195**	7.38	0.034	**0.629**	**0.185**	5.88	**0.090**

Table 5-8: Median performance of eight model approaches (rows) obtained from 50 random splits of the AML data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 10-fold CV and either the CVPL (four columns on the left) or IBSC tuning criterion (four columns on the right). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

LAS, SPC and RID approaches show the best results. The shrinkage-based methods achieve the highest performance for models tuned by CVPL. The SPC, LAS and GBS approaches perform best for models validated by IBSC.

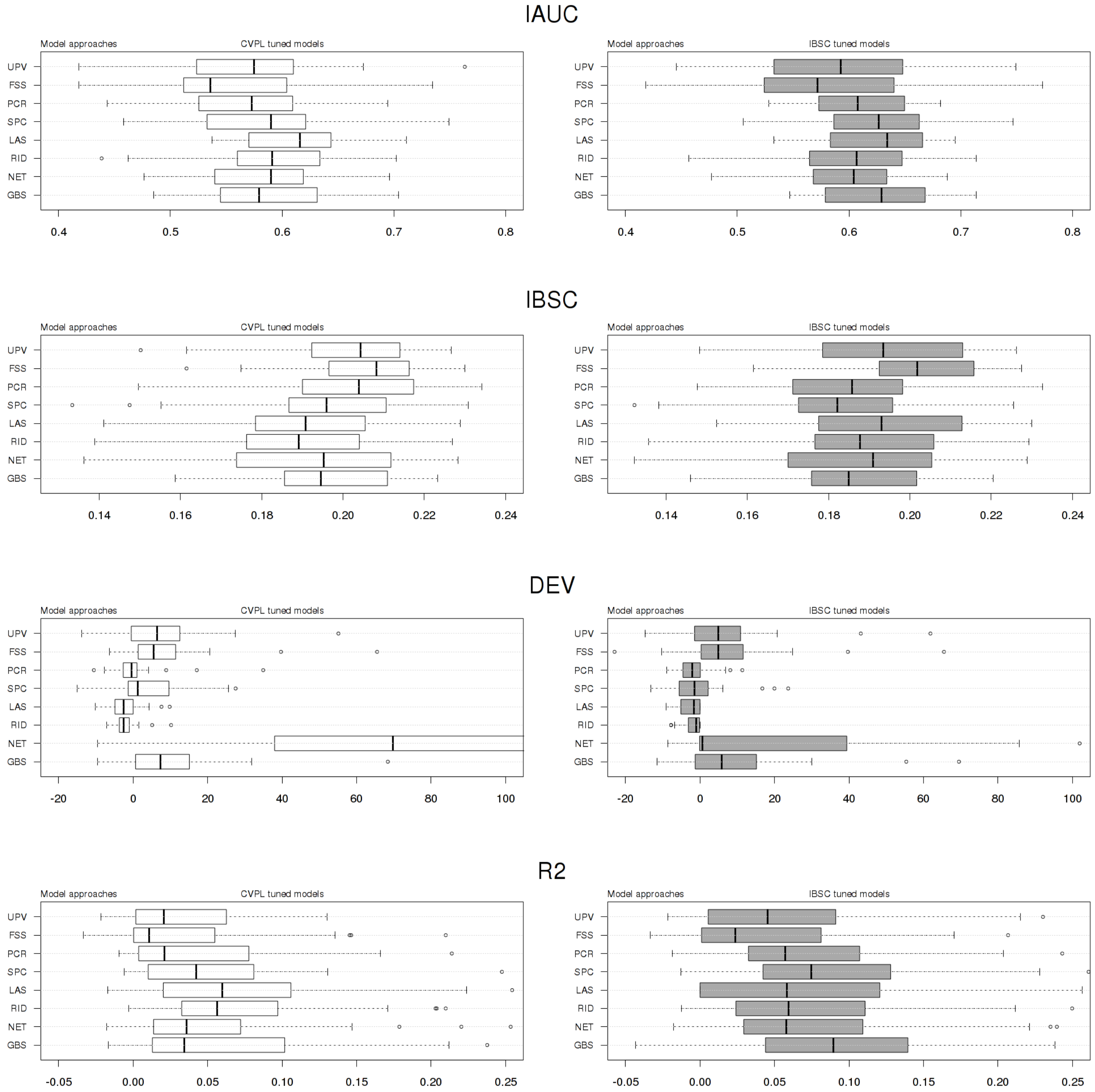

Figure 5-5: Performance of eight model approaches obtained from 50 random splits of the AML data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2. The survival models are validated by 10-fold CV and either the CVPL or IBSC tuning criterion. High prediction performance is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Results described in detail:

- Only minor differences appear between the results of the eight model approaches. The median IAUC values range from 0.54 (FSS) to 0.62 (LAS) for models validated by CVPL and between 0.57 (FSS) and 0.63 (LAS) for models tuned by IBSC. LAS, RID, NET, SPC and GBS perform marginally better than UPV, PCR and FSS. The variance of the IAUC values is almost equal (IQR around 0.1).

- The median IBSC performance values range from 0.21 (FSS) to 0.19 (RID) for models selected by the CVPL criterion and from 0.20 (FSS) to 0.18 (SPC) for models tuned by the IBSC criterion. Shrinkage methods, variance-based and ensemble approaches are marginally outperforming subset selection methods. The variances of the IBSC values are homogeneous and exhibit an IQR around 0.02. Only NET approaches achieve an IQR of 0.04.

- Only PCR, SPC, LAS and RID present negative median deviance values. FSS, UPV and GBS are marginally and NET far above 0. All model approaches show an IQR lower than 20. Only NET achieves an IQR around 40.

- Survival models achieve low median R2 values < 0.1. Only SPC, LAS and RID are above 0.04 for models validated by CVPL. SPC and GBS are above 0.06 for models tuned by IBSC. Shrinkage-, variance-based and ensemble methods marginally surpass subset selection techniques. The variance of the R2 values is homogeneous (IQR around 0.1).

Analysis of the primary objective for the preselected dataset

The second part of the primary objective is dedicated to the prediction performance of ten model techniques using the AML dataset with 500 preselected features. The results are displayed in table 5-9 and figure 5-6.

	CVPL				IBSC			
	IAUC	IBSC	DEV	R2	IAUC	IBSC	DEV	R2
UPV	0.575	0.204	7.13	0.020	0.600	0.193	6.32	0.063
FSS	0.574	0.203	3.32	0.035	0.652	0.187	1.51	0.111
PCR	**0.693**	**0.161**	**-8.94**	**0.188**	0.696	**0.156**	**-9.02**	**0.183**
SPC	0.651	0.182	-2.79	0.100	0.670	0.163	-5.93	0.140
PLS	**0.706**	**0.159**	**-9.48**	**0.190**	**0.697**	**0.161**	**-8.75**	0.179
LAS	0.685	0.169	-5.25	0.148	**0.697**	**0.161**	-5.63	**0.185**
RID	**0.698**	**0.159**	**-8.75**	**0.172**	**0.702**	**0.159**	**-7.33**	**0.180**
NET	0.649	0.170	49.97	0.118	0.682	**0.161**	-0.07	0.159
RSF	0.679	0.175	5.77	0.104	0.685	0.170	3.15	0.130
GBS	0.603	0.195	6.52	0.056	0.670	0.168	6.32	0.160

Table 5-9: Median performance of ten model approaches (rows) obtained from 50 random splits of the AML data including 500 features. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 10-fold CV and either the CVPL (four columns on the left) or IBSC tuning criterion (four columns on the right). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

The PLS, PCR, RID and LAS approaches reveal high prediction performance. SPC, RSF and NET show average results. The UPV, FSS and GBS strategies present poor performance. The NET approach achieves high variation of the deviance values. The performance values of the other approaches show homogeneous variances.

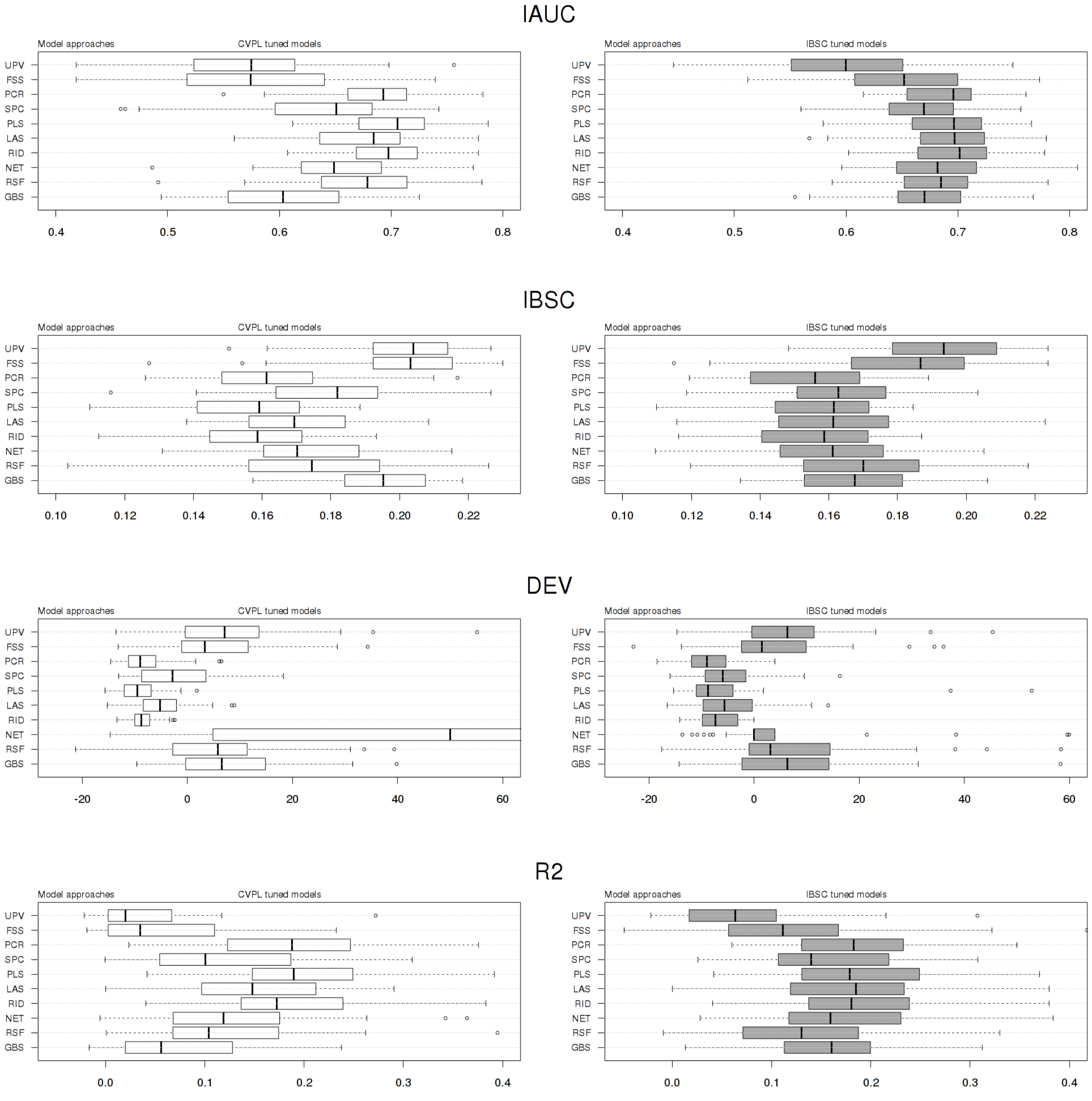

Figure 5-6: Performance of ten model approaches obtained from 50 random splits of the AML data including 500 features. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2. The survival models are validated by 10-fold CV and either the CVPL or IBSC tuning criterion. High prediction performance is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Results described in detail:

- High median IAUC values are presented performing PLS (0.71 for models tuned by CVPL and 0.70 for models validated by IBSC), PCR (0.69 and 0.70), RID (0.70 and 0.70), LAS (0.69 and 0.70), RSF (0.68 and 0.69), NET (0.65 and 0.68) and SPC (0.65 and 0.67). Lower values are achieved by GBS (0.60 and 0.67), UPV (0.58 and 0.60) and FSS (0.57 and 0.65). The IQR is homogeneous between 0.05 and 0.1.

- The median IBSC performance values range from 0.18 to 0.16 using RID, PLS, PCR, SPC, LAS, NET and RSF. They are minimally higher for UPV and FSS (0.20). The IQR of the performance values ranges from 0.02 to 0.04.

- The median deviance values are negative applying PCR, SPC, PLS, LAS and RID. Small positive values are achieved using UPV, FSS, RSF and GBS. A high median deviance result is detected for NET using models tuned by CVPL (49.97). The IQR of the deviance values ranges from 5 to 20. NET reveals a high IQR above 40.

- PLS, PCR and RID achieve high median R2 values for models tuned by CVPL (between 0.17 and 0.19) as well as PLS, PCR, RID and LAS for models validated by IBSC (between 0.18 and 0.19). The variance of the R2 values is homogeneous (IQR around 0.10).

Conclusion

The AML data consist of low signal genes. Variance and shrinkage-based methods including a high number of features perform best. Ensemble methods and subset selection techniques achieve a lower performance.

Using the preselected data the direction-derived (PCR, SPC) and shrinkage-based (RID, LAS and NET) approaches reveal an explicitly higher performance than applied to the full dataset. Subset selection techniques present a low performance. The role of the ensemble methods is not clearly defined. GBS shows a poor performance. RSF provides intermediate results.

The variance of the performance values is almost equal. Only the NET approach and the subset selection methods FSS and UPV show a negligibly higher variability.

The NET approach achieves high prediction accuracy for both datasets. Using the preselected data the PCR approach works similarly to the SPC technique. It shows a high performance and can even outperform the SPC approach. Survival models including a high number of features show the highest performance for the AML data. RID, NET, LAS and SPC approaches achieve accurate predictions using the full AML dataset and PLS, RID, PCR using the preselected dataset.

5.4.2.2 Analysis of the secondary objectives

The secondary objectives are dedicated to the examination of the tuning criteria and to resampling techniques. The performance of the validation strategies is assessed in terms of the IAUC, IBSC, DEV and R2 criteria.

Analysis of resampling techniques

The first analysis refers to the evaluation of 5-, 10- and 20-fold cross-validation. The results are presented in table 5-10 and figure 5-7. The median values of the performance criteria for the resampling techniques are shown in table 5-10. The best resampling technique for each model fitting technique and for each performance criterion is presented in bold numbers. Figure 5-7 presents the results for each model fitting technique and for each performance value in a boxplot matrix. The resampling techniques are analyzed using the full dataset. Only the values from the PLS and RSF approach are computed using the AML dataset with 500 preselected genes.

	IAUC			IBSC			DEV			R2		
	5-fold	**10-fold**	**20-fold**	**5-fold**	**10-fold**	**20-fold**	**5-fold**	**10-fold**	**20-fold**	**5-fold**	**10-fold**	**20-fold**
UPV	0.576	**0.583**	0.582	0.201	0.200	**0.196**	5.95	**5.20**	5.43	0.027	**0.032**	0.027
FSS	0.543	0.555	**0.585**	0.212	0.205	**0.201**	6.72	5.22	**5.20**	0.006	0.017	**0.033**
PCR	**0.597**	0.589	0.585	**0.192**	0.195	0.193	**-1.73**	-1.17	-0.85	**0.049**	0.037	0.039
SPC	0.608	**0.611**	0.596	0.191	0.190	**0.186**	**0.17**	0.46	2.43	**0.059**	**0.059**	0.053
PLS	0.688	**0.704**	0.673	0.159	0.161	**0.158**	-8.40	**-9.30**	-6.94	0.170	**0.181**	0.148
LAS	0.610	**0.630**	0.609	0.194	0.192	**0.189**	**-2.52**	-2.24	-2.00	0.054	0.060	**0.061**
RID	0.592	**0.600**	0.573	0.190	**0.188**	0.192	-1.59	**-2.05**	-1.11	0.055	**0.058**	0.042
NET	0.584	**0.597**	0.569	**0.192**	0.193	0.195	**35.72**	38.15	45.26	0.046	**0.049**	0.035
RSF	0.664	**0.681**	0.643	**0.171**	0.172	0.178	**3.75**	4.45	8.23	0.114	**0.119**	0.074
GBS	0.598	0.603	**0.608**	**0.190**	0.191	0.193	**4.96**	6.01	7.96	0.055	0.061	**0.068**

Table 5-10: Median performance of the resampling methods 5-, 10- and 20-fold cross-validation obtained from 50 random splits of the AML data. Survival predictions are made based on ten model approaches (rows). The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The resampling technique with the highest performance is shown in bold numbers and is characterized by high IAUC and R2 as well as low IBSC and DEV values.

The performance measurements don't allow conclusions about the best tuning setting. Only small differences of the prediction performances are detected between different validation samples. With respect to IAUC and the R2 values 10-fold cross-validation outperforms 5- and 20-fold cross-validation. In terms of the IBSC performance criterion 5- and 20-fold cross-validation surpass 10-fold cross-validation. Regarding deviance values 10- and 20-fold cross-validation show marginally poorer results that 5-fold cross-validation. The variance between the resampling techniques is almost equal.

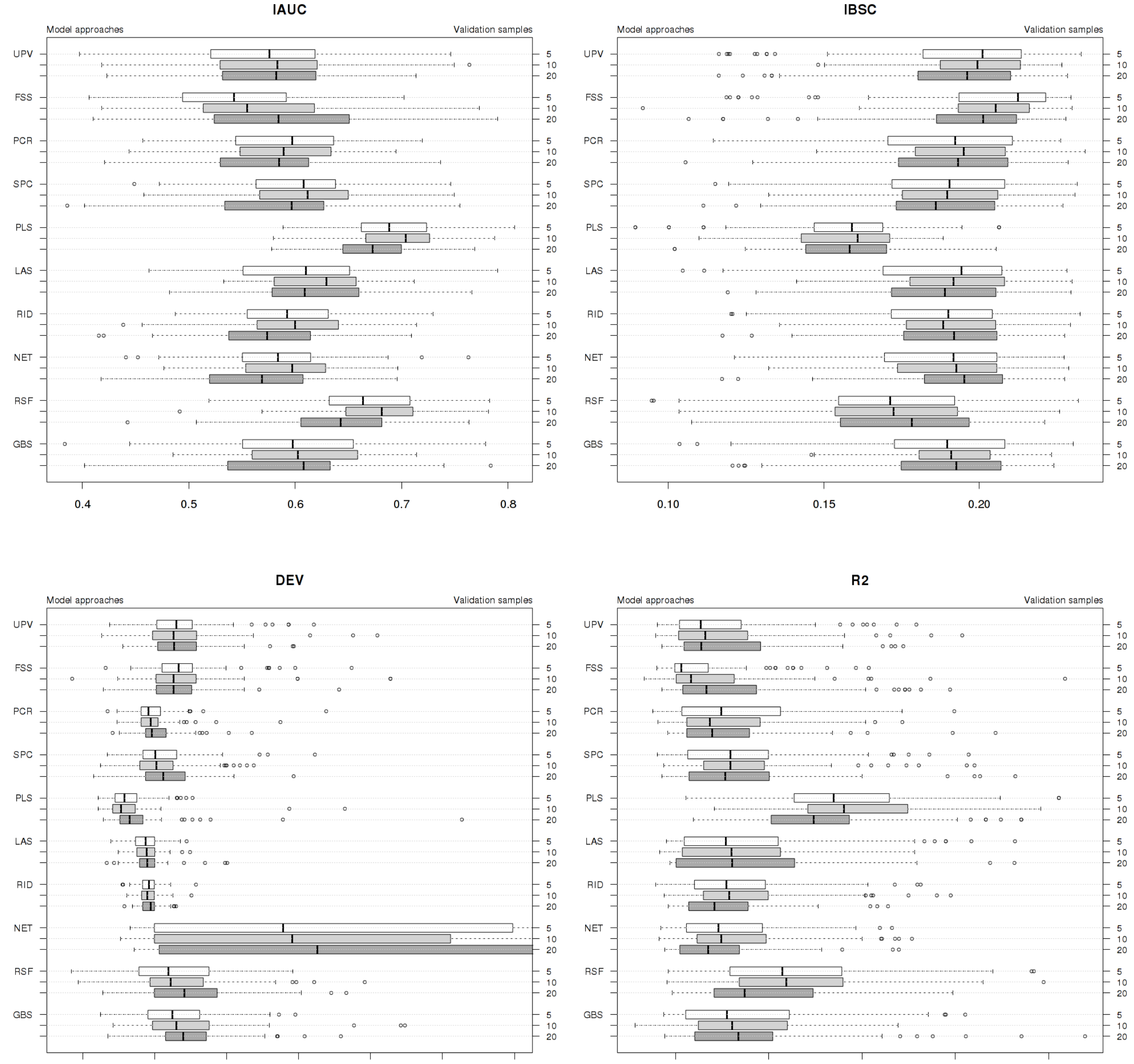

Figure 5-7: Performance of the resampling methods 5-, 10- and 20-fold cross-validation obtained from 50 random splits of the AML data. Survival predictions are made based on ten model approaches. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (boxes). High prediction performance is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Results described in detail:

- **With respect to median IAUC values:** 20-fold cross-validation performs best for 2 approaches, 10-fold cross-validation for 7 model fitting technique and 5-fold CV for 1 model approach.

- **Regarding median IBSC results:** 20-fold cross-validation achieves the highest prediction accuracy for 5 model approaches, 10-fold cross-validation for 1 model technique and 5-fold cross-validation for 4 approaches. The differences of median IAUC and the IBSC values between the two best resampling techniques are lower 0.02.

- **In terms of median DEV values:** 20-fold cross-validation shows the highest performance for 1 model approach, 10-fold cross-validation for 3 and 5-fold cross-validation for 6 approaches. The differences of median DEV values between the two best resampling techniques is lower than 2.5.

- **With regard to median R2 values:** 20-fold cross-validation presents the highest prediction accuracy for 3 approaches, 10-fold cross-validation for 6 and 5-fold cross-validation for 2 model fitting techniques. The differences of median R2 values between the two best resampling techniques are lower than 0.02.

Analysis of the tuning criterion

A secondary objective of this work is the evaluation of the tuning criteria IBSC and CVPL. Table 5-11 and figure 5-8 display the results of the secondary research question. Table 5-11 shows the median values of the performance metrics for the tuning criteria. The best validation criterion for each model approach and for each performance metric is presented in bold numbers. Figure 5-8 displays the results for each model fitting technique and for each performance value in a boxplot matrix. The resampling techniques are applied to the full dataset. Only the values of the PLS and RSF approach are computed using the dataset with 500 variables.

	IAUC		IBSC		DEV		R2	
	CVPL	**IBSC**	**CVPL**	**IBSC**	**CVPL**	**IBSC**	**CVPL**	**IBSC**
UPV	0.567	**0.593**	0.205	**0.193**	6.62	**4.77**	0.018	**0.043**
FSS	0.540	**0.571**	0.208	**0.203**	**5.65**	5.74	0.012	**0.022**
PCR	0.571	**0.607**	0.203	**0.182**	-0.25	**-2.24**	0.023	**0.065**
SPC	0.584	**0.617**	0.197	**0.182**	1.84	**-0.63**	0.039	**0.072**
PLS	**0.690**	0.681	**0.159**	0.161	**-8.69**	-7.57	**0.174**	0.166
LAS	0.605	**0.622**	0.191	**0.190**	-2.28	**-2.41**	0.058	**0.062**
RID	0.584	**0.595**	0.192	**0.189**	**-2.26**	-0.35	0.049	**0.054**
NET	0.575	**0.588**	0.194	**0.192**	68.65	**12.10**	0.035	**0.049**
RSF	0.660	**0.672**	0.177	**0.172**	6.86	**4.82**	0.094	**0.116**
GBS	0.586	**0.625**	0.195	**0.186**	**6.67**	7.01	0.035	**0.078**

Table 5-11: Median performance of the tuning criteria CVPL and IBSC obtained from 50 random splits of the AML data. Survival predictions are made based on ten model approaches (rows). The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The tuning criterion with the highest performance is shown in bold numbers and is characterized by high IAUC and R2 as well as low IBSC and DEV values.

The IBSC tuning criterion surpasses the CVPL criterion for almost all model approaches. Only small differences exist between the validation criteria. The variation of the results seems to be marginally higher for IBSC that for the CVPL criterion.

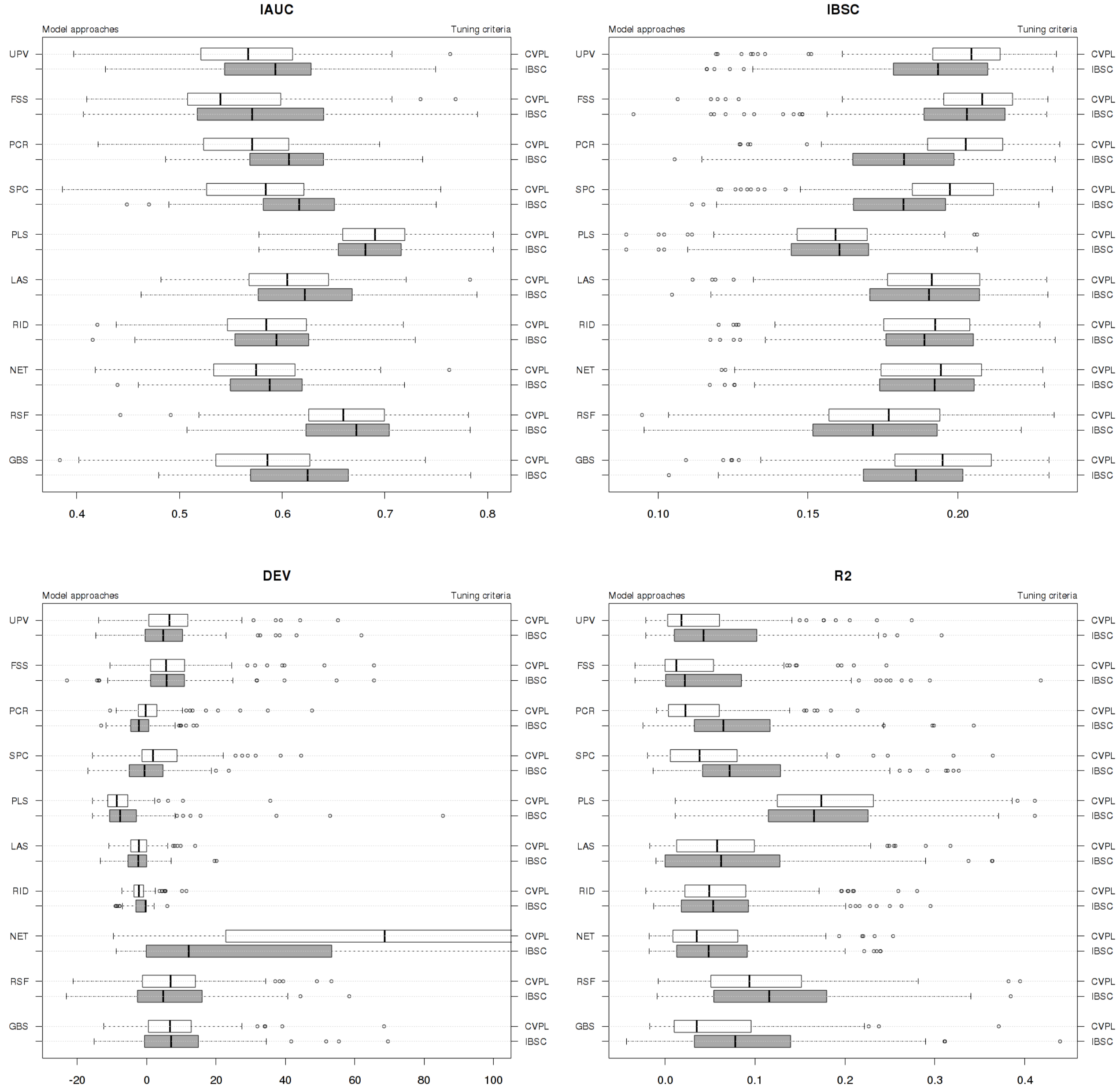

Figure 5-8: Performance of the tuning criteria CVPL and IBSC obtained from 50 random splits of the AML data. Survival predictions are made based on ten model approaches. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (boxes). High prediction performance is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Results described in detail:

- **With reference to median IAUC performance values:** The IBSC tuning criterion outperforms CVPL for 9 model approaches (differences range from 0.01 to 0.04). CVPL is better than the IBSC validation criterion only for PLS approaches (difference 0.01). The variation between the two tuning criteria is almost equal.

- **With respect to the median IBSC performance values:** The IBSC tuning criterion surpasses CVPL results for 9 of 10 model approaches (differences up to 0.02). Using PLS approaches the results are almost equal.

- **In terms of median deviance values:** The IBSC validation criterion achieves the highest performance for 6 approaches and CVPL for 4 model techniques. The differences are small (lower than 3 for almost all approaches). Only the NET approach achieves high differences (> 50). The IQR of the tuning criteria is almost equal except for the NET approach.

- **With reference to median R2 values:** The IBSC tuning criterion has a higher accuracy than CVPL for 9 (differences <= 0.05) and lower performance for one approach (differences < 0.01).

5.4.2.3 Analysis of the exploratory objective

The exploratory objective of this work examines features, which are selected by model fitting procedures.

GENE	UPV	FSS	PCR	SPC	PLS	LAS	RID	NET	GBS	TOTAL	NAPP
F726	**13**	**14**	0	**5**	0	**28**	**31**	**32**	0	**123**	6
F3759	**9**	3	0	4	0	**5**	**7**	0	0	**28**	5
F2773	**13**	4	0	2	0	3	3	0	0	**25**	5
F562	**7**	2	0	2	0	3	4	**5**	0	**23**	6
F2627	**5**	1	0	0	0	3	**9**	3	0	**21**	5
F3232	**11**	0	0	3	0	3	4	0	0	**21**	4
F2921	**10**	4	0	1	0	3	2	0	0	**20**	5
F351	**8**	4	0	3	0	2	0	0	0	**17**	4
F103	**7**	2	0	4	0	3	0	0	0	**16**	4
F286	0	0	0	0	**16**	0	0	0	0	**16**	1
F2724	1	0	0	0	0	**5**	4	**5**	0	**15**	4
F105	0	0	0	0	**12**	0	0	0	0	**12**	1
F972	0	0	0	0	0	0	0	0	**12**	**12**	1
F297	3	2	0	3	0	1	2	0	0	**11**	5
F2650	1	0	1	0	0	3	**5**	1	0	**11**	5
F466	0	0	0	0	0	0	0	0	**11**	**11**	1
F749	0	0	0	0	0	0	0	0	**11**	**11**	1
F685	1	1	0	0	0	1	0	**6**	0	**9**	4
F2643	3	3	0	0	0	1	1	0	0	**8**	4
F2667	4	2	0	0	0	1	1	0	0	**8**	4
F3223	1	0	0	0	0	1	3	3	0	**8**	4
F4675	2	1	0	1	0	2	2	0	0	**8**	5
F458	0	0	0	0	**8**	0	0	0	0	**8**	1
F178	2	0	0	3	0	0	2	0	0	**7**	3
F3017	3	3	0	0	0	0	0	1	0	**7**	3
F422	0	0	0	0	**7**	0	0	0	0	**7**	1
F4721	0	0	0	0	0	**7**	0	0	0	**7**	1
F4722	0	0	0	0	0	**7**	0	0	0	**7**	1
F400	0	0	0	0	0	0	0	0	**7**	**7**	1
F352	2	1	0	3	0	0	0	0	0	**6**	3
F1541	1	0	0	0	0	1	3	1	0	**6**	4
F2666	2	0	0	1	0	1	2	0	0	**6**	4
F3093	3	0	0	1	0	2	0	0	0	**6**	3
F157	0	0	0	0	0	0	0	0	**6**	**6**	1

Table 5-12: Genes included in 50 survival models using the AML data. The frequency matrix presents in how many cases the genes (rows) are included in the survival models. The analysis is based on nine model approaches (columns). The survival models are validated by 10-fold CV and the CVPL tuning criterion. The table is sorted in decreasing order of the frequency (TOTAL). Presented is the number of model techniques (NAPP) from which a gene was selected at least once. Frequencies above 4 are shown in bold numbers.

The survival models are fitted using 50 data samples of size 2/3* N . The models are validated by 10-fold cross-validation and the CVPL tuning criterion. The survival models are investigated using the full dataset. The PLS approach is applied to the dataset with 500 variables. Using FSS and UPV approaches all features in the prediction model are analyzed. Performing all other approaches only the two variables with the highest “relevance score” are used. This ensures a balanced number of features.

The genes are numbered from 1 to 4724. Table 5-12 presents the frequency of single genes selected by nine model approaches. The column “TOTAL” represents the total frequency over all model approaches (<= 450). The last column “NAPP” represents the number of approaches, which select a single gene at least one time (NAPP). Genes are only presented if they are included in the survival models more than 5 times.

Seven genes are included in the model at least 20 times (4.44 % of all survival models). One gene was selected 123 times (27.33 %; F726), one 28 (6.22 %; F3759), one 25 (5.56 %; F2773), one 23 (5.11 %; F562), two 21 (4.67 %; F2627, F3232) and one 20 times (4.44 %; F2921) in total. 17 genes are incorporated in the survival models at least 11 times (2.44 %).

The 7 most frequently used genes are selected from at least 4 model approaches (UPV, FSS, LAS and RID). The 17 most often used genes are chosen from 5, 4 or only 1 approach.

Results described in detail:

- **The 7 most often used genes** are included in models fitted by UPV, LAS and RID approaches. 6 of these genes are chosen from FSS and SPC, 3 from NET and 0 from PCR, PLS and GBS techniques.

- **Genes selected from each model approach at least 5 times:**
  - Models fitted by the UPV approaches incorporate 9 genes (5 to 13 times), FSS 1 gene (14 times), PCR 0 genes, SPC 1 gene (5 times), PLS 4 genes (7 to 16 times), LAS 5 genes (5 to 28 times), RID 4 genes (5 to 31 times), NET 4 genes (5 to 32 times) and GBS 5 genes (6 to 12 times).
  - Only one gene (F726) is selected by both subset selection approaches (UPV, FSS) and by shrinkage approaches (LAS, RID, NET). Two genes (F726, F3759) are chosen by UPV as well as LAS techniques. Three features (F726, F3759, F2627) are selected by both UPV and RID techniques. Two features (F726, F562) are chosen by UPV as well as NET approaches.

- Models fitted by the PLS approach incorporate 4 genes (F286, F105, F458, F422) and GBS includes 5 features that are not selected by other models. The results of the PCR approach seems to be arbitrary as PCR is an unsupervised technique.

- **The most often selected gene** is included in 27 % of the survival models (F726; 123/450 cases). All other genes are incorporated in less than 6.22 % (28/450) of the models.

5.4.3 Results for the DLBCL data

The DLBCL data (Diffuse large-B-cell lymphoma data; Rosenwald et al., 2002) include 240 patients. The primary objective is examined using either 5000 genes (full dataset) or 500 preselected features. Exploratory and secondary objectives are investigated using the full dataset. The PLS and RSF approaches are only applied to the DLBCL dataset with 500 variables.

5.4.3.1 Analysis of the primary objectives

The first part of the primary objective refers to the prediction performance of eight model approaches applied to the full DLBCL dataset including 5000 features. The performance of UPV, FSS, PCR, SPC, LAS, RID, NET and GBS approaches is assessed with the integrated area under the ROC curve (IAUC), the integrated Brier score (IBSC), the deviance (DEV) and the explained variation (R2) based on the Brier score. The prediction models are validated by either the cross-validated log partial likelihood (CVPL) or the integrated Brier score (IBSC) and by 10-fold cross-validation.

The results are presented in table 5-13 and figure 5-9. Median values of the performance metrics for the model approaches are presented in table 5-13. The best three approaches for the performance measurements are shown in bold numbers. The four right columns represent the results for the IBSC tuning criterion and the four left columns the values for the CVPL tuning criterion. Figure 5-9 displays the results by a boxplot matrix. The four right boxes are dedicated to models tuned by IBSC and the four left boxes to models validated by CVPL.

	CVPL				IBSC			
	IAUC	IBSC	DEV	R2	IAUC	IBSC	DEV	R2
UPV	0.555	0.216	8.37	0.007	0.560	**0.211**	5.15	0.012
FSS	0.563	**0.211**	6.73	0.011	0.561	**0.210**	3.86	**0.016**
PCR	0.564	0.218	**0.00**	0.016	0.540	0.213	**-0.39**	0.009
SPC	0.542	0.222	5.12	0.003	0.555	**0.210**	**-0.62**	0.015
LAS	**0.579**	**0.213**	**-1.68**	**0.023**	**0.586**	0.216	**-0.14**	0.006
RID	**0.581**	0.214	**-2.86**	**0.028**	**0.576**	0.216	-0.11	**0.021**
NET	**0.578**	0.217	42.36	**0.021**	0.562	0.221	76.14	0.012
GBS	**0.578**	**0.213**	6.59	0.018	**0.580**	**0.211**	6.67	**0.020**

Table 5-13: Median performance of eight model approaches (rows) obtained from 50 random splits of the DLBCL data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 10-fold CV and either the CVPL (four columns on the left) or IBSC tuning criterion (four columns on the right). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Only small differences appear between the model approaches. Shrinkage-based approaches perform well with respect to IAUC, subset selection techniques related to IBSC and variance-based methods with regard to deviance results.

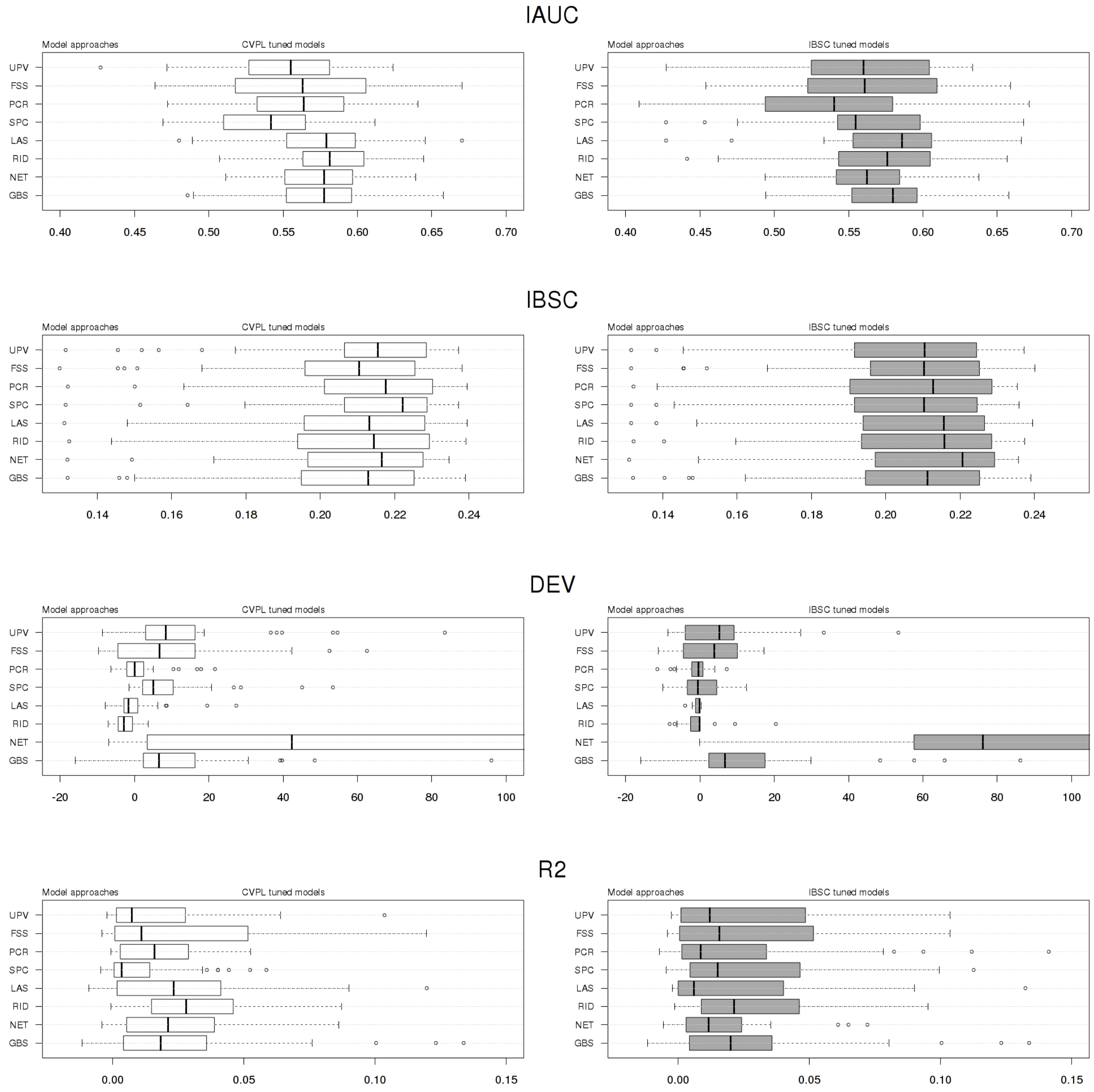

Figure 5-9: Performance of eight model approaches obtained from 50 random splits of the DLBCL data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2. The survival models are validated by 10-fold CV and either the CVPL or IBSC tuning criterion. High prediction performance is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Results described in detail:

- **With regard to median values of the IAUC criterion:** The shrinkage-based approaches (LAS, RID and NET range from 0.56 to 0.59) and GBS (0.58) reveal the highest performance. UPV, FSS (0.56) as well as PCR and SPC (range from 0.54 to 0.56) achieve the same performance level. The variability of the IAUC results is almost equal (IQR ranges from 0.05 to 0.1).

- **With respect to median IBSC values:** The FSS and GBS (0.21) approaches perform best. UPV, PCR, SPC, LAS, RID and NET (between 0.21 and 0.22) achieve the same level of performance.

- Only PCR, SPC, LAS and RID present negative **median deviance values**. FSS, UPV and GBS are marginally and NET far above 0. The IQRs range from 2 to 20. Only the IQR of the NET approach was much higher.

- The survival models achieve low **median R2 values** < 0.03. LAS, RID and NET values are between 0.02 and 0.03 for models tuned by CVPL. RID and GBS show R2 values of 0.02 for models validated by the IBSC criterion. The IQR of the R2 values (0.02 to 0.05) is homogeneous.

For survival models based on CVPL the shrinkage methods marginally outperform the other approaches and for models validated by IBSC the subset and variance-based approaches achieve the highest performance.

Analysis of the primary objective for the preselected DLBCL data

The second part of the primary analysis is dedicated to the prediction performance of ten model techniques using the DLBCL dataset with 500 preselected features. The results are displayed in table 5-14 and Figure 5-10.

	CVPL				IBSC			
	IAUC	**IBSC**	**DEV**	**R2**	**IAUC**	**IBSC**	**DEV**	**R2**
UPV	0.557	0.216	8.37	0.008	0.561	0.211	5.15	0.017
FSS	0.563	0.211	6.36	0.013	0.563	0.209	3.86	0.028
PCR	**0.726**	0.172	**-25.76**	0.219	0.717	0.166	**-23.57**	0.212
SPC	0.651	0.195	-6.85	0.099	0.646	0.186	-6.92	0.100
PLS	0.724	**0.168**	-24.62	**0.225**	**0.738**	**0.151**	**-22.21**	**0.261**
LAS	0.617	0.205	-2.53	0.056	0.606	0.208	-0.26	0.013
RID	**0.745**	**0.154**	**-26.53**	**0.260**	**0.742**	**0.154**	**-24.31**	**0.255**
NET	**0.750**	**0.151**	**-25.63**	**0.270**	**0.726**	**0.161**	-8.05	**0.226**
RSF	0.632	0.205	27.04	0.046	0.657	0.205	34.35	0.067
GBS	0.585	0.211	6.59	0.024	0.589	0.208	8.98	0.027

Table 5-14: Median performance of ten model approaches (rows) obtained from 50 random splits of the DLBCL data including 500 features. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 10-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

RID, NET, PLS and PCR show high-prediction performances. SPC, LAS and RSF achieve average and GBS, UPV and FSS poor performances. The performance values of the model approaches show homogeneous variances.

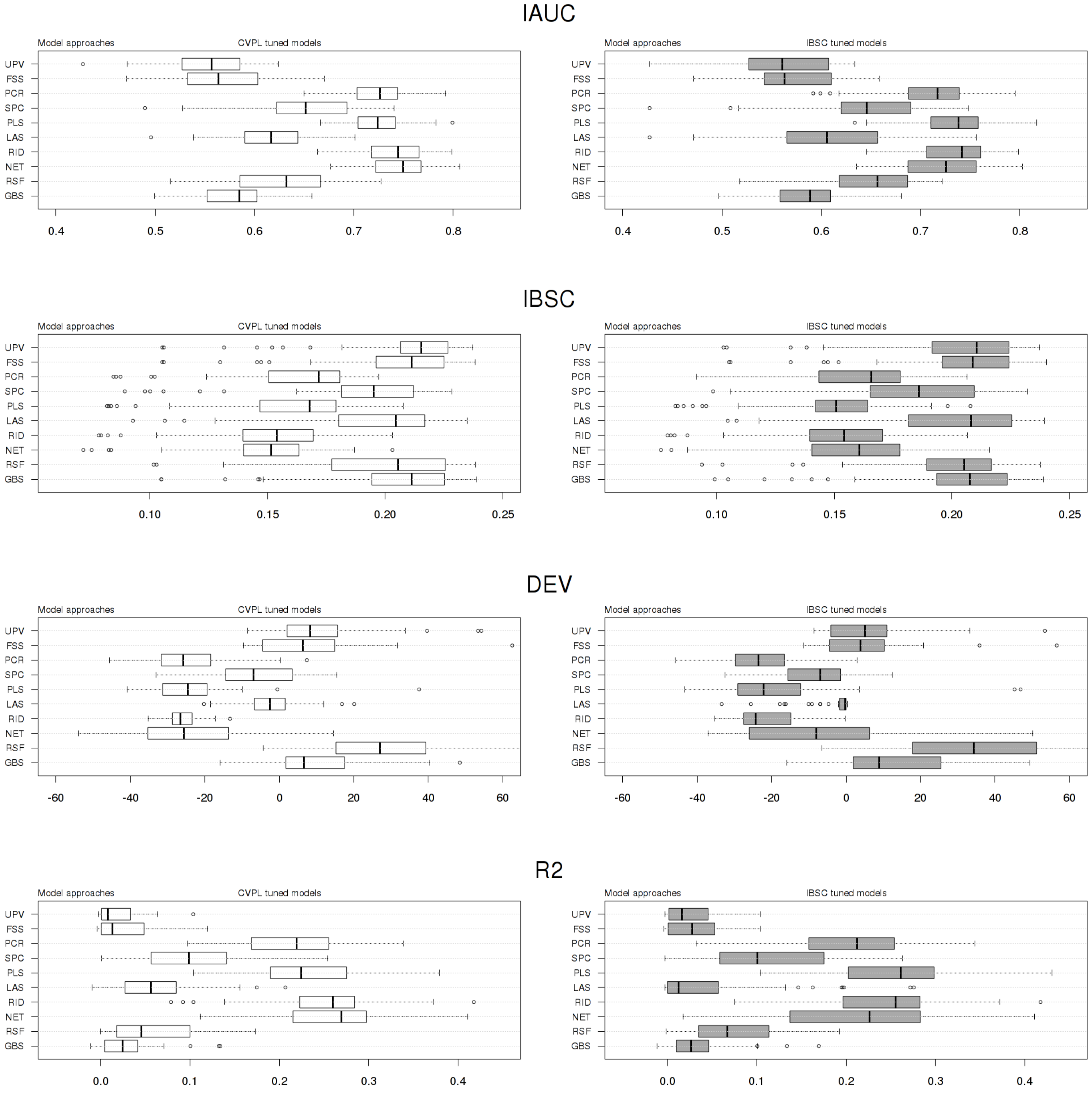

Figure 5-10: Performance of eight model approaches obtained from 50 random splits of the DLBCL data including 500 features. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2. The survival models are validated by 10-fold CV and either the CVPL or IBSC tuning criterion. High prediction performance is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Results described in detail:

- High median IAUC values are revealed using NET (0.75/0.73 for models tuned by CVPL/IBSC), RID (0.75/0.74), PCR (0.73/0.72) and PLS (0.72/0.74). Average performance is achieved applying SPC (0.65), RSF (0.63/0.66) and LAS (0.62/0.61). GBS (0.59), UPV (0.56) and FSS (0.56) show poor performances. The IQR is homogeneous and below 0.10.

- The median IBSC performance values are below 0.20 performing RID (0.15 for models tuned by CVPL/IBSC), NET (0.15/0.16), PLS (0.17/0.15) and PCR (0.17) and around 0.2 using SPC (0.20/0.19), RSF, GBS, FSS, LAS (0.21) and UPV (0.22/0.21). The IQR ranges from 0.02 to 0.05.

- The median deviance values are around -25 applying PCR, PLS and RID approaches and marginally below 0 using LAS. The median deviance results are minimally above 0 performing UPV, FSS and GBS and above 27 applying RSF. The IQR of the deviance values ranges from 5 to 20 for almost all approaches. The IQR is above 20 when using GBS, RSF and the NET approaches.

- The median values of the explained variation range from 0.20 to 0.31 performing PCR, PLS, RID and NET. All other approaches range below 0.11. The IQR of the R2 values is lower than 0.2 for all model approaches.

Conclusion

The variance-based and penalized techniques achieve the highest performance. RID is the best model approach. The PLS method only evaluated on the preselected dataset shows accurate predictions. The NET technique shows a similar level of performance.

The PCR approach achieves average results using the full and high-performance using the preselected dataset. As already demonstrated for the AML data, the PCR surpasses SPC results using the preselected data.

Feature extraction approaches exhibit a considerably higher performance using the preselected data than for the full dataset. The UPV, FSS and GBS values achieve the same performance level.

The GBS technique shows a high performance using the full and poor performance for the preselected data. The NET approach reveals a much higher performance using the preselected data than for the full dataset.

The variance of the performance values is homogeneous. Only the NET approach achieves a minimally higher variance regarding the deviance criterion.

Using the full dataset the performance of the approaches is almost equal. Shrinkage approaches and GBS minimally outperform the other methods. For the preselected data the extraction methods surpass model selection approaches. The shrinkage and derived-direction techniques show high performance and outperform both ensemble methods and subset selection techniques.

5.4.3.2 Analysis of the secondary objectives

The secondary research questions, the performance of tuning criteria and of resampling techniques, are examined in this section. The accuracy of the tuning strategies is evaluated in terms of IAUC, IBSC, DEV and R2 criteria.

Analysis of resampling techniques

The first research question refers to the evaluation of 5-, 10- and 20-fold cross-validation. The results are shown in table 5-15 and figure 5-11.The median values of the performance metrics for the resampling techniques are presented in table 5-15. The best resampling technique for each model approach and for each performance metric is presented in bold numbers. Figure 5-11 presents the results for each model technique and for each performance metric in a boxplot matrix. The resampling techniques are analyzed using the full dataset. Only the values from the PLS and RSF approach are computed for the DLBCL dataset with 500 preselected genes.

	IAUC			**IBSC**			**DEV**			**R2**		
	5-fold	**10-fold**	**20-fold**	**5-fold**	**10-fold**	**20-fold**	**5-fold**	**10-fold**	**20-fold**	**5-fold**	**10-fold**	**20-fold**
UPV	0.548	0.556	**0.568**	0.219	0.212	**0.211**	7.80	6.11	**5.34**	0.012	0.009	**0.016**
FSS	0.555	0.561	**0.587**	0.219	0.210	**0.208**	3.36	5.31	**-0.97**	0.010	0.012	**0.033**
PCR	**0.555**	0.550	0.552	0.222	**0.217**	**0.217**	-0.14	**-0.28**	-0.27	0.010	0.011	**0.012**
SPC	0.548	0.550	**0.561**	0.220	0.215	**0.211**	3.93	2.68	**2.12**	0.011	0.008	**0.019**
PLS	0.729	0.729	**0.738**	0.160	0.157	**0.153**	-20.03	-23.06	**-25.72**	0.226	0.245	**0.258**
LAS	**0.585**	0.579	**0.585**	0.221	0.214	**0.213**	-0.33	-0.29	**-1.07**	0.012	0.016	**0.022**
RID	0.585	0.579	**0.595**	0.218	0.215	**0.212**	-1.98	-1.43	**-3.00**	0.024	0.025	**0.037**
NET	0.568	0.568	**0.571**	0.218	0.218	**0.216**	71.81	68.89	**50.90**	0.014	0.017	**0.019**
RSF	0.638	**0.646**	0.641	0.207	0.205	**0.204**	28.61	31.55	**28.57**	0.055	**0.057**	**0.057**
GBS	0.573	0.578	**0.583**	0.219	**0.212**	0.213	8.60	**6.61**	7.23	0.025	0.019	**0.027**

Table 5-15: Median performance of the resampling methods 5-, 10- and 20-fold cross-validation obtained from 50 random splits of the DLBCL data. Survival predictions are made based on ten model approaches (rows). The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The resampling technique with the highest performance is shown in bold numbers and is characterized by high IAUC and R2 as well as low IBSC and DEV values.

The performance of 20-fold cross-validation surpasses the prediction accuracy of 5- and 10-fold cross-validation. The differences of the performance values are very small. The variance between the resampling techniques is almost equal.

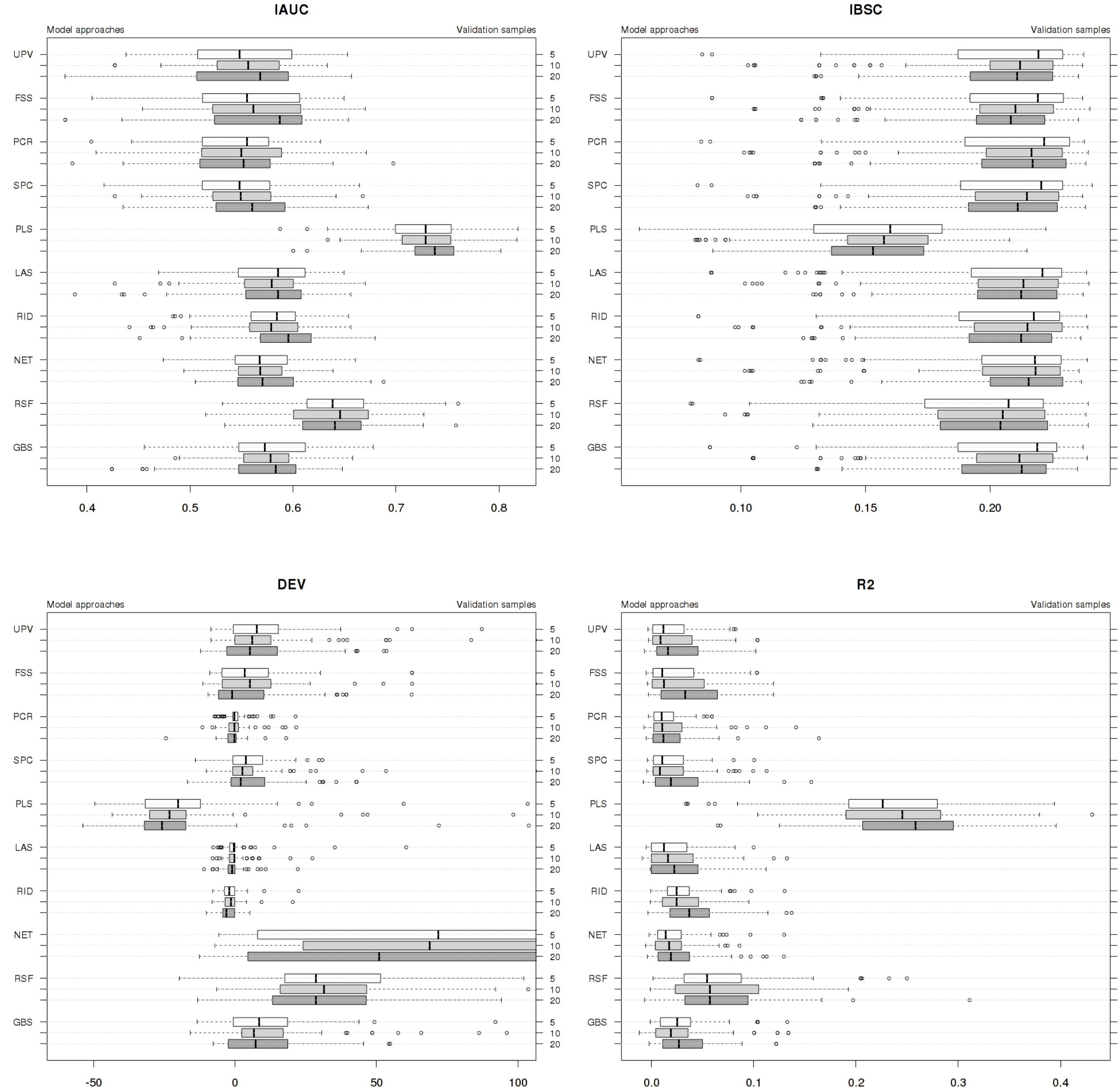

Figure 5-11: Performance of the resampling methods 5-, 10- and 20-fold cross-validation obtained from 50 random splits of the DLBCL data. Survival predictions are made based on ten model approaches. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (boxes). High prediction performance is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Results described in detail:

- **With respect to median IAUC values:** 20-fold cross-validation performs best for 8 approaches, 10-fold cross-validation for 1 model fitting technique and 5-fold CV for 2 model approaches.

- **Regarding median IBSC results:** 20-fold cross-validation achieves the highest prediction accuracy for 9 model approaches, 10-fold cross-validation for 2 model techniques and 5-fold cross-validation for 0 approaches. The differences of the median IAUC and the IBSC values between the two best resampling techniques are lower 0.01.

- **In terms of median DEV values:** 20-fold cross-validation shows the highest performance for 8 model approaches and 10-fold cross-validation for 2 approaches. The differences of the median DEV values between the two best resampling techniques is lower than 3.

- **With regard to median R2 values:** 20-fold cross-validation presents the highest prediction accuracy for 10 approaches, 10-fold cross-validation for 1 model fitting technique. The differences of the median R2 values between the two best resampling techniques are lower than 0.02.

Analysis of the tuning criterion

The evaluation of the tuning criteria IBSC and CVPL is a secondary objective of this work. Figure 5-12 and table 5-16 present the results of the secondary research question. Table 5-16 displays the median values of the performance measurements for the tuning criteria. The best validation metric for each model approach and for each performance metric is displayed in bold numbers. Figure 5-12 presents the results for each model technique and for each performance metric in a boxplot matrix. The resampling techniques are applied to the full dataset. Only the values from the PLS and RSF approach are computed using the dataset with 500 variables.

	IAUC		**IBSC**		**DEV**		**R2**	
	CVPL	**IBSC**	**CVPL**	**IBSC**	**CVPL**	**IBSC**	**CVPL**	**IBSC**
UPV	0.553	**0.561**	0.218	**0.213**	8.19	**4.27**	0.012	**0.017**
FSS	**0.571**	**0.571**	**0.211**	**0.211**	4.07	**2.22**	0.020	**0.021**
PCR	**0.562**	0.539	0.220	**0.217**	-0.02	**-0.35**	**0.014**	0.009
SPC	0.544	**0.556**	0.221	**0.211**	6.11	**-0.24**	0.006	**0.021**
PLS	0.724	**0.740**	0.166	**0.150**	**-24.34**	-22.55	0.222	**0.261**
LAS	0.579	**0.587**	0.217	**0.214**	**-1.09**	-0.29	**0.020**	0.014
RID	**0.589**	0.583	**0.215**	**0.215**	**-2.86**	-0.12	**0.029**	0.026
NET	**0.578**	0.562	**0.216**	0.220	**33.46**	79.71	**0.022**	0.013
RSF	0.635	**0.649**	0.211	**0.201**	**28.03**	31.52	0.047	**0.067**
GBS	**0.578**	**0.578**	**0.214**	**0.214**	**6.79**	7.19	0.024	**0.025**

Table 5-16: Median performance of the tuning criteria CVPL and IBSC obtained from 50 random splits of the DLBCL data. Survival predictions are made based on ten model approaches (rows). The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The tuning criterion with the highest performance is shown in bold numbers and is characterized by high IAUC and R2 as well as low IBSC and DEV values.

The IBSC criterion marginally surpasses the CVPL results except for the RID and NET approach. The variance between the tuning criteria is almost equal.

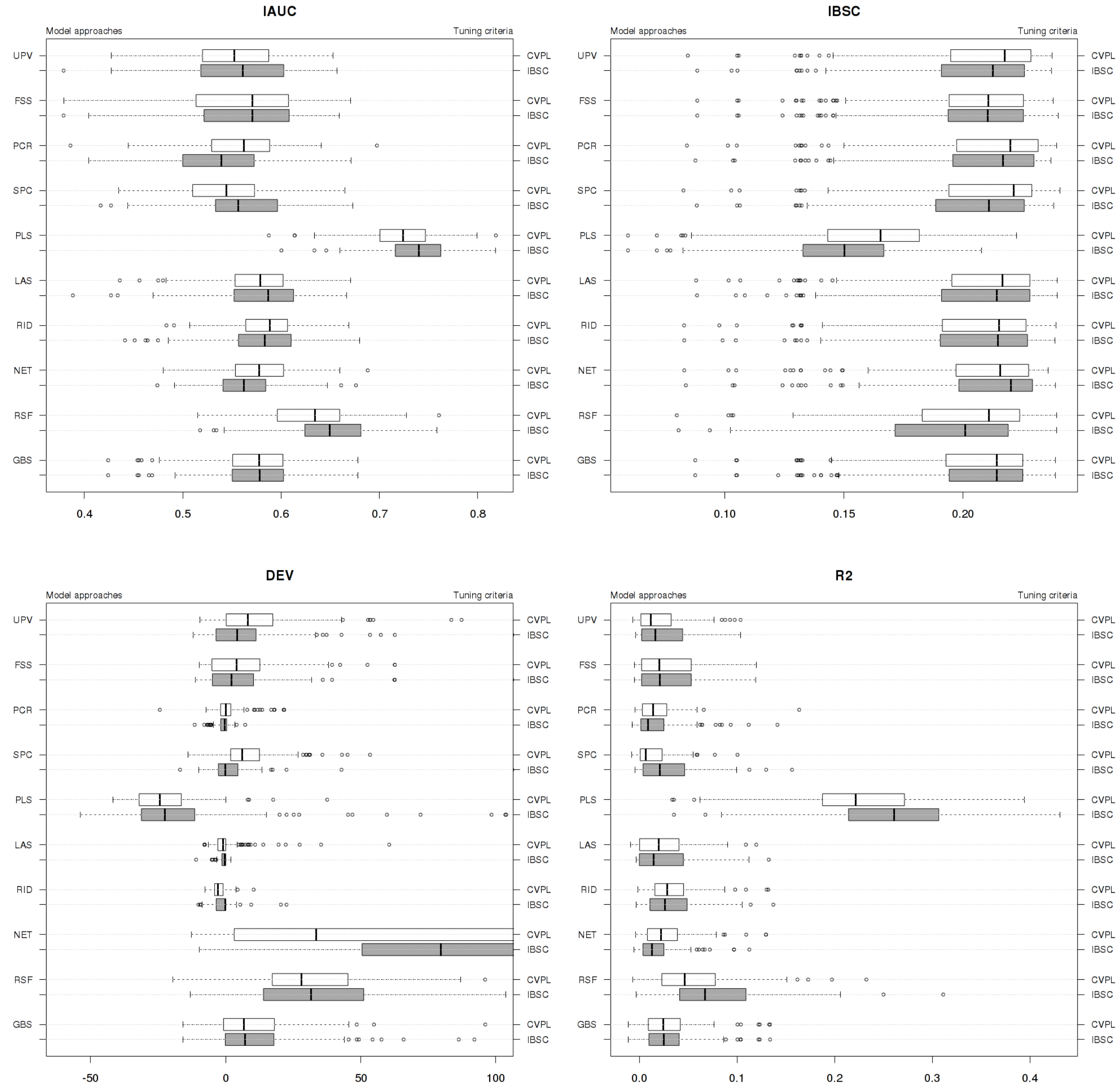

Figure 5-12: Performance of the tuning criteria CVPL and IBSC obtained from 50 random splits of the DLBCL data. Survival predictions are made based on ten model approaches. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (boxes). High prediction performance is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Results described in detail:

- **With reference to median IAUC performance values:** The IBSC tuning criterion outperforms CVPL for 5 model approaches (differences range from 0.01 to 0.02). CVPL surpasses the IBSC validation criterion only for 3 model approaches (differences range from 0.01 to 0.02).

- **With respect to the median IBSC performance values:** The IBSC tuning criterion surpasses CVPL results for 6 model approaches (differences < 0.02). CVPL outperforms IBSC results only for NET approaches (difference < 0.01).

- **In terms of median deviance values:** The IBSC validation criterion achieves the highest performance for 4 approaches and CVPL for 6 model techniques. The differences are small (lower than 7 for almost all models). Only the NET approach achieves high differences (46).

- **With reference to median R2 values:** The IBSC tuning criterion has a higher accuracy than CVPL for 6 (differences < 0.04) and a lower performance for 4 approaches (differences < 0.01).

5.4.3.3 Analysis of the exploratory objective

The exploratory objective of this work examines features, which are selected by model fitting procedures.

GENE	UPV	FSS	PCR	SPC	PLS	LAS	RID	NET	GBS	TOTAL	NAPP
F1245	**35**	**27**	0	**14**	0	**31**	**26**	**8**	0	**141**	6
F2241	**9**	1	0	3	0	**7**	**15**	**5**	0	**40**	6
F124	0	0	0	0	0	0	0	0	**28**	**28**	1
F243	0	0	0	0	**26**	0	0	0	0	**26**	1
F4450	4	**6**	0	0	0	4	4	2	0	**20**	5
F1597	**5**	0	0	1	0	4	**7**	2	0	**19**	5
F4	1	3	0	1	0	2	0	1	**10**	**18**	6
F335	1	1	0	0	0	0	2	**10**	3	**17**	5
F944	3	1	0	1	0	2	2	2	**5**	**16**	7
F1214	**5**	0	1	1	0	3	2	1	0	**13**	6
F2373	**7**	3	0	1	0	2	0	0	0	**13**	4
F2469	4	3	0	2	0	2	0	2	0	**13**	5
F389	1	0	0	0	0	1	1	**6**	3	**12**	5
F2031	1	0	0	0	0	2	3	**6**	0	**12**	4
F388	2	0	0	2	0	2	1	2	2	**11**	6
F1259	3	2	0	2	0	2	2	0	0	**11**	5
F2163	2	1	0	0	0	2	4	1	0	**10**	5
F4722	1	1	0	0	0	2	4	1	0	**9**	5
F3381	**6**	0	0	2	0	0	0	0	0	**8**	2
F121	0	0	1	0	0	0	0	0	**7**	**8**	2
F516	1	1	0	1	0	1	1	1	1	**7**	7
F528	1	1	0	0	0	1	0	3	1	**7**	5
F665	3	0	0	1	0	0	0	0	3	**7**	3
F952	4	0	0	1	0	0	0	0	2	**7**	3
F1951	3	1	0	1	0	2	0	0	0	**7**	4
F4307	0	0	0	1	0	0	3	3	0	**7**	3
F158	0	0	0	0	**7**	0	0	0	0	**7**	1
F274	0	0	0	0	**7**	0	0	0	0	**7**	1
F937	1	0	0	1	0	1	1	1	1	**6**	6
F1555	4	1	0	0	0	1	0	0	0	**6**	3
F3948	**5**	0	0	1	0	0	0	0	0	**6**	2
F1698	0	1	0	0	0	0	2	3	0	**6**	3
F771	0	0	0	1	0	1	1	2	1	**6**	5

Table 5-17: Genes included in 50 survival models using the DLBCL data. The frequency matrix presents in how many cases the genes (rows) are included in the survival models. The analysis is based on nine model approaches (columns). The survival models are validated by 10-fold CV and the CVPL tuning criterion. The table is sorted in decreasing order of frequency (TOTAL). Presented is the number of model techniques (NAPP) from which a gene was selected at least once. Frequencies above 4 are shown in bold numbers.

The survival models are fitted using 50 data samples of size 2/3* N . The models are tuned by 10-fold cross-validation and the CVPL tuning criterion. The survival models are examined using the full DLBCL dataset. The PLS approach is applied to the dataset with 500 variables. For FSS and UPV approaches all features in the survival model are analyzed. For all other approaches only the two variables with the highest "relevance score" are used. This ensures a balanced number of features.

The features are numbered from 1 to 5000. Table 5-17 shows the frequency of single genes selected by nine model approaches. The column "TOTAL" represents the total frequency over all model approaches (<= 450). The last column "NAPP" represents the number of approaches, which select a single gene at least one time (NAPP). Genes are only presented if they are incorporated in more than 5 survival models.

Seventeen genes are included in the survival models at least 10 times (2.22 % of all survival models) and five features are selected at least 20 times (4.44 %). One feature was included 141 times (31.33 %; F1245), one 40 (8.89 %; F2241), one 28 (6.22 %; F124), one 26 (5.78 %; F243) and one 20 times (4.44 %; F4450).

Similar to the AML results, the 17 most frequently selected genes are included in 4 to 7 approaches (NAPP) or are selected by only one approach (PLS and GBS). The following interpretations only take cell frequencies above 5 into account.

Results described in detail:

- Survival models fitted by the UPV approach incorporate 7 genes (5 to 35 times), by NET 5 features (5 to 10 times), by GBS 4 genes (5 to 28 times), by PLS and RID 3 genes (7 to 26 times). Survival models selected by the FSS and LAS approach include 2 genes (6 to 27 times and 7 to 31 times respectively), by SPC 1 gene (14 times) and by PCR 0 genes.

- UPV, FSS, LAS, RID and NET included 2 of the 5 most often selected genes; SPC, PLS and GBS only 1 of these 5 genes.

- Only one gene (F1245) is incorporated by both subset selection approaches (UPV, FSS). Two features (F1245, F2241) are included by all shrinkage-based techniques (LAS, RID, NET).

- Prediction models selected by PLS and GBS approaches only include a few genes. The PLS approach mainly incorporated three genes (F243, F158, F274) that are never selected by other models. These results indicate that other model fitting techniques are less accurate as they do not include the "PLS genes". The PLS approach achieves a high prediction performance. GBS and NET include 3 features (F124, F4, F121 and F335, F389, F2031 respectively) that are never or seldom used by other model techniques.

- **The two most often selected genes** are included in 31 % (F1245) and in 9 % of the prediction models (F2241). All other genes are included in less than 6.22 % of the models.

5.5 Discussion

The prediction performance of model approaches for high-dimensional data is strongly associated with the presence of high-signal genes. The generated dataset consists of a low number of highly significant and many random variables. Survival models built from feature selection approaches include a high number of correct variables and a low number of noise variables. The prediction models achieve a high performance and low variance. The use of subset selection techniques is preferable to feature extraction models.

Using the two microarray datasets (AML and DLBCL dataset) only one gene is frequently selected by the model fitting techniques. The performance of the survival models is low and differences between the model approaches are small. Since microarray data consist of a large set of genes that is weakly associated with survival, the feature extraction methods that include a high number of variables (PLS, SPC, PCR, RID and NET), seem to marginally surpass the other model approaches.

The feature extraction approaches achieve the highest performance for the microarray data (AML and DLBCL dataset). These are PLS, RID, NET and PCR for the preselected data. PLS was only assessed for datasets with 500 variables, but provides the highest prediction performance of all model approaches. SPC and PCR achieve accurate predictions for the preselected microarray data. Using the full dataset the PCR reveals poor performance. The results are partly consistent with the literature as Bovelstad et al. (2007) and van Wieringen et al. (2009) demonstrated high prediction performance of PLS and the ridge regression. The lasso and the supervised principal components regression on the other hand present high performance particularly for the full microarray data.

Feature selection techniques are marginally outreached for the microarray datasets. Only LAS and GBS achieve high to average prediction performances. The subset selection techniques are minimally outperformed by extraction methods. This coincides with results reported in other publications (Bovelstad et al., 2007, and van Wieringen et al., 2009).

This work comprises GBS, NET and RSF approaches that were seldom compared with other model approaches. The GBS and NET show high and the RSF poor performance. A high variance is detected for the RSF results.

The performance of the prediction models is higher for the preselected than for the full dataset. In this work 500 genes are preselected that are univariately associated with survival. Using preselected datasets the feature extraction approaches perform much better than feature selection methods. PCR, RID and NET have an explicitly higher performance for the preselected data. The model performances of UPV, FSS, LAS and GBS approaches are similar for the preselected and the full dataset.

Using the generated data with 1000 variables the prediction performance of the model approaches is quite different, but for the dataset with 500 variables the prediction accuracy is almost equal. Using the full microarray datasets small differences of the prediction performance appear between the model techniques. For preselected microarray data large differences are apparent.

To sum up in a precise and succinct way: For high signal data the SPC, FSS and UPV perform best and feature selection methods, particularly the subset selection approaches obtain high prediction performance. For low signal data PLS, SPC, RID and LAS achieve accurate predictions. The variance-based and shrinkage techniques reveal the highest performance. For ensemble approaches only GBS can achieve the same performance level as the other approaches. RSF shows poor results and unstable models. Subset selection techniques have high performance using high-signal data and are only marginally surpassed for the microarray data. The variance of the results is homogeneous except for RSF and NET approaches.

The prediction performance of validation strategies for survival models from gene expression data offers an almost uniform picture. For the generated high-signal data (sample size of the training data N = 100 observations) 20-fold cross-validation performs marginally better than 5- and 10-fold cross-validation. Using the real data the following results are achieved: 5- and 10-fold cross-validation marginally exceed 20-fold cross-validation using the AML data (sample size of training data N = 77 patients). 20-fold cross-validation performs marginally better than 5- and 10-fold cross-validation for the DLBCL dataset (sample size of training data N = 160 patients).

These results are consistent with the findings of Subramanian and Simon (2011), although the research questions and the setup of the comparison study are different. A high number of cross-validation samples (> 10-fold: e.g. LOOCV) for high-signal data achieve a high prediction performance. Using low-signal data and small samples (N ~80) the accuracy of the predictions is high for 5-fold cross-validation or 10-fold cross-validation. For training data of high N (N > 100) the prediction performance weakly depends on the number of validation samples. The DLBCL data show other results than the study of Subramanian and Simon (2011). 20-fold cross-validation performs marginally better than 5- and 10-fold cross-validation.

The performance of the IBSC tuning criterion is minimally higher than of the CVPL criteria. The variances are homogeneous. The CVPL tuning criteria outperforms IBSC only for the AML data using PLS and the shrinkage-based methods.

Model approaches applied to the generated high-signal data reveal a high prediction accuracy. The prediction performance is associated with the proportion of high-signal variables in the survival models. FSS, PLS, UPV and SPC detect a high number of correct variables (>= 95 %). The GBS method (90 % success rate) only shows average performance, RID and NET (detection rate of 89 and 87 % respectively) a high detection rate but low performance. PCR achieves low performance and a low success rate.

The tertiary objective of this work refers to features included in the survival prediction models. The AML and DLBCL data present a uniform picture. Only one gene is frequently included in the survival models and a high number of genes are rarely selected.

Chapter 6

6 Cure Models

Objectives

Chapter 6 is dedicated to survival prediction models for populations of mixed frail and cured patients. The models are based on microarray data. In chapter 6.1 Cox proportional hazards mixture cure models and accelerated failure time mixture cure models are introduced. In chapter 6.2 the study objectives are presented. The survival prediction procedure is introduced in chapter 6.3. In chapter 6.4 the datasets and in chapter 6.5 the software packages are described. The prediction performance of the Cox proportional hazards model, the Cox mixture cure and the AFT mixture cure models are compared in chapter 6.6.

Background and Introduction

In medical research the Cox proportional hazards model is a commonly used regression model to examine the effects of risk factors on survival. Censored observations in Cox models are considered as partly missing outcome information caused by short follow up periods of the patients.

Some other type of censoring may be applied to patients with serious diseases. Patients remain event-free for a long follow-up period and are assumed to be medically cured. For a population of partly frail and partly healed patients the Cox proportional hazards model may not be appropriate, since it does not account for cure. Survival cure models are suitable for the analysis of mixed frail and insusceptible populations.

Two classes of survival cure models are described in the statistical literature, the mixture cure models (Berkson and Gage, 1952) and the bounded cumulative hazards models (Yakovlev and Tsodikov, 1996).

Important representatives of the bounded cumulative hazards models are the proportional hazards mixture models and the gamma frailty models. The proportional hazards mixture model is based on the Cox regression model with the cumulative hazard function $\Delta(t)$ bounded to a value $\theta>0$. As time increases the cumulative hazard $\Delta(t)\rightarrow\theta$. The survival function of the proportional hazards mixture model is given by $S(t)=e^{-\theta F(t)}$, where $F(t)$ represents a cumulative distribution function of a random variable, $F(t)=\Delta(t)/\theta$ and $F(0)=0$ with $\theta>0$ and $t\geq 0$.

Berkson and Gage (1952) introduced the mixture cure model for populations of both frail and cured patients. The survival models are used “to estimate the cure rate of treatment and the survival rate of uncured patients at the same time” (Cai et al., 2012). The mixture cure model consists of a latency and an incidence part. The logistic regression model as well as the Cox model can be used to model the cure and the survival part (Cai et al., 2012).

Mixture cure models were developed by Farewell (1977, 1982) and Sy and Taylor (2000). Farewell (1977, 1982) modeled the cure and survival part of mixture cure models using logistic regression models and Weibull models. Peng and Dear (2000) used the logistic regression and the Cox regression model. Liu et al. (2012) introduced model selection techniques for semiparametric cure models. The SCAD (smoothly clipped absolute deviation) and the lasso approach were performed.

The second part of this thesis is dedicated to survival models from microarray data applied to mixed populations of long- and short term survivors. The Cox proportional hazards models, the Cox proportional hazards mixture cure models (CCM) and the accelerated failure time mixture cure models (ACM) are compared in terms of prediction performance.

The prediction accuracies of Cox proportional hazards and mixture cure models fitted to microarray data have never been compared in the statistical literature as far as this is known. Cox and cure models for clinical data or generated low-dimensional data were investigated. Exemplary the works of Perperoglou (2006) and Sy and Taylor (2000) are mentioned.

Perperoglou (2006) examined survival models for long-term survivors in cancer research. The data included clinical and pathological variables like the age at diagnosis, the tumor size, the status of the lymph nodes, grading, the administration of chemo- and radio therapy and the hormonal receptor status. Perperoglou examined the prediction accuracy of the standard Cox model, the frailty models (the Burr model and relaxed Burr model), the Cox proportional hazards model with time-varying effects (reduced rank model) and the semiparametric cure rate mixture model (a combination of the Cox PH model and the logistic regression model). The performance of the survival models was assessed with the Brier score and the explained variation R2 at some time points. Product limit estimates represented reference values for the survival estimates of the prediction models. The reduced rank model reached the highest performance with respect to the Brier score for early (<= 4 years) and the relaxed Burr model for late time points (5 to 10 years). Cure and Cox PH models performed almost equally and surpassed the Kaplan-Meier estimates. The reduced rank models (for late terms) and mixture cure models achieved the lowest, the standard Cox model medium and the Burr model the highest level of performance. The Cox PH model showed higher prediction accuracy than the cure model at late terms.

Sy and Taylor (2000) compared the parameter estimates of the Cox PH, the PH mixture cure and the Weibull cure model using clinical data and examined the parameter estimates of the mixture cure models for generated data. The simulated data exhibited mild or heavy censoring (0.10 and 0.40) and different cure rates (0.20 to 0.80). They were generated from a combined Weibull and logistic regression model ("logistic-exponential mixture model"). The models were evaluated by the mean squared error (MSE) of the parameter estimates. For mild censoring the PH mixture model and the Weibull mixture cure models achieved similar result for the incidence part of the model. For heavy censoring the PH cure model revealed a lower MSE than the Weibull cure model. The cure rate estimates of the PH cure model exhibited a lower or equal MSE compared to the Weibull cure model. Weibull cure models showed accurate estimates for baseline survival and the coefficients of the survival model.

Sposto (2002) compared parametric cure models and the standard Cox PH model based on pediatric cancer data that exhibit early events and a high cure rate. He suggested that parametric survival models are appropriate and reasonable for the analysis of cure data, as they allow an interpretation of long- and short-term effects on survival. Nevertheless they do not generally achieve a higher performance level than the Cox regression PH model for the case that the proportional hazards assumption is met.

Chapter 6 is dedicated to survival prediction models from microarray data in the case of long-term survivors. The prediction accuracy of the standard Cox model (COX), the Cox mixture cure model (CCM) and the accelerated failure time mixture cure model (ACM) are compared. A new survival prediction procedure is introduced (see 6.3 for details). It includes new model fitting techniques, which are derived from two approaches of chapter 4. The prediction procedure is used to tune, fit and assess the survival models. It contains:

1) model fitting approaches appropriate for mixture cure models,
2) tuning strategies to validate the survival models and
3) performance metrics to assess the accuracy of the survival predictions.

Chapter 6 of this thesis is arranged as follows:

In chapter 6.1 semiparametric mixture cure models, especially the Cox proportional hazards mixture cure model and the accelerated failure time mixture cure model are introduced and an EM algorithm for parameter estimation is described.

Chapter 6.2 contains the main research questions of the second part of this thesis. The new survival prediction procedure is presented in chapter 6.3. Datasets and software packages used to compare the survival models are described in chapter 6.4 and 6.5.

The prediction performance of mixture cure models and the Cox proportional hazards model are compared in chapter 6.6. Genes associated with survival and cure are presented.

6.1 Mixture cure models

Survival cure models are applicable for mixed populations of individuals that are susceptible and insusceptible to events. Susceptible patients, referred to as frail or non-cured, are assumed to experience an event if they are followed up for a long time. Nevertheless for some susceptible patients an event does not appear within the follow-up period and they are referred to as being censored. Insusceptible patients, also termed being immune, healed, cured or long-term survivors, remain event-free and the individuals are being censored after a certain cut-off point in the long-term follow-up.

Mixture cure models consist of an incidence and latency part that can be modeled separately by binary and survival regression models. Cai et al. (2012), Peng and Dear (2000) and Sy and Taylor (2000) introduced Cox mixture cure models and accelerated failure time mixture cure models. The following considerations are based on these three publications. The mixture cure model is defined as follows:

Assume that T represents a random variable that accounts for the survival time of the patients, z and x are the covariates that are related to incidence and latency, $\pi(z)$ is the probability of incidence and $S(t|x)$ the conditional survival probability of the uncured individuals given x . The survival function of the mixture cure model is given by:

$$S^c(t|x,z)=(1-\pi(z))+\pi(z)S(t|x) \text{ .}$$

The cure rate is represented by $1-\pi(z)$ and the rate of susceptible patients by $\pi(z)$. If all patients are frail, $\pi(z)=1$, the mixture cure model is reduced to a survival model $S(t|x)$ and if all patients are cured, the survival becomes 1.

Components of the mixture cure model

The parameter values of the incidence and latency part of the Cox proportional hazards mixture cure model (CCM) and the accelerated failure time mixture cure model (ACM) are estimated iteratively by the logistic and by the Cox and AFT regression model (Cai et al., 2012).

1) **Logistic regression model:** The covariate-dependent cure rate $1-\pi(z)$ of the mixture cure models is described by logistic regression models, where z are the explanatory variables and ζ represents the parameter vector. The incidence rate $\pi(z)$ is given by: $\pi(z)=\dfrac{\exp(\zeta' z)}{1+\exp(\zeta' z)}$. Alternatives to the logistic regression model like probit regression for instance are not considered in this text.

2a) **The Cox proportional hazards model** was described in chapter 2.2. The survival function for the non-cured at time t is given by: $S(t|x)=[S_0(t)]^{\exp(\beta' x)}$, where S_0 represents the baseline survival, x are the covariates and β are the parameter values.

2b) The accelerated failure time models (AFT) were described by Kalbfleisch and Prentice (2002). Accelerated failure time models can be expressed by the survival function $S(t|x)$ at time t given the covariates x : $S(t|x)=S_0(t\exp(\beta' x))$, where S_0 represents the baseline survival function and β are the parameter estimates. The term $e^{\beta' x}$ represents the "acceleration factor", a multiplicative effect on the survival time. The parameter values can be estimated by the linear-rank-test-based method (Kalbfleisch and Prentice, 2002).

In AFT models the logarithmic event time $\log(T)$ can be expressed by the mean μ , the coefficients β , the covariables x and the error term σW : $\log(T)=\mu+\beta' x+\sigma W$,where W is described by the Weibull distribution in this work and σ represents the shape parameter.

Parameter estimates

Cai et al. (2012) introduced an EM algorithm to calculate parameter values for the latency and incidence part of mixture cure models, whereas the likelihood function is estimated in the expectation step (E-step) and the parameter values in the maximization step (M-step). The following description was made by Cai et al. (2012) and includes M runs ($m=1,\ldots,M$).

Assume that $O=(t_i,\delta_i,z_i,x_i)$ are the observed data of patient $i=1,\ldots,N$, t_i are survival times, δ_i is the censoring status, which takes the value 1 if i is censored and 0 otherwise and z_i and x_i are the covariates related to the logistic and survival part of the mixture cure model.

$\Theta=(\zeta,\beta,S_0(t))$ are the parameter values of the logistic regression model, the survival model and the baseline survival respectively. The hazard and survival functions are denoted by h and S . The random variable Y takes the value $Y=1$ if an event will likely and $Y=0$ if an event will never occur.

If the y_i could be observed, the complete likelihood would be given by:

$$L(\zeta,\beta,O,y)=\prod_{i=1}^{N}[1-\pi(z_i)]^{(1-y_i)}\pi(z_i)^{y_i}h(t_i|Y=1,x_i)^{\delta_i y_i}S(t_i|Y=1,x_i)^{y_i}\,.$$

The log-likelihood function can be separated in one part that is related to the incidence l_{inc} and one part that is associated with the latency l_{lat} of the mixture cure model:

$$l_{inc}(\zeta,O,y)=\sum_{i=1}^{N}y_i\log(\pi(z_i))+(1-y_i)\log(1-\pi(z_i))\text{ and}$$

$$l_{lat}(\beta,O,y)=\sum_{i=1}^{N}y_i\delta_i\log(h(t_i|Y=1,x_i))+y_i\log(S(t_i|Y=1,x_i))\,.$$

In the **expectation step** the expected log-likelihood function $E[l(\zeta,\beta,O,y)]$ is calculated from the parameter estimates $\hat{\zeta}^{(m)}$, $\hat{\beta}^{(m)}$ and the unobserved y by: $w_i=E(y_i|O,\hat{\Theta}^{(m)})$.

The w_i are obtained by:

$$w_i = \delta_i + (1-\delta_i)\frac{\pi(z_i)S(t_i|Y=1,x_i)}{(1-\pi(z_i)) + \pi(z_i)S(t_i|Y=1,x_i)}|_{(O,\hat{\Theta}^{(m)})}.$$

The complete log-likelihood consists of $E(l_{inc}) + E(l_{lat})$, where y_i is estimated by w_i.

1) In the **maximization step** the parameter values of the incidence and latency part of the mixture model are estimated. The coefficients of the logistic model are obtained by the expected log-likelihood $E(l_{inc})$ that is given by: $E(l_{inc}) = \sum_{i=1}^{N} w_i^{(m)} \log(\pi(z_i)) + (1-w_i^{(m)})\log(1-\pi(z_i))$.

2) The **parameter values of the survival part** can be estimated by:
$E(l_{lat}) = \sum_{i=1}^{N} \delta_i \log(w_i^{(m)} h(t_i|Y=1,x_i)) + w_i^{(m)} \log(S(t_i|Y=1,x_i))$.

2a) In case of the **Cox regression model** the **parameter estimates** (Cai et al., 2012, based on Peng and Dear, 2000, and Sy and Taylor, 2000) are obtained by the expected log-likelihood $E(l_{lat})$ with the hazard function $h(t_i|Y=1,x_i) = h_0(t_i)\exp(\beta' x_i)$ and the survival function $S(t_i|Y=1,x_i) = S_0(t_i)^{\exp(\beta' x_i)}$, where $h_0(t_i)$ and $S_0(t_i)$ are the baseline hazard function and the baseline survival function.

The parameter values are estimated by the log partial likelihood method of a survival model with known coefficient $\log(w_i)$. The baseline survival function is recalculated by the Breslow method, where $\hat{S}_0(t|Y=1)$ are set to 0 for $t > t^*$ so that $\hat{S}_0(t|Y=1)$ does not fall to 0 for $t \rightarrow \infty$. The variable t^* represents the last event time. The baseline survival is obtained by:
$\hat{S}_0(t|Y=1) = \exp(-\sum_{k:t_k \le t} \frac{d_{t_k}}{\sum_{i \in R_{t_k}} w_i^{(m)} e^{\hat{\beta}(m)' x_i}})$, where t_k are sorted survival times, d_{t_k} and R_{t_k} are the number of deaths and the individuals at risk, respectively.

2b) The **parameter values in the AFT model** are obtained by a rank-based estimation technique presented by Zhang and Peng (2007). Computational details to estimate the parameter values β and the baseline survival S_0 can be found in Cai et al. (2012).

6.2 Study objectives

This chapter is dedicated to survival prediction models from gene expression data in case of long-term survivors. A new prediction procedure is presented to tune, fit and evaluate cure and Cox models based on microarray data. The prediction performance of the standard Cox proportional hazards (COX), the Cox proportional hazards mixture cure (CCM) and the accelerated failure time mixture cure models (ACM) are compared. Genes associated with survival and cure are examined.

The **primary research questions** of this chapter are:

- Which survival model leads to the highest prediction performance?
- Can mixture cure models surpass the prediction performance of the Cox proportional hazards model?
- Does the prediction accuracy remain constant over the follow up interval or is it varying between early and late terms?
- Do the results achieve high or low variation?

The **exploratory objectives** refer to features that are included in the survival prediction models. The research questions are tested using microarray data and generated cure data.

Exploratory objectives, which are examined based on the generated cure data:

- Does the prediction procedure choose the correct variables?
- Do the mixture cure models select the real long- and short term effects on survival?

Research questions investigated on the real data:

- Which genes affect survival?
- Which features have an impact on long- and short term survival?
- Which genes are incorporated in both Cox models and mixture cure models?

6.3 Model fitting and performance assessment

The primary and exploratory research questions are investigated by a new survival prediction procedure. The prediction procedure described in chapter 3.4 suits the research questions from the last chapter but is not fully applicable here. The main reason is that mixture cure models are used to model the data. The study population consists of mixed frail and cured subjects. This may lead to a non-proportional situation and the Cox regression model may not be appropriate to model the data.

Mixture cure models consist of an incidence and latency part. Model fitting techniques have to take both model parts into account.

The new survival prediction procedure is carried out as follows:

- The data are randomly split into training and test data at a ratio of 2:1. The training data are used to fit the prediction model and the test data to assess the survival estimates.

- Two model fitting techniques are applied: a modified version of the univariate selection approach in case of the generated survival data and an adapted supervised principal components technique for microarray data. 10-fold cross-validation is used to tune the number of features and principal components in the survival models. The model size is selected from the integrated Brier score. For Cox proportional hazards models one tuning parameter is selected and in the mixture cure model two tuning parameters are chosen (one for the survival and one for the cure part).

- The performance of the survival predictions is assessed using the test data. The prediction accuracy is evaluated with the integrated area under the curve, the integrated Brier score and the explained variation R2 based on Graf et al. (1999).

The results of the previous chapter will be the cornerstones for the new survival prediction procedure: A modified supervised principal components approach and an adapted univariate selection technique are used to predict survival from high-dimensional data. Supervised principal components regression demonstrated a high prediction performance for microarray data and univariate selection achieved accurate predictions for generated high-signal data.

A modified supervised principal components approach is used in this work. It uses 250 genes, which are univariately related to survival and 250 genes univariately associated with cure. Preliminary studies demonstrated that survival predictions achieved stable results, high performance and low computational costs, if:

- The subset of genes used to compute the principal components is not higher than 250.
- A maximum of 10 covariables were included in each part of the prediction model.

Results are not presented here.

A new survival prediction procedure

A new survival prediction procedure is introduced that tunes and fits mixture cure and Cox regression models. It includes model approaches, which are based on univariate effects on survival and cure to develop the survival models.

The prediction procedure works as follows, details are given below:

1) The survival status and survival time of the patients are used to estimate their **event probability** $\hat{y}_i$ with $i=1,2,\ldots,N$. The calculation algorithm of the $\hat{y}_i$ is derived from Peng's S-Plus package *semicure*.
2) **Effects on survival:** Univariate Cox and univariate AFT models are applied to each gene. The survival models include a term with logarithmic event probabilities $\log(\hat{y}_i)$. For each gene the hypothesis: $H_0:\hat{\beta}_j=0$ versus $H_A:\hat{\beta}_j\neq 0$ is tested, where $\hat{\beta}_j$ are the parameter estimates for each gene x_j $(j=1,\ldots,p)$ in the survival model. The most significant genes from the Cox models are included in the Cox mixture cure model. Features with the lowest p-values in AFT models are incorporated in accelerated mixture cure models.
3) **Effects on cure:** Univariate logistic regression models for each gene x_j are used to examine effects on the event probabilities $\hat{y}_i$. For each gene the hypothesis: $H_0:\hat{\zeta}_j=0$ versus $H_A:\hat{\zeta}_j\neq 0$ is tested ($\hat{\zeta}_j$ are the parameter estimates for each gene in the logistic regression model). The genes with the lowest p-values of the Wald test are assumed to be highly related to cure and are included in the latency part of the mixture cure models, both ACM and CCM.
4) The most significant genes from 2) and 3) are used to fit the mixture cure models.

Estimation of the event probabilities

The event probabilities $\hat{y}_i$ with $i=1,\ldots,N$ are derived from a code sequence of Peng's S-Plus package *semicure* (see *http://www.math.mun.ca/~ypeng/research/semicure/*, retrieved: July, 30, 2012). The $\hat{y}_i$ are computed by the following rules:

- If the patients experience an event, their event probability is 1 ($\hat{y}_i=1$).
- If they are censored after the last observed event, they are likely to be cured and an extremely low event probability ($\hat{y}_i=0.0001$) is allocated.
- Subjects that are censored before the last event take an uncure probability that depends on their observation time t_i . It is estimated from a linear decreasing function of time t . This function assigns the event probability $\hat{y}_i=1$ for the earliest t_1 and $\hat{y}_i=0.0001$ for the latest survival time t_n with $n\leq N$. The t_N represents the last observation time of the data.

Steps of the survival prediction procedure

The new survival prediction procedure works as follows. The data are randomly divided into a learning and a test dataset. The models size is selected by cross-validation using the learning data. Two tuning parameters are applied. One represents the number of variables in the latency, the other one the number of variables in the incidence part. The prediction model is fitted to the whole training data. The survival predictions are assessed on the basis of the test data.

A) Model tuning (using training data)

A1) For each combination of the tuning parameters λ_1, λ_2 = 1, 2,...,10 (number of the variables in the latency and incidence parts of the cure model) the following steps are performed:

I) 10-fold cross-validation is used to tune the survival models using the training data. Within each cross-validation sample the subsequent steps are applied:

i. Determine effects on survival and cure using the procedures described in items 2) and 3) (see details above) for $k-1$ validation samples. The p genes related to survival and to cure x_j and z_j , $j=1,\ldots, p$, are sorted by significance. $\hat{\beta}_1, \ldots, \hat{\beta}_{\lambda_1}$ (latency part) and $\hat{\zeta}_1, \ldots, \hat{\zeta}_{\lambda_2}$ (incidence part) are the parameter estimates of the mixture cure model. See chapter 6.1 for details.

ii. Predict survival for the k -th validation sample. The survival predictions are computed by the survival cure function $S^c(t|x, z)=(1-\pi(z))+\pi(z)S(t|x)$. $\pi(z)$ is the incidence function and $S(t|x)$ the survival function of the frail patients (chapter 6.1).

iii. Compute the integrated Brier score (IBSC) of the survival estimates using the k -th validation sample.

II) Calculate the median IBSC values of the ten values from iii.

A2) Choose the model size $(\hat{\lambda}_1, \hat{\lambda}_2)$ with the lowest median IBSC value.

B) Parameter estimation (using training data)
Fit the mixture cure model of size $(\hat{\lambda}_1, \hat{\lambda}_2)$ based on the processes explained in 2) and 3) using the whole training data. The parameter values $\hat{\beta}_1, \ldots, \hat{\beta}_{\hat{\lambda}_1}$ (latency part) and $\hat{\zeta}_1, \ldots, \hat{\zeta}_{\hat{\lambda}_2}$ (incidence part) are estimated by the EM algorithm described in chapter 6.1.

C) Model assessment (using test data)
The **survival estimates** on the basis of the test data are computed by the survival cure function $S^c(t|x, z)=(1-\pi(z))+\pi(z)S(t|x)$.

Realization

- Mixture cure models and pure Cox models are compared in this work. The new prediction procedure is used to fit and assess mixture cure models. The prediction procedure described in chapter 4.1 is applied to Cox models. The Cox PH models are tuned by 10-fold cross-validation and the IBSC tuning criteria. The model size is selected by the tuning parameter $\lambda=1,2,\ldots,10$ (number of covariables in the Cox model).

- The survival models are examined regarding microarray data and generated cure data. Univariate selection (UPV approach) is used to predict survival using the simulated data. For the mixture cure models the procedures presented in items 2) and 3) are used. For Cox models the UPV approach of chapter 4 is applied.

- In case of microarray data the survival models are fitted by the SPC approach. For each model part the 250 most significant univariate effects (items 2 and 3) are used to form the principal components. The prediction procedure works in the same way as described above but instead of single genes the principal components are used to predict survival. The pure Cox model uses the SPC approach of the last chapter. The principal components are built from the 250 most significant survival effects. Only the number of principal components in the survival model is tuned.

- The survival predictions $S^c(t|x, z)$ from the mixture cure models and $S(t|x)$ from the Cox PH model are evaluated in terms of the integrated Brier score IBSC (Graf et al., 1999), the integrated area under the receiver operating characteristic curve IAUC (Chambless and Diao, 2006) and the median explained variation R2 (Graf et al., 1999) across time t (in years). The prediction accuracy at certain time points t is assessed with the Brier score $BSC(t)$, the area under the ROC curve $AUC(t)$ and the explained variation $R2(t)$. Details are given in chapter 4.2.

- The area under the ROC curve at time t, $AUC(t)$, is determined by the event probability $1-S^c(t|x, y)$ and $1-S(t|x)$. The integrated AUC is calculated as a weighted sum of $AUC(t_k)$, where t_k with $k=1,\ldots, K$ are unique and sorted observation times.

Excursion

Evaluation of the survival prediction procedure for mixture cure models

This excursion is intended as an examination of the prediction procedure, which was described in the previous section. The aim of this supplement is to validate the parameter estimates from the prediction procedure with the estimates of the Cox mixture cure models.

The parameter estimates are assessed for the VDX data (Wang et al. 2005, see details below). The following procedures are applied for each feature:

- Univariate logistic regression models measuring the effects on the event probabilities $\hat{y}_i$.
- Univariate Cox regression models for non-survivors based on $\hat{y}_i$.
- Semiparametric Cox mixture cure models, where the same feature is incorporated in the latency and the incidence part of the model.

The parameter estimates and the p-values of the genes from logistic regression models and from the incidence part of the mixture cure models as well as the results from Cox models and the latency part of the mixture cure models are compared. The results of this comparison are summarized in table 6-1. The VDX data consist of 5000 variables. 4952 of 5000 regression models show reliable results.

The first analysis is dedicated to the p-values of the logistic regression model and of the incidence part of the mixture cure model. Concordance between the p-values of both models classified as < 0.05 or >= 0.05 and < 0.10 or >= 0.10 is observed in more than 97 % of the variables. The median absolute difference between the p-values was 0.033 for all models and 0.009 for variables, for which at least one of the p-values is below 0.10. The subgroup of genes with $p < 0.10$ in one or both models is presented separately, since genes with p-values in both models >= 0.10 are not included in a prediction model. The median parameter estimates are 0.009 and 0.013 for models with one or two p-values below 0.10.

The median differences of the intercept are < 0.001 for the p-values of the models and about 0.046 for the parameter estimates.

The second analysis refers to the latency part of the mixture cure model and the results from univariate Cox PH models. In about 94 % of the genes both of them are either significant ($p < 0.05$) or non-significant ($p >= 0.05$). In 93 % of the features both of them are either significant or non-significant with respect to the significance level of 0.10 ($p < 0.10$ versus $p >= 0.10$). The median difference of the p-values is 0.056 (median absolute difference of parameter estimates 0.014) and 0.019 (median absolute difference of parameter estimates 0.025) for models with at least one p-value below 0.10.

The results of this comparison study allow the conclusion that the parameter estimates from the new prediction procedure are “similar” to estimates from Cox mixture cure models. The new approach can perform model selection and can save computational time.

Comments to the prediction procedure applied to accelerated failure time mixture cure models (results are not shown here): Similar results were achieved regarding the incidence part of the survival models but obvious differences (p-values and parameter estimates) became apparent in the latency part. Irrespectively of this the same model approach is applied to Cox cure models and accelerated failure time mixture cure models.

It will be shown later that survival predictions based on the accelerated failure time mixture cure models strongly account for cure-related variables even in the latency part of the model. Nevertheless accelerated failure time mixture cure models will not generally be surpassed by Cox mixture cure models, although the new technique does not seem to fit perfectly for these models.

covariable	incidence part	latency part
p-values < 0.05 *	4831 (97.56 %)	4651 (93.92 %)
p-values < 0.10 *	4810 (97.13 %)	4598 (92.85 %)
Δ p-values **	0.009 [0.003, 0.021]	0.019 [0.006, 0.045]
Δ parameter estimates **	0.013 [0.006, 0.022]	0.025 [0.009, 0.034]
Δ p-values ***	0.033 [0.013, 0.064]	0.056 [0.021, 0.115]
Δ parameter estimates ***	0.009 [0.004, 0.016]	0.014 [0.007, 0.025]
intercept		
Δ p-values **	0.000 [0.000, 0.000]	
Δ parameter estimates **	0.047 [0.046, 0.049]	
Δ p-values ***	0.000 [0.000, 0.000]	
Δ parameter estimates ***	0.046 [0.045, 0.047]	

* concordance expressed by N (%)

** differences expressed by median [1.quartile, 3.quartile], where at least one model shows p < 0.10

*** differences expressed by median [1.quartile, 3.quartile] of all models

Table 6-1a: Parameter estimates and p-values obtained from the new model approach and the Cox mixture cure model (reference model). 5000 survival models, one for each gene, are fitted using the VDX data (details see chapter 6.4): The same gene is included in the latency and incidence part of each model. The covariables (and the intercept) of the incidence (left column) and the covariables of the latency part (right column) are compared. Analysis of the covariables: 1. (row 3-4) Number of variables in the two model approaches, of which both of them are either significant or not significant (p < 0.05 and < 0.10 respectively). 2. (row 5-8) Differences (median plus 25 % and 75 % percentiles) of the parameter estimates and the p-values (all genes and genes with p < 0.10) in absolute numbers. Analysis of the intercept: 3. (row 10-13) Analogous to 2.

	general information
dataset	VDX
number of covariables	5000 (100 %)
number of covariables with results	4952 (99.04 %)
running time	~ 10 hrs

Table 6-1b: General information about the algorithm to compare the new model approach and the Cox mixture cure model. Information about the dataset and running time of the procedure.

6.4 Datasets

The main objectives of this work are examined for generated cure data and for two breast-cancer datasets. These are the Netherlands Cancer Institute breast cancer data (NKI) by van't Veer et al. (2002) and breast cancer data (VDX) from the Erasmus Medical Center in Rotterdam (Netherlands) by Wang et al. (2005). The Netherlands breast cancer data (NKI) show data gaps, which are imputed by the k-nearest neighbor algorithm from Hastie et al. (1999). See chapter 5.4 for details.

Netherlands Cancer Institute breast cancer data

Van't Veer et al. (2002) published clinical and molecular data of 98 patients with primary breast cancer. The primary endpoint of the study was distant metastasis free survival. 34 patients developed metastasis within 5 years, 64 patients stayed event-free for more than 5 years including BRCA1 and BRCA2 carriers. The raw data included about 25000 human genes.

The gene expression data are publicly available on the website of the “Division of Molecular Carcinogenesis” from the Netherlands Cancer Institute via the URL *http://bioinformatics.nki.nl/data/van-t-Veer_Nature_2002/* (retrieved: November 30, 2012). Phenotype data were published in the “Nature” journal via the supplementary info site of the article “Gene expression profiling predicts clinical outcome” (van't Veer et al., 2002). The data are available on the website *http://www.nature.com/nature/journal/v415/n6871/abs/415530a.html* (retrieved: November 30, 2012).

The data were prepared for analysis taking into account the following considerations: Features with more than 10 % (10 values) and samples with more than 10 % (2418) missing values were excluded from the analysis. The k-nearest neighbor algorithm by Hastie et al. (1999) was applied via the R package *impute* to substitute missing values. 5000 features with the highest variance were considered for data analysis. The gene expression data were standardized, therefore each feature has mean 0 and variance 1.

Erasmus Medical Center breast cancer data

Wang et al. (2005) provided survival and genomic data of 286 breast cancer patients and recorded the distant metastasis free survival up to 14 years. The lymph nodes of the patients were unaffected and treatment was not administered before or after the procedure. 107 patients experienced a metastatic event. Their median follow up time was 2 years and 4 months. The median follow up time was 8 years and 8 months for metastasis free patients.

The gene expression dataset is available for public use from the GEO database with the accession number GSE2034. The data can be directly accessed via the URL *http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2034* (retrieved: November 25, 2012).

Benjamin Haibe-Kains provided Wang's breast cancer data in an R data file “wang2005.RData” with complete survival and molecular data. The dataset is available at *http://www.ulb.ac.be/di/map/bhaibeka/survcompaper/* (retrieved: November 25, 2012). The data include 286 observations and 22283 gene expression variables. 5000 genes with the highest variance (analogous to Subramanian and Simon, 2011) were considered for data analysis. The data were standardized to show mean 0 and variance 1.

The generated cure data (CGE)

Three cure datasets are generated including 200 observations and 1000 variables. The variables are standard normally distributed, whereas three variables (x_1, x_2, x_3) and six variables ($x_1, x_2, \ldots, x_6$) respectively are assumed to be related to survival and cure.

The event time and the event status are generated by the following considerations:

1) The parameter values (coefficients of the cure ζ and the survival part β) and the cure rate ($1-\pi_0$) of the survival model are set beforehand. The values are shown below.

2) The covariate dependent cure rate ($1-\pi(z)$) is calculated from a logistic regression model: The intercept of the regression model is obtained by $\zeta_0 = (\log((1-\pi_0)/\pi_0))$. A linear combination of the covariate effects (LK_ζ) is given by $LK_\zeta = \zeta_0 + \sum_{j=1}^{3[6]} x_j * \zeta_j$. One dataset includes 6 variables related to survival. Two datasets include 3 variables associated with the outcome. The covariate dependent cure rate is determined by $1-\pi = \exp(LK_\zeta)/(1+\exp(LK_\zeta))$.

3) 200 uniform random values ranging between 0 and 1 are drawn representing the survival probabilities p of the subjects. The Weibull survival times d are derived from p :

   - $d = \infty$ if $1-\pi \geq p$ for the cured subjects.

   - $d = (-\log((p-(1-\pi))/\pi)/\eta)^{(1/\gamma)}$ for the uncured patients.

   The time variable d for the frail patients follows a Weibull distribution with the parameter values η with $0<\eta\leq 1$ and $\gamma>0$. The η are assumed to depend on the covariates x_j : $\eta = \eta_0 * \exp(LK_\beta)$, where

   $$LK_\beta = \sum_{j=1}^{3[6]} x_j * \beta_j \, .$$

   The Weibull parameters γ and η_0 , which are the scale and the shape parameters of the Weibull distribution, have to be set beforehand.

4) 200 uniform random values ranging between 0 and 1 are drawn representing the censoring probabilities p_c of the observations. Censoring times c are generated by $c = p_c * a + f$, where a represents the maximal accrual and f the minimal follow up time.

5) The survival times t are obtained by $t = min(c, d)$. An observation is assumed to be censored if the censoring time c is shorter than the survival time d and a failure is assumed if vice versa. The Kaplan-Meier plot of the generated cure data is shown in the results section.

Specifications to generate the datasets

Three datasets are generated by the algorithm (1-5) described above. The coefficients (related to cure and survival), the cure rate, the two Weibull parameters, the maximal accrual time and the minimum follow up time are set beforehand.

1) **The first dataset (CGE) is generated from the following specifications:**

- The coefficients of the cure (ζ) and the survival part (β): $\zeta_1 = 2$, $\zeta_2 = -2$, $\zeta_3 = 3$ and $\beta_1 = -3$, $\beta_2 = 2$, $\beta_3 = 1$.
- The cure rate ($1-\pi_0$) is set to 50 %.
- The Weibull parameters η_0 and γ are set to 1.
- The maximal accrual time (a) and the minimal follow up time (f) are set to 2 and 5.

2) **The second dataset (I3V) includes three variables x_1 , x_2 , x_3 related to survival and cure. The effects of the variables on latency and incidence are opposite. This means that increasing values of the variables decrease survival time but increase cure:**

- The coefficients of the cure (ζ) and the survival part (β): $\zeta_1 = 2$, $\zeta_2 = 1.5$, $\zeta_3 = 1$ and $\beta_1 = 2$, $\beta_2 = 2$, $\beta_3 = 2$.
- The cure rate ($1-\pi_0$) is set to 30 %.
- The Weibull parameters η_0 and γ are set to 1 and 1.4.
- The maximal accrual time (a) and the minimal follow up time (f) are set to 5 and 5.

3) **The third dataset (I6V) includes six variables. x_1 , x_2 , x_3 are related to survival and x_4 , x_5 , x_6 are associated with cure.** The coefficients of the cure (ζ) and the survival part (β): $\zeta_1=\zeta_2=\zeta_3 = 0$, $\zeta_4 =2$, $\zeta_5 = 1.5$, $\zeta_6 = 1$ and $\beta_1 = 2$, $\beta_2 = 2$, $\beta_3 = 2$, $\beta_4=\beta_5=\beta_6 = 0$. Furthermore the data are generated from the same settings as the dataset I3V.

6.5 Software

The simulated cure data are generated by SAS (Statistical Analysis System; version 9.2) macro code. R (open source software for statistical computing, version 2.14.0) is used to prepare the datasets, to perform statistical analysis and to present the results in graphs and tables. Algorithms for the survival prediction procedure and the SAS programming steps are performed on three computer server of the “Medizinische Universität Wien”. Graphs and tables are created on a Linux desktop computer.

The R *base* and *survival* package is used to validate the survival models and to estimate the parameter values of logistic, Cox and AFT models. Mixture cure models are obtained by the R package *smcure*. Algorithms to tune and assess the models are based on the R packages *survcomp* and *ipred*.

Software packages for cure models

Many cure rate models in R and SAS demand a high working memory and a long runtime. The S-Plus package *semicure* by Peng (2003) includes semiparametric cure models. De Castro et al. (2010) and Stasinopoulos and Rigby (2007) introduced cure models via the GAMLSS framework. GAMLSS is a model class of generalized additive models which can be transformed to long term survival models (see R package *gamlss.cens* for details). The R package *nltm* by Garibotti and Tsodikov provides proportional hazards proportional hazards cure models and gamma frailty models.

Two further popular implementations are: Corbiere and Joly (2007) published the SAS macro *PSPMCM* version 1.1 for (semi-) parametric cure models. Finally the function *strsmix* for mixture cure models is available in the statistical software STATA.

6.6 Results of the comparison study based on cure data

Chapter 6.6 is dedicated to the prediction accuracy of the accelerated failure time mixture cure model (ACM), the Cox proportional hazards cure model (CCM) and the standard Cox proportional hazards model (COX). The survival models are applied to microarray data of long-term survivors. The new survival prediction procedure is used to fit mixture cure models and the prediction algorithms from chapter 3.4 are applied to develop Cox regression models. The performance of the survival models is assessed with the integrated area under the ROC curve (IAUC), the integrated Brier score (IBSC), median R2 values and with the AUC, BSC and R2 measurements at time $t=1,2,...$ for time in years.

The exploratory research questions are dedicated to genes which are related to survival and/or to cure. Influential genes on survival are identified by heuristics (see chapter 5.3 for details). In case of the modified SPC approach 10 genes with the highest "relevance score" regarding survival and cure are considered for analysis. For the generated dataset the explanatory variables in the survival model are not aggregated. Therefore effects on survival can be identified directly. All variables in the survival model are considered in the analysis of the exploratory objective.

The results of the comparison study are presented in three sections. The first and second are dedicated to real, the third to three generated cure datasets. Each section contains the analysis of the research questions and additional information about tuning parameter values of the survival models and the runtime of the algorithms.

6.6.1 Results of the Netherlands breast cancer data

The primary and exploratory research questions are applied to the Netherlands breast cancer (NKI) dataset published by van't Veer et al. (2002). The data consist of 98 patients and 5000 genes. The adapted SPC approach, described in chapter 6.3, is used to predict survival from mixture cure and Cox regression models.

6.6.1.1 Analysis of the primary objectives

The overall prediction accuracies of the COX, CCM and ACM models are assessed with IAUC, IBSC and the median R2 value across all t in the test sample. Table 6-2 presents the median as well as the first and third quartile of the performance values from 50 random splits of the NKI data. The survival models are shown in the columns and the performance criteria in the rows of the table.

	CCM	ACM	COX
IAUC	**0.730 [0.666, 0.778]**	0.720 [0.669, 0.770]	**0.730 [0.672, 0.768]**
IBSC	**0.170 [0.156, 0.184]**	0.171 [0.149, 0.185]	0.179 [0.161, 0.193]
R2	**0.155 [0.047, 0.210]**	0.150 [0.054, 0.248]	0.115 [0.027, 0.208]

Table 6-2: Performance of the survival models CCM, ACM and COX obtained from 50 random splits of the NKI data. The prediction accuracy is measured by IAUC, IBSC and R2. The models are assessed by the median as well as the first and third quartile of the performance values. High prediction performance is characterized by high IAUC as well as R2 and low IBSC values.

The results of the model comparison study are presented by boxplots to examine the spread of the performance values for the three survival models. Figure 6-1 presents the performance values (IAUC, IBSC and R2) for the CCM, ACM and COX model.

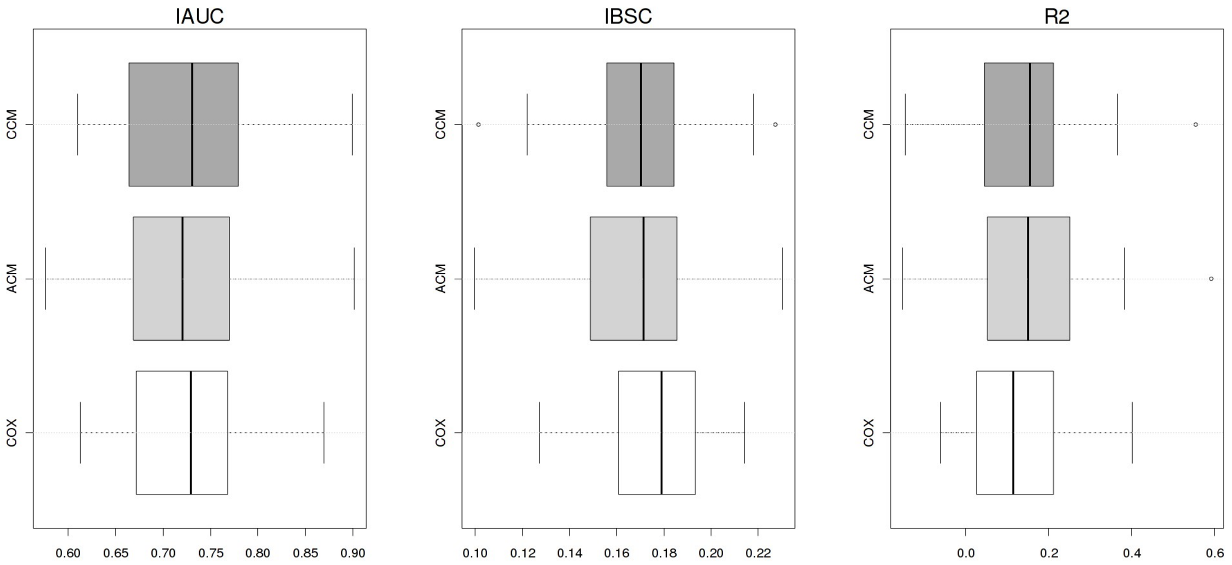

Figure 6-1: Prediction performance of the survival models CCM, ACM and COX measured by IAUC, IBSC and R2 (boxes). The values are obtained from 50 random splits of the NKI data. Models with high performance are characterized by high IAUC and R2 as well as low IBSC values.

The Cox mixture cure model marginally surpasses the accelerated failure time mixture cure model and the standard Cox model. No survival models achieve a noticeable variability of the performance values.

Results described in detail:

- The IAUC values of CCM and COX models are equal (0.730) and are marginally higher compared to ACM (0.720). CCM shows a minimally higher inter quartile range IQR (0.112) than ACM (0.101) and COX (0.096). Considering only the third quartile of the results CCM (0.778) reaches a higher performance level than COX (0.768). COX performs almost as good as ACM (0.770).

- The IBSC measurements of CCM and ACM are similar (0.170 and 0.171) and are lower than COX values (0.179). The IQR of the models is equivalent: CCM (0.028), ACM (0.036) and COX (0.032). Assessing only the first quartile of the IBSC values ACM (0.149) surpasses CCM (0.156) and COX (0.161) results.

- The median explained variation R2 is higher for CCM (0.155) than for ACM (0.150) and COX (0.115) models. The IQR shows low differences: CCM (0.163), ACM (0.194) and COX (0.181). ACM (0.248) surpasses CCM (0.210) and COX (0.208) regarding the 75th percentile of the R2 results.

Time-related results

A further analysis of the primary objective refers to the prediction performance of the CCM, ACM and COX models at early and late terms. The accuracy of the survival estimates is measured by the area under the ROC curve, $AUC(t)$, the Brier Score, $BSC(t)$, and the explained variation, $R2(t)$, at time t, where t represents years.

The survival curve of the NKI data is presented in figure 6-2. The survival function decreases until year 5 and then reaches a plateau. The maximal follow up is 13 years.

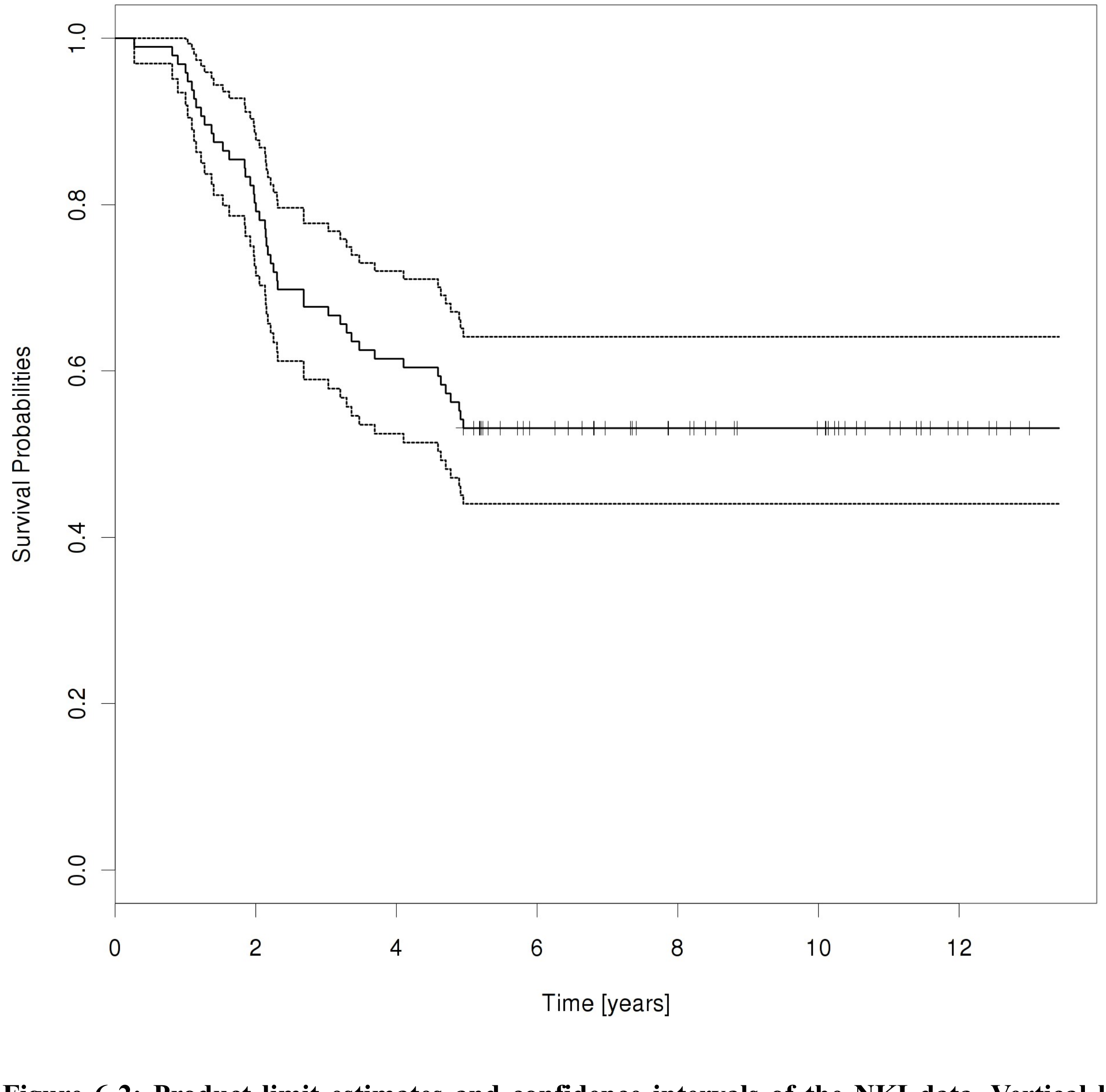

Figure 6-2: Product limit estimates and confidence intervals of the NKI data. Vertical lines represent censored observations.

Table 6-3 presents the prediction performance of the survival models at the time points $t=1,2,\ldots,12$ years measured by $AUC(t)$, $BSC(t)$ and $R2(t)$.

Year	**AUC**			**BSC**			**R2**		
	CCM	**ACM**	**COX**	**CCM**	**ACM**	**COX**	**CCM**	**ACM**	**COX**
Yr 1	0.613	0.556	**0.648**	0.047	0.047	**0.046**	-3.163	-6.264	**-0.311**
Yr 2	0.725	0.737	**0.750**	0.168	**0.157**	0.160	0.015	**0.071**	0.039
Yr 3	0.748	0.739	**0.751**	0.201	0.202	**0.200**	**0.110**	0.106	0.098
Yr 4	**0.739**	0.733	0.731	**0.217**	0.220	0.222	**0.096**	0.087	0.071
Yr 5	**0.747**	0.739	0.738	**0.221**	0.227	0.225	**0.118**	0.112	0.104
Yr 6	**0.745**	0.740	0.738	0.209	**0.208**	0.217	0.166	**0.168**	0.138
Yr 7	**0.745**	0.739	0.738	**0.212**	0.214	0.222	**0.154**	0.143	0.116
Yr 8	**0.745**	0.742	0.738	0.206	**0.205**	0.215	0.178	**0.183**	0.147
Yr 9	**0.745**	0.742	0.738	**0.174**	**0.174**	0.186	**0.307**	0.305	0.260
Yr 10*	**0.745**	0.742	0.738	0.159	**0.156**	0.173	0.363	**0.377**	0.312
Yr 11*	**0.745**	0.742	0.738	0.148	**0.143**	0.159	0.410	**0.432**	0.366
Yr 12*	**0.744**	0.737	0.737	0.152	**0.147**	0.164	0.393	**0.414**	0.347

Table 6-3: Performance of the survival models CCM, ACM and COX (single columns) obtained from 50 random splits of the NKI data. The prediction accuracy is measured by median values of $AUC(t)$, $BSC(t)$ and $R2(t)$ (global columns) at time $t=1,2,\ldots,12$. The highest performance is characterized by the highest AUC and R2 as well as the lowest BSC values. They are shown in bold numbers.
***At late time points the measurements might become slightly inaccurate as late observation times are not included in some test samples due to the random selection procedure.**

The standard Cox model shows high prediction performance in the years 1, 2 and 3. CCM and ACM models outperform the Cox model in the years 4 to 12. CCM seems to reach a higher level of performance than ACM within the period of year 4 to 9. ACM surpasses CCM in the period between year 10 to 12. For cure data the models addressing cures achieve a higher performance than pure Cox in the late term. The prediction models exhibit negative R2 values in year 1, which means they show poorer performance than the null model. This may be due to a low number of events in the first year.

Results described in detail:

- Regarding $AUC(t)$ values: COX models outperform CCM and ACM in year 1 (difference at least 0.035). COX reach a marginally higher level of performance than mixture cure models in year 2 and 3 (difference at least 0.013 and 0.003). Within the period of year 4 to 12 CCM is superior to ACM (difference 0.003 to 0.008) and COX (difference 0.007 to 0.009). ACM achieves a higher performance than COX (difference 0.000 to 0.004).

- With respect to $BSC(t)$ results: ACM and CCM models perform almost as good as COX models in the years 1 and 3 (difference <= 0.002). Mixture cure models outperform the COX model within the period of year 4 to 12. In the time interval from year 4 to 9 CCM performs almost as good as ACM (differences up to 0.006) and is superior over COX (differences 0.004 to 0.012). In the period between year 10 to 12 the ACM models outperform CCM (differences 0.003 to 0.005) and COX (differences 0.016 to 0.017).

- Referring to $R2(t)$ values: COX models perform best in year 1, ACM in year 2 and CCM in year 3. In the time span from year 3 to 9 CCM performs better than COX (0.012 to 0.047) and is superior over ACM (differences up to 0.009) except the years 6 and 8. Within the period of year 10 to 12 ACM outperforms CCM (differences 0.014 to 0.022) and COX (0.065 to 0.067).

Model size and running time of the algorithms

This subsection presents information regarding the model comparison study like the runtime of the algorithms and the size of the survival models.

The SPC approach was used to predict survival from NKI data. The number of principal components (1 to 10) in the survival and cure part of the prediction models was tuned. Table 6-4 shows the median number of principal components in the survival models. ACM and CCM models incorporate a median number of 4 components. 3 are related to latency and 1 is associated with incidence. The standard Cox regression models only include a single component.

The last row of table 6-4 presents the runtime of the survival prediction procedure based on ACM, CCM and COX models. It largely depends on the type of the survival model and the sample size. The computations of the COX models take 5 minutes, of the CCM models 6 minutes and 30 seconds and of the ACM models more than two hours (2 h, 7 min, 30 sec). This arises from the fact that parameter estimates in ACM models with a high number of variables are extremely time-consuming.

A single survival prediction procedure based on the Cox mixture cure model includes the calculation of 55000 univariate Cox models (R function *coxph*), 55000 univariate logistic regression models (R function *glm*) and 1002 mixture cure models (R function *smcure*). The prediction procedure based on Cox survival models contains 55000 univariate Cox models and 102 mixture cure models.

no. of PCs	**CCM**	**ACM**	**COX**
survival	3.00 [1.00, 5.00]	3.00 [1.00, 6.75]	1.00 [1.00, 2.00]
cure	1.00 [1.00, 2.00]	1.00 [1.00, 1.00]	-
running time [min]*	6.52 [6.49, 6.54]	127.55 [113.23, 142.83]	5.25 [5.24, 5.26]

Table 6-4: Number of principal components (median plus 25 % and 75 % percentiles) in the Cox (COX) and the incidence and latency parts of mixture cure models (CCM, ACM). The running time (median as well as 25 % and 75 % percentiles) of the algorithms for one prognosis based on cure and Cox models is shown in the last row of the table.
***Server specifications: Intel Core i7 CPU 2.80 GHz, 8 GB RAM.**

6.6.1.2 Analysis of the exploratory objectives

The impact of genes on survival and cure for standard Cox and mixture cure models is the exploratory objective of this work. As supervised principal components regression is a feature extraction strategy the influential genes are identified by heuristic procedures. A score value is assigned to each gene, which expresses the impact on incidence and latency. The 10 most influential genes on survival and cure for each cure model and the 10 most significant genes on survival for Cox regression models are considered for this analysis.

Table 6-5 presents genes that are included in the survival and cure parts of the prediction models. The genes are shown by decreasing counts (TOTAL). The incidence (.INC) and latency parts (.LAT) of ACM and CCM and the COX model are presented in the columns and the genes are shown in the rows. The results are based on 50 survival models, hence the cell frequency of a single gene in table 6-5 can be up to 50 and can reach a total of 250 for each gene (TOTAL). Genes that are selected at least 15 times in total are listed.

Six genes (number 3358, 3940, 4715, 3868, 4158 and 2012) are included in the survival models more than 50 times (every fifth model in total) and seven genes (number 4269, 4936, 4527, 193, 4854, 3716 and 3801) between 30 and 42 times in total.

The following findings can be reported:

- The top six genes show cell frequencies of at least 10 for the incidence parts of ACM and CCM as well as COX models. The latency part of CCM does not achieve cell frequencies above 4.

- ACM reveals the highest cell frequencies and 20 genes are selected at least 10 times (incidence part: 14 genes, latency: 6 genes). CCM (incidence: 12 genes, latency: 0 genes) and COX (12 genes) only include 12 genes at least 10 times.

- For ACM models the same genes are included in the latency and incidence part of the survival models.

- The incidence parts of ACM and CCM include mainly the same genes. The survival part of ACM and COX models incorporate identical genes. The cell frequencies within the top 6 features range from 6 to 14 (ACM) and from 10 to 17 (COX).

- The ACM, the incidence part of CCM and the COX model used the same genes to predict survival. Only one gene of the latency part of ACM (1447) and two genes of COX (4398 and 2507) are selected at least 10 times but are rarely chosen (one or two times) by the other models. Hence some important genes in COX models are only weakly related to other model parts (195, 905, 4398 and 2507).

Six genes were identified to play a large role for survival prediction. They are included in 20 % to 36 % of the survival models or model parts respectively (34-54 % related to incidence and 12 %-23 % related to latency). The results do not allow a clear distinction between genes associated with survival and others related to cure.

GENE	CCM.INC	CCM.LAT	ACM.INC	ACM.LAT	COX	TOT	TOT.INC	TOT.LAT
3358	23	4	31	14	17	**89**	54	35
3940	20	1	23	12	16	**72**	43	29
4715	18	4	24	7	13	**66**	42	24
3868	15	0	22	11	13	**61**	37	24
4158	15	1	21	9	10	**56**	36	20
2012	15	2	19	6	10	**52**	34	18
4269	16	0	16	2	8	**42**	32	10
4936	10	0	10	8	10	**38**	20	18
4527	10	0	12	10	6	**38**	22	16
193	4	0	5	10	19	**38**	9	29
4854	8	0	11	7	8	**34**	19	15
3716	9	2	12	2	8	**33**	21	12
3801	6	0	10	7	7	**30**	16	14
195	1	3	2	5	18	**29**	3	26
4167	7	0	9	5	7	**28**	16	12
2965	8	4	8	3	4	**27**	16	11
2769	12	0	8	4	3	**27**	20	7
345	8	0	10	5	4	**27**	18	9
3424	13	2	7	3	1	**26**	20	6
48	13	4	6	3	0	**26**	19	7
4119	5	2	8	2	4	**21**	13	8
3924	3	1	4	6	7	**21**	7	14
2426	7	0	7	4	3	**21**	14	7
1540	8	0	10	1	2	**21**	18	3
1363	6	0	8	2	5	**21**	14	7
4667	6	0	6	3	4	**19**	12	7
905	0	0	0	6	13	**19**	0	19
4398	0	1	0	0	16	**17**	0	17
4602	4	0	7	2	3	**16**	11	5
3248	5	0	6	2	3	**16**	11	5
2507	1	0	2	2	11	**16**	3	13
3844	7	1	5	1	1	**15**	12	3
2467	7	3	4	1	0	**15**	11	4
1447	1	1	1	10	2	**15**	2	13

Table 6-5: Participation of single genes in survival prediction obtained from 50 random splits of the NKI data. The frequency matrix presents in how many cases the genes (rows) are involved in the survival prognosis. The analysis is based on COX models and mixture cure models (ACM and CCM with latency and incidence parts). The table is sorted in decreasing

order of frequency (TOT). TOT.LAT and TOT.INC show in how many cases each gene is related to survival and to cure in total.

6.6.2 Results of the Erasmus Medical Center breast cancer data

In this section the main research questions of chapter 6 are examined for the Erasmus Medical Center breast cancer data (VDX) by Wang et al. (2005). The data consist of 286 patients and 5000 features. The new survival prediction procedure coupled with the modified supervised principal components approach is used to predict survival from microarray data.

6.6.2.1 Analysis of the primary objectives

The overall prediction accuracy of the COX, CCM and ACM models is measured by IAUC, IBSC and median R2 value for time t . Table 6-6 presents the median plus the first and third quartile of the performance values from 50 random splits of the VDX data.

	CCM	ACM	COX
IAUC	**0.666 [0.638, 0.690]**	0.658 [0.609, 0.700]	0.643 [0.611, 0.680]
IBSC	0.203 [0.190, 0.215]	**0.202 [0.189, 0.217]**	0.210 [0.192, 0.223]
R2	**0.029 [0.000, 0.049]**	0.025 [-0.007, 0.086]	0.017 [-0.023, 0.052]

Table 6-6: Performance of the survival models CCM, ACM and COX obtained from 50 random splits of the VDX data. The prediction accuracy is measured by IAUC, IBSC and R2. The models are assessed by the median plus the first and third quartile of the performance values. High prediction performance is characterized by high IAUC and R2 as well as low IBSC values.

Figure 6-3 presents the results by boxplots. The prediction performance of the CCM, ACM and COX models are evaluated with IAUC (left box), IBSC (box in the center) and R2 (right box).

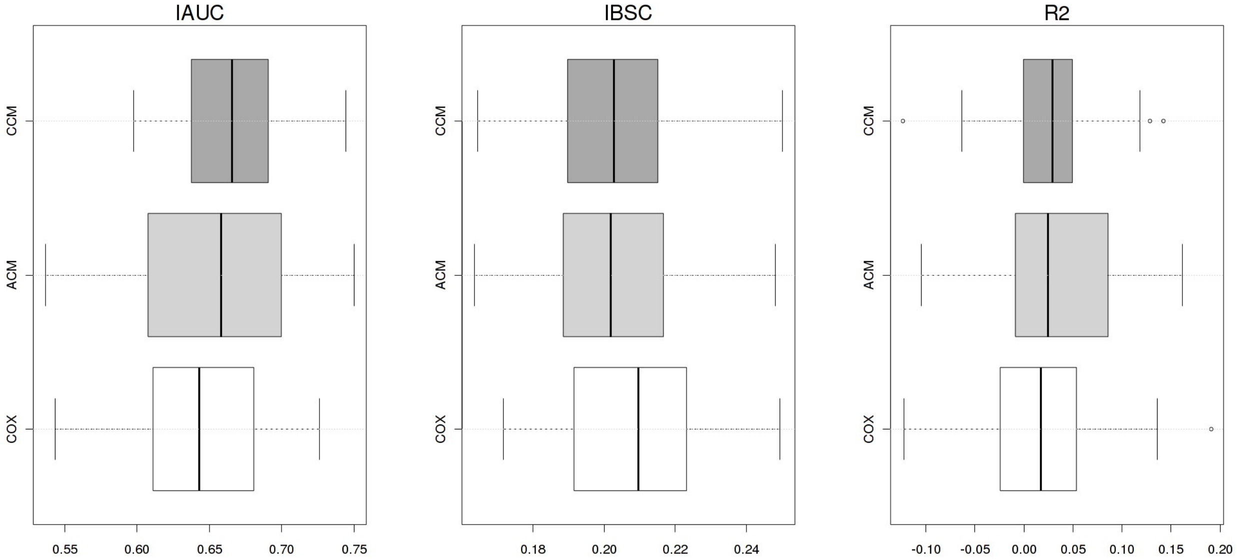

Figure 6-3: Prediction performance of the survival models CCM, ACM and COX measured by IAUC, IBSC and R2 (boxes). The values are obtained from 50 random splits of the VDX data. Models with high performance are characterized by high IAUC and R2 as well as low IBSC values.

The Cox mixture cure model outreaches the standard Cox model and the accelerated failure time mixture cure model. The two mixture cure models reach the same level of performance regarding median IBSC values. The CCM model shows the lowest and ACM the highest variability of the performance values.

Results described in detail:

- CCM (0.666) has a higher median IAUC value than ACM (0.658) and COX (0.643). ACM shows a higher IQR of IAUC (0.091) than COX (0.069) and CCM (0.052). CCM (0.690) and ACM (0.700) are superior over COX (0.680) based on the third quartile of the IAUC values.

- The median IBSC values of CCM and ACM (0.203 and 0.202) are lower than the median COX value (0.210). The IQR of the mixture cure models is marginally lower than the IQR of COX models: CCM (0.025), ACM (0.028) and COX (0.031). Considering only the first quartile of the IBSC values ACM (0.189) outperforms CCM (0.190) and COX (0.192) results.

- The median values of the explained variation R2 are low: CCM (0.029), ACM (0.025) and COX (0.017). The IQR of the R2 values is lower for CCM (0.049) than for COX (0.075) and ACM (0.093) models. With respect to the 75th percentile of the R2 results ACM (0.086) outreaches COX (0.052) and CCM (0.049).

Time-related results

The second analysis of the primary hypothesis investigates the prediction accuracy of the survival estimates related to time t by the area under the ROC curve, $AUC(t)$, the Brier Score, $BSC(t)$, and the explained variation, $R2(t)$.

Survival estimates of the VDX data are shown in figure 6-4. Patients are observed for a maximum of 14 years. The "cure plateau" is reached between 6 and 7 years of follow up.

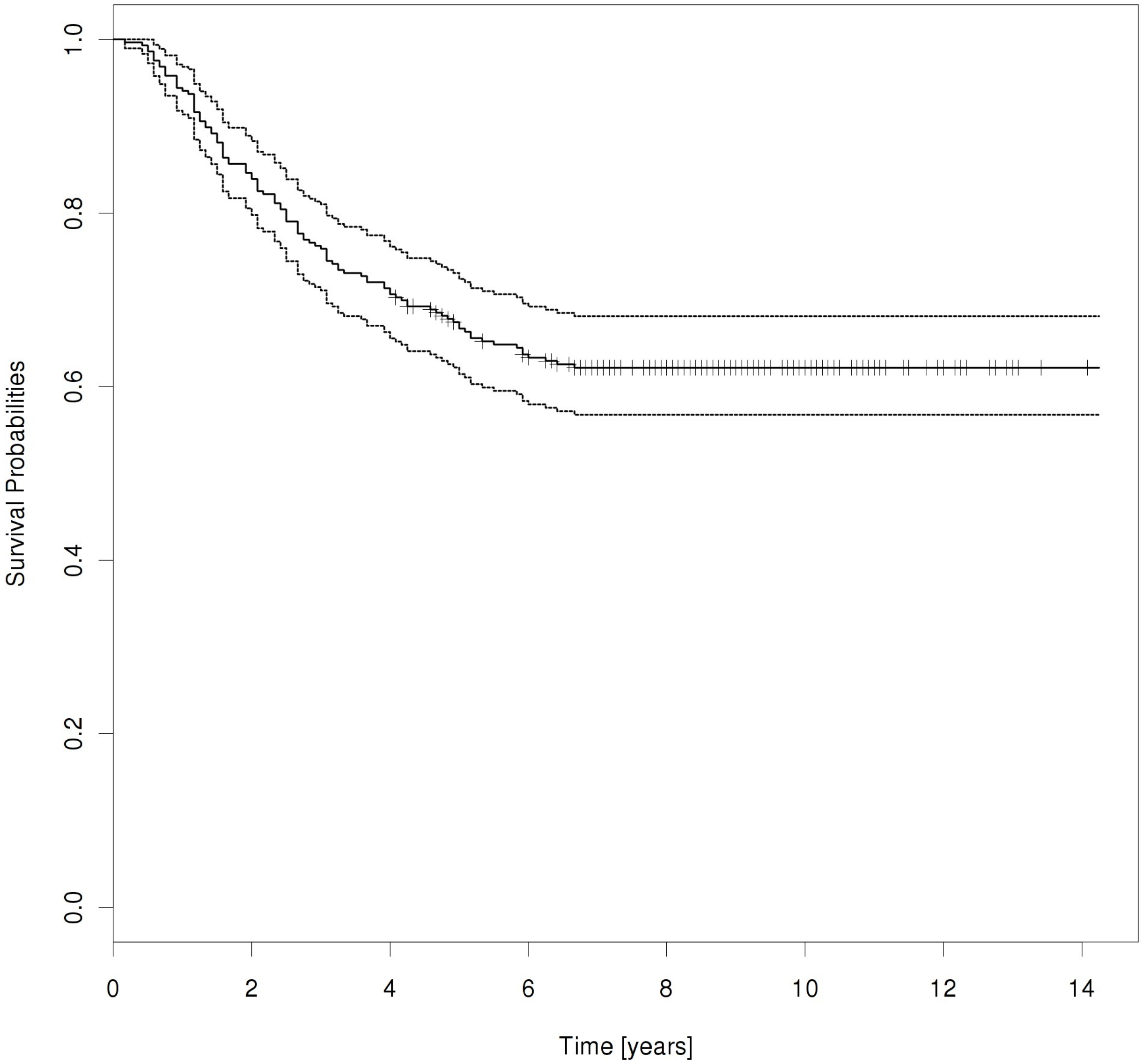

Figure 6-4: Product limit estimates and confidence intervals of the VDX data. Vertical lines represent censored observations.

Table 6-7 presents the values of $AUC(t)$, $BSC(t)$ and $R2(t)$ for the standard Cox and mixture cure models.

Year	**AUC**			**BSC**			**R2**		
	CCM	**ACM**	**COX**	**CCM**	**ACM**	**COX**	**CCM**	**ACM**	**COX**
Yr 1	**0.802**	0.676	0.671	**0.053**	0.055	0.055	**0.050**	0.031	0.016
Yr 2	**0.733**	0.661	0.652	**0.128**	0.134	0.134	**0.088**	0.058	0.050
Yr 3	**0.713**	0.657	0.656	**0.165**	0.175	0.175	**0.115**	0.072	0.074
Yr 4	**0.681**	0.650	0.652	**0.191**	0.198	0.198	**0.084**	0.055	0.061
Yr 5	**0.663**	0.655	0.652	**0.210**	0.212	0.215	**0.058**	0.057	0.052
Yr 6	0.654	**0.662**	0.655	0.224	**0.221**	0.226	0.045	**0.061**	0.042
Yr 7	0.644	**0.661**	0.650	0.232	**0.227**	0.235	0.023	**0.043**	0.015
Yr 8	0.644	**0.661**	0.650	0.235	**0.231**	0.240	0.008	**0.024**	-0.007
Yr 9	0.644	**0.660**	0.650	0.234	**0.232**	0.240	0.011	**0.025**	-0.009
Yr 10	0.644	**0.660**	0.650	0.242	**0.241**	0.246	-0.022	**-0.009**	-0.031
Yr 11*	0.644	**0.659**	0.650	0.255	**0.254**	0.266	-0.074	**-0.062**	-0.115
Yr 12*	0.644	**0.659**	0.650	**0.269**	**0.269**	0.280	-0.135	**-0.118**	-0.177
Yr 13*	0.644	**0.655**	0.650	0.307	**0.304**	0.313	-0.293	**-0.258**	-0.318
Yr 14*	0.628	**0.645**	0.643	0.432	**0.426**	0.481	-0.822	**-0.757**	-1.028

Table 6-7: Performance of the survival models CCM, ACM and COX (single columns) obtained from 50 random splits of the VDX data. The prediction accuracy is measured by median values of $AUC(t)$, $BSC(t)$ and $R2(t)$ (global columns) at time $t=1,2,\ldots,14$. The highest performance is characterized by the highest AUC and R2 as well as the lowest BSC values. They are shown in bold numbers.
***At late time points the measurements might become slightly inaccurate as late observation times are not included in some test samples due to the random selection procedure.**

The Cox mixture cure model exhibits high prediction accuracy in the first five years and the accelerated failure time mixture models in the last nine years. ACM and COX reach the same performance level within the period of year 1 to 5 and CCM surpasses COX results in the period between year 6 and year 14 with $BSC(t)$ and $R2(t)$ but not using $AUC(t)$. Compared to the NKI dataset the $AUC(t)$ values are lower, $R2(t)$ are much lower and $BSC(t)$ values are higher.

Results described in detail:

- Regarding $AUC(t)$ values: The CCM models surpass COX and ACM results with declining differences in the period from year 1 to year 5 (at least: 0.126 [year 1], 0.072, 0.056, 0.029 and 0.008 [year 5]). The differences between ACM and COX are < 0.01 in the first 5 years. In the time span from year 6 to 10 ACM surpasses COX and CCM (difference between 0.007 and 0.017). The COX model outperforms CCM results in the time interval from year 6 to 14 (difference 0.001 to 0.015).

- Within the period from year 1 to 5 the median $BSC(t)$ value of CCM is superior over ACM and COX models (differences 0.002 to 0.01). The ACM and COX reach the same level of performance (difference < 0.003). In the time span from year 6 to 14 ACM marginally surpasses CCM (differences < 0.006) and COX (differences <= 0.055). Hence CCM values are superior over COX values (differences 0.002 to 0.049).

- CCM reach a higher level of performance than ACM (difference 0.001 to 0.043) and COX (0.006 to 0.041) models with regard to the R2 values in the time span from year 1 to year 5. The results of COX and ACM in the time span from year 1 to year 5 are ambiguous. Within the time period from year 6 to 14 ACM outperforms CCM (difference between 0.012 and 0.065) and COX models (difference between 0.019 and 0.271). Hence CCM outreaches COX in the time interval from years 6 to year 14 (differences 0.003 to 0.206). In the period between the years 10 and 14 the R2 values are negative, which means the survival models are inferior towards the null model. The null model represents a survival model without covariates.

Model complexity and running time of the algorithms

The runtime of the survival prediction procedure and the model size of CCM, ACM and COX models are presented in this section. Survival estimates from the SPC approach were applied to the VDX data. The median number of components in the prediction models is presented in table 6-8.

CCM models incorporate 2 principal components (one for latency and one for incidence), ACM models contain 3 principal components (two versus one) and standard Cox models one principal component.

The runtime of the prediction procedure based on the standard Cox models takes almost 6 minutes, for the CCM model 10 minutes and more than eight hours computing the ACM model. This takes up a lot more time than for the NKI data. The VDX data include much more observations (286 subjects compared to 97 patients in the NKI data).

no. of PCs	**CCM**	**ACM**	**COX**
survival	1.00 [1.00, 2.00]	2.00 [1.00, 4.00]	1.00 [1.00, 3.75]
cure	1.00 [1.00, 1.75]	1.00 [1.00, 2.00]	-
running time [min]*	10.09 [10.00, 10.21]	488.35 [468.21, 508.53]	5.82 [5.81, 5.83]

Table 6-8: Number of principal components (median as well as 25 % and 75 % percentiles) in the Cox (COX) and the incidence and latency parts of mixture cure models (CCM, ACM). The running time (median plus 25 % and 75 % percentiles) of the algorithms for one prognosis based on cure and Cox models is shown in the last row of the table.
***Server specifications: Intel Core i7 CPU 2.80 GHz, 8 GB RAM.**

6.6.2.2 Analysis of the exploratory objectives

Since survival prediction using the VDX data is applied by the supervised principal components technique the impact of features on survival and cure is achieved by heuristic procedures. The 10 genes with the highest impact on cure and the 10 features mostly related to survival for each survival model are considered for analysis.

Table 6-9 presents genes that are related to incidence (.INC) and latency parts (.LAT) for ACM and CCM and to survival for COX models by decreasing counts (TOTAL). Single genes are shown in the rows of the table. Each gene can be included in the models (or model parts) for a total of 250 times (TOTAL) and a single cell frequency can be up to 50. Genes selected at least 15 times in total are listed.

Eleven genes (number 3032, 3535, 3843, 2496, 2497, 3180, 3246, 3839, 3648, 3071 and 3161) are included in the survival models more than 50 times (every fifth model in total) and five genes (number 513, 3952, 3912, 2652 and 2495) between 31 and 38 times in total.

The following findings can be reported:

- The top eleven genes are included in the incidence parts of ACM and CCM 15 to 37 times as well as in the COX models 8 to 23 times. They are incorporated in the latency part of CCM 0 times and in ACM 3 to 11 times.

- CCM selects 25 genes at least 10 times (incidence part: 13 genes, latency: 12 genes), ACM 21 genes at least 10 times (incidence: 14 genes, latency: 7 genes) and COX 11 genes at least 10 times.

- The top 11 genes are frequently incorporated in the latency and incidence part of the ACM models. The CCM model allows a clear distinction between genes exclusively related to survival and other features solely associated with cure.

- For cure models the results of the incidence parts are almost equal, since individual genes are selected by the same prediction procedure. The same genes are incorporated in the survival part of ACM and in the standard COX model. They are included in COX models more often. The genes in the survival part of CCM do not correspond with the genes selected in the survival parts of ACM and the standard COX model.

Eleven genes can be detected that have an impact on more than 20 % of the models (model parts). Two genes are even included in 39 % of all model parts. A high number of genes are included in ACM, the incidence part of CCM and COX models. The results allow a clear distinction between genes exclusively associated with survival and others solely related to cure only for CCM models.

GENE	CCM.INC	CCM.LAT	ACM.INC	ACM.LAT	COX	TOT	TOT.INC	TOT.LAT
3032	37	0	37	11	23	**108**	74	34
3535	30	0	34	11	22	**97**	64	33
3843	28	0	30	9	22	**89**	58	31
2496	29	0	28	9	18	**84**	57	27
2497	28	0	30	9	16	**83**	58	25
3180	24	0	27	8	19	**78**	51	27
3246	26	0	22	7	9	**64**	48	16
3839	15	0	19	5	17	**56**	34	22
3648	19	0	21	4	11	**55**	40	15
3071	21	0	18	4	11	**54**	39	15
3161	20	0	20	3	8	**51**	40	11
513	4	1	3	19	11	**38**	7	31
3952	12	0	12	3	7	**34**	24	10
3912	9	0	11	5	6	**31**	20	11
2652	0	31	0	0	0	**31**	0	31
2495	10	0	10	3	8	**31**	20	11
2882	0	29	0	0	0	**29**	0	29
371	0	29	0	0	0	**29**	0	29
3402	0	28	0	0	0	**28**	0	28
4232	4	6	2	8	7	**27**	6	21
3781	6	0	8	5	8	**27**	14	13
2737	4	0	2	10	10	**26**	6	20
479	0	26	0	0	0	**26**	0	26
4057	0	25	0	0	0	**25**	0	25
3589	8	0	9	4	3	**24**	17	7
1903	0	24	0	0	0	**24**	0	24
670	0	24	0	0	0	**24**	0	24
3438	8	0	8	3	4	**23**	16	7
518	0	21	0	0	0	**21**	0	21
1863	7	0	9	2	2	**20**	16	4
1902	0	19	0	0	0	**19**	0	19
626	4	0	3	10	2	**19**	7	12
197	2	6	1	5	5	**19**	3	16
3057	4	0	4	3	6	**17**	8	9
3053	5	0	5	3	4	**17**	10	7
123	2	3	1	7	4	**17**	3	14
1822	6	0	7	2	1	**16**	13	3
381	1	0	1	11	3	**16**	2	14
4800	1	13	0	1	0	**15**	1	14
4356	2	0	0	10	3	**15**	2	13
8	0	12	0	2	1	**15**	0	15

Table 6-9: Participation of single genes in survival prediction obtained from 50 random splits of the VDX data. The frequency matrix presents in how many cases the genes (rows) are

involved in the survival prognosis. The analysis is based on COX models and mixture cure models (ACM and CCM with latency and incidence parts). The table is sorted in decreasing order of frequency (TOT). TOT.LAT and TOT.INC show in how many cases each gene is related to survival and to cure in total.

6.6.3 Results for the generated cure datasets

The main objectives of this thesis are examined for three generated cure datasets (CGE, I3V, I6V). The data include 200 observations and 1000 variables. The parameter values depicting almost identical or opposite effects of the variables on survival and cure. This section accentuates the selection of variables into the model parts of the mixture cure models.

The following three cases are examined in this subchapter:

1) **Three variables are related to latency and incidence.** Three variables (x_1, x_2, x_3) are related, 997 random variables ($x_4, \dots, x_{1000}$) are not associated with survival and cure. The effects on survival and cure are expressed by the coefficients of the cure ($\zeta_1, \zeta_2, \zeta_3$) and the survival part ($\beta_1, \beta_2, \beta_3$).

2) **Three variables are associated with latency and incidence. The effects of the variables on latency and incidence are opposite.** Three variables (x_1, x_2, x_3) are associated and 997 random variables ($x_4, \dots, x_{1000}$) are not associated with survival and cure.

3) **Three variables are related to latency and other three variables to incidence.** Three variables (x_1, x_2, x_3) are associated with survival, three variables (x_4, x_5, x_6) are related to cure and 994 random variables ($x_7, \dots, x_{1000}$) are not associated with survival or cure.

A detailed description of the data is given in chapter 6.4. An adapted univariate selection approach is used to develop survival models for all generated datasets.

6.6.3.1 Three variables related to latency and incidence

Chapter 6.6.3.1 is dedicated to investigations regarding the generated dataset with 200 observations and 1000 variables. Three variables represent effects on latency and incidence. The coefficients of the cure (ζ) and the survival part (β) are set to: $\zeta_1 = 2$, $\zeta_2 = -2$, $\zeta_3 = 3$ and $\beta_1 = -3$, $\beta_2 = 2$, $\beta_3 = 1$ respectively.

6.6.3.1.1 Analysis of the primary objectives

The primary objective of this chapter, the prediction performance based on the COX, CCM and ACM model, is analyzed using the IAUC, IBSC and median R2 values. Table 6-10 presents the median as well as the first and third quartile of the performance values from 50 random splits of the CGE data.

	CCM	**ACM**	**COX**
IAUC	0.902 [0.867, 0.919]	**0.917 [0.892, 0.931]**	0.875 [0.843, 0.908]
IBSC	0.130 [0.111, 0.157]	**0.114 [0.102, 0.131]**	0.139 [0.119, 0.173]
R2	0.442 [0.354, 0.510]	**0.493 [0.424, 0.556]**	0.423 [0.270, 0.480]

Table 6-10: Performance of the survival models CCM, ACM and COX obtained from 50 random splits of the CGE data. The prediction accuracy is measured by IAUC, IBSC and R2. The models are assessed by the median plus the first and third quartile of the performance values. High prediction performance is characterized by high IAUC and R2 as well as low IBSC values.

A graphical representation of the results is shown in figure 6-5.

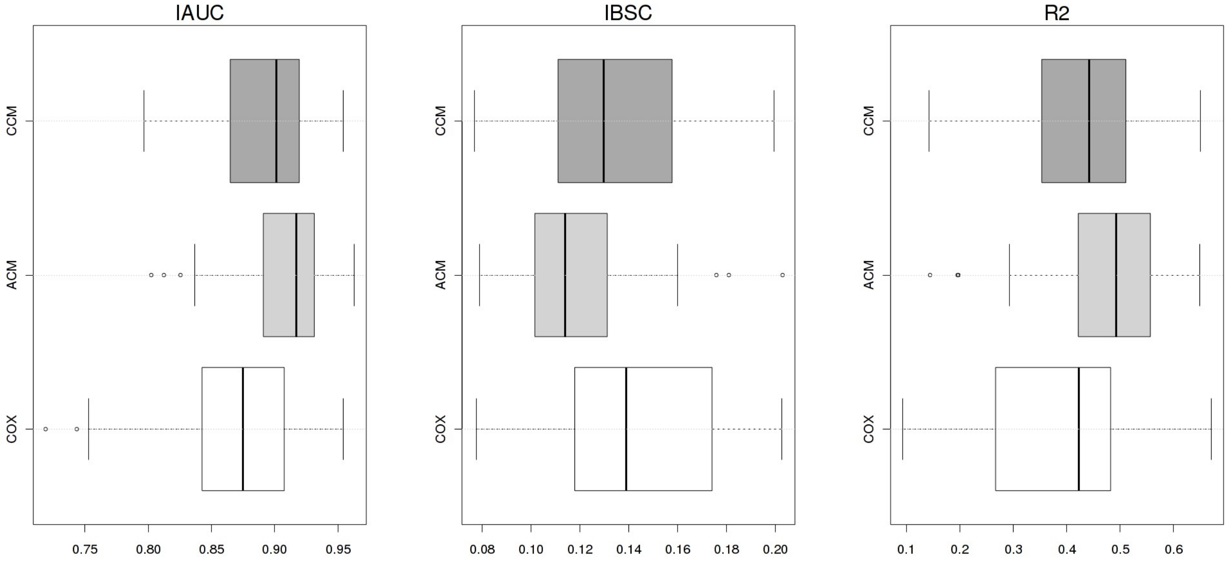

Figure 6-5: Prediction performance of the survival models CCM, ACM and COX measured by IAUC, IBSC and R2 (boxes). The values are obtained from 50 random splits of the CGE data. Models with high performance are characterized by high IAUC and R2 as well as low IBSC values.

The accelerated failure time mixture cure model surpasses the Cox mixture cure model, which achieves a higher level of performance than the standard Cox model. ACM results exhibit a low variation and COX models show the highest variability.

Results described in detail:

- ACM achieves a higher median value of IAUC (0.917) than CCM (0.902) and COX (0.875). The IQR is lower for ACM (0.039) than for CCM (0.052) and COX (0.065).

- ACM shows a lower median IBSC value (0.114) than CCM (0.130) and COX (0.139). ACM models (IQR: 0.029) provide a lower variability than CCM (IQR: 0.046) and COX (IQR: 0.054) models.

- ACM (0.493) reveals a higher median R2 value than CCM (0.442) and COX (0.423). The IQR of the R2 results is lower for ACM (0.132) than for CCM (0.156) and COX (0.210).

Time-related results

The second analysis of the primary objective is assessed with the area under the ROC curve, $AUC(t)$, the Brier Score, $BSC(t)$, and the explained variation, $R2(t)$, at time t, where t is referred to as years on account of simplicity.

Survival estimates of the CGE data are presented by Kaplan-Meier plots in figure 6-6. The survival probabilities decrease heavily in the first year and the cure plateau is reached at year 5. The performance criteria are calculated at year $t=1,2,\ldots,7$.

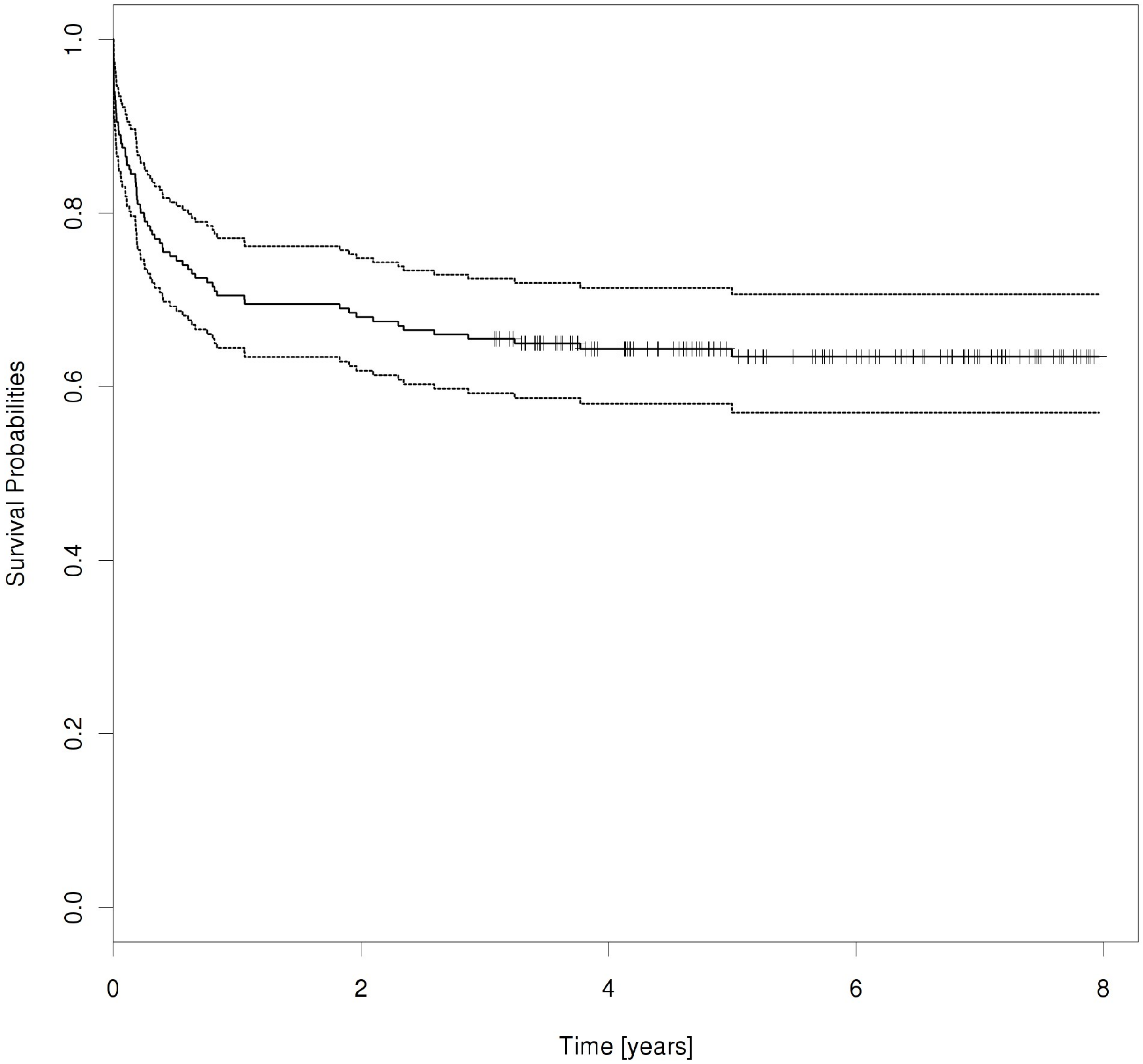

Figure 6-6: Product limit estimates and confidence intervals of the generated cure data. Vertical lines represent censored observations.

The values of the performance criteria $AUC(t)$, $BSC(t)$ and $R2(t)$ are shown in table 6-11. The survival models are presented in single columns and the time in the rows.

Year	**AUC**			**BSC**			**R2**		
	CCM	**ACM**	**COX**	**CCM**	**ACM**	**COX**	**CCM**	**ACM**	**COX**
Yr 1	0.900	**0.916**	0.880	0.119	**0.108**	0.129	0.475	**0.498**	0.422
Yr 2	0.905	**0.919**	0.880	0.118	**0.108**	0.134	0.482	**0.514**	0.421
Yr 3	0.892	**0.909**	0.866	0.129	**0.117**	0.144	0.448	**0.494**	0.399
Yr 4	0.887	**0.906**	0.862	0.135	**0.122**	0.151	0.430	**0.479**	0.379
Yr 5	0.882	**0.901**	0.854	0.148	**0.129**	0.165	0.387	**0.458**	0.328
Yr 6	0.877	**0.901**	0.854	0.145	**0.130**	0.159	0.399	**0.455**	0.352
Yr 7*	0.877	**0.900**	0.854	0.150	**0.128**	0.152	0.380	**0.464**	0.382

Table 6-11: Performance of the survival models CCM, ACM and COX (single columns) obtained from 50 random splits of the CGE data. The prediction accuracy is measured by median values of $AUC(t)$, $BSC(t)$ and $R2(t)$ (global columns) at time $t=1,2,\ldots,7$. The highest performance is characterized by the highest AUC and R2 as well as the lowest BSC values. They are shown in bold numbers.
***At late time points the measurements might become slightly inaccurate as late observation times are not included in some test samples due to the random selection procedure.**

The accelerated failure time mixture cure model outperforms CCM and COX models. The differences regarding model performance increase over time. The Cox cure model reaches a higher level of performance than the Cox model.

Results described in detail:

- ACM achieves the highest $AUC(t)$ values, whereas the differences between the $AUC(t)$ values increase over time (CCM: 0.016 to 0.024 and COX: 0.036 to 0.047). The CCM outperforms COX models (differences between 0.020 and 0.028).

- With respect to $BSC(t)$ values: The ACM outperforms CCM and COX. The differences to CCM increase over time (0.010 to 0.022), whilst the differences to COX increase up to year 5 (0.021 to 0.036) and then slightly decrease (to 0.024). The CCM outperforms COX models (differences range from 0.010 to 0.017) for year 1 to 6 and the survival models reach the same level of performance in year 7.

- Based on $R2(t)$ values: The ACM models outperform CCM and COX. The differences regarding prediction performance increase over time (CCM: 0.023 to 0.084 and COX: 0.076 to 0.130). For the last two years the values of COX reach the performance level of CCM results. The CCM outperforms COX (differences between 0.047 and 0.061 except year 7).

Model complexity and running time of the algorithms

This section is dedicated to the runtime of the model prediction procedure and the number of variables included in CCM, ACM and COX models. The survival prediction procedure coupled with the modified univariate selection approach (UPV) was used to predict survival from the generated cure data. The median number of variables included in the prediction models is shown in table 6-12.

CCM models incorporate 5.5 variables (latency: 2 and incidence: 3.5), ACM comprises even 9 variables (5 versus 4) and the standard Cox models 4 variables, whilst 6 variables would be appropriate for mixture cure and 3 variables for Cox models.

A single survival prediction procedure based on the standard Cox models takes 1 minute. For CCM 6 minutes and 30 seconds are needed and for the ACM model more than thirteen hours. The reason is the high number of objects (200) and the lower power of the computer server (see specifications in table 6-12).

no. of variables	**CCM**	**ACM**	**COX**	**CORRECT**
survival	2.00 [2.00, 3.00]	5.00 [4.00, 6.00]	4.00 [2.00, 5.75]	3
cure	3.50 [3.00, 4.00]	4.00 [3.00, 5.00]	-	3
running time [min]*	6.46 [5.32, 6.79]	789.11 [770.86, 803.40]	1.07 [1.06, 1.08]	-

Table 6-12: Number of variables (median plus 25 % and 75 % percentiles) in the Cox (COX) and the incidence and latency parts of mixture cure models (CCM, ACM). CORRECT represents the number of correct variables. The running time (median as well as 25 % and 75 % percentiles) of the algorithms for one prognosis based on cure and Cox models is shown in the last row of the table.
***Server specifications: Intel Core 2 CPU 2.66 GHz, 8 GB RAM.**

6.6.3.1.2 Analysis of the exploratory objectives

The exploratory objective of this chapter examines whether the survival prediction procedure (based on the three survival models) selects the correct survival model (variables 1, 2, 3) or uses noise.

Table 6-13 presents in how many cases the correct variables (three variables 1, 2 and 3) and noise variables are included in the survival models. The frequency table is based on 50 prediction models. Therefore a gene can be included 50 times in each model part and 250 times in total (TOT). The correct variables (1, 2, 3) and the noise variables (OTHERS) are shown in the rows of the table. The cure (.INC) and survival parts (.LAT) of ACM and CCM as well as the results of the COX model are shown in the columns of table 6-13.

VARS	CCM.INC	CCM.LAT	ACM.INC	ACM.LAT	COX	TOT	TOT.INC	TOT.LAT	β	ζ
1	50	49	50	50	50	**249**	100	149	-3	2
2	49	39	49	50	44	**231**	98	133	2	-2
3	42	5	46	41	33	**167**	88	79	1	3
OTHERS	59	67	65	110	98	**399**	124	275	-	-

Table 6-13: Participation of single genes in survival prediction obtained from 50 random splits of the CGE data. The frequency matrix presents in how many cases the genes (rows) are involved in the prediction models. The analysis is based on COX models and mixture cure models (ACM and CCM with latency and incidence parts). VARS 1, 2 and 3 (rows) represent the variables associated with survival and cure. OTHERS describe the number of noise variables in the survival models. TOT.LAT and TOT.INC show in how many cases each variable is related to survival and to cure in total. The two columns on the right represent the true parameter values of the variables related to survival and cure.

Variables that are associated with survival and cure are frequently detected by the survival prediction procedure (frequencies > 33). The third variable in the survival part of CCM models is included in only 5 prediction models. This is due to the weak effect of the third variable on survival.

The following findings can be reported:

- The first variable was incorporated in almost all models (249/250), the second in 231 models and the third variable in two of three survival models (167).

- ACM includes a higher number of correct variables (286/300 = 95 %) than COX (127/150 = 85 %) and CCM (234/300 = 78 %). Nevertheless ACM includes 175 noise variables (38 % of all model variables), COX 98 (44 %) and CCM 126 (35 %) noise variables.

- The incidence part of the survival models had a higher detection rate of correct variables (95 %: ACM 97 % and CCM 94 %) than the survival parts of the prediction models (80 %: ACM 94 %, CCM 62 % and COX 85 %).

The prediction models included a high number of correct variables and a high proportion of noise variables. ACM detected a high proportion of the correct variables and CCM had a low failure rate but incorporated variable 3 in only 5 of 50 survival models (low effect of the third variable on survival).

6.6.3.2 Three variables associated with survival and cure (oppositely related)

The research questions are investigated regarding a generated dataset with 200 observations and 1000 variables. Three variables (x_1 , x_2 , x_3) represent the effects on latency and incidence. The data are generated from a SAS macro described in chapter 6.4 with the following specifications: The parameter values of the variables related to cure are $\zeta_1 = 2$, $\zeta_2 = 1.5$ and $\zeta_3 = 1$, the coefficients associated with survival are $\beta_1 = 2$, $\beta_2 = 2$ and $\beta_3 = 2$. The Kaplan-Meier curve of the data (I3V) is shown below.

6.6.3.2.1 Analysis of the primary objectives

The prediction performance of the COX, CCM and ACM model is assessed with the IAUC, IBSC and median R2 values. Table 6-14 presents the median plus the first and third quartile of the performance values from 50 random splits of I3V.

	CCM	**ACM**	**COX**
IAUC	0.786 [0.724, 0.833]	**0.837 [0.775, 0.908]**	0.726 [0.693, 0.761]
IBSC	0.169 [0.149, 0.186]	**0.145 [0.120, 0.165]**	0.195 [0.190, 0.203]
R2	0.151 [0.071, 0.219]	**0.289 [0.163, 0.530]**	0.044 [0.017, 0.070]

Table 6-14: Performance of the survival models CCM, ACM and COX obtained from 50 random splits of the I3V data. The prediction accuracy is measured by IAUC, IBSC and R2. The models are assessed by the median plus the first and third quartile of the performance values. High prediction performance is characterized by high IAUC and R2 as well as low IBSC values.

A graphical representation of the results is shown in figure 6-7.

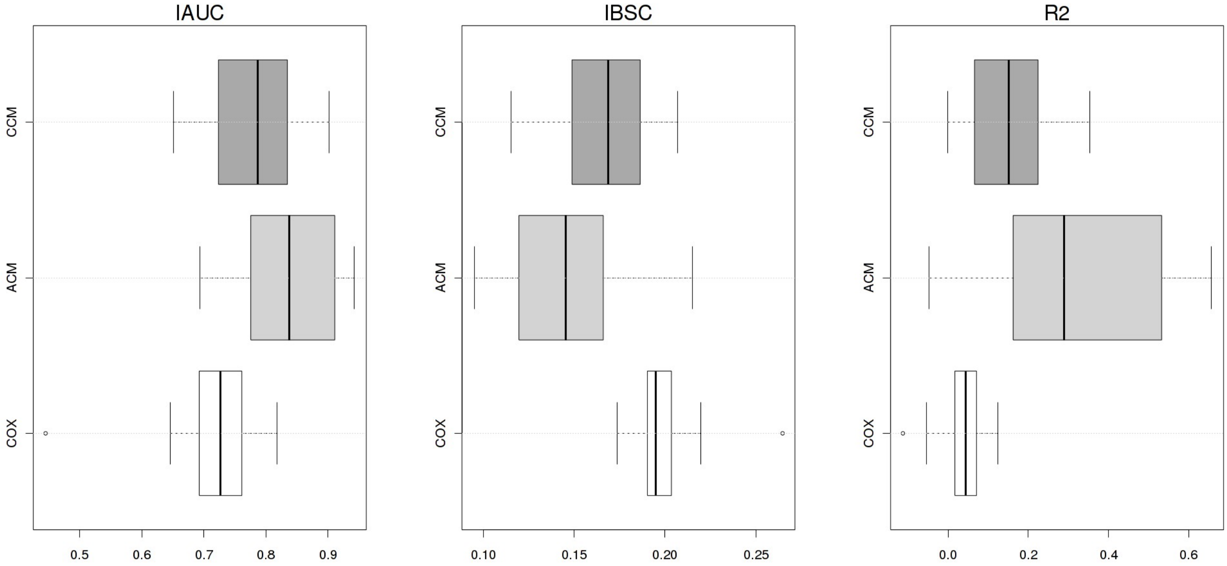

Figure 6-7: Prediction performance of the survival models CCM, ACM and COX measured by IAUC, IBSC and R2 (boxes). The values are obtained from 50 random splits of the I3V data. Models with high performance are characterized by high IAUC and R2 as well as low IBSC values.

The accelerated failure time mixture cure model outperforms the Cox mixture cure model, which surpasses the standard Cox model. The performance values of the accelerated failure time mixture cure model have a higher variation than the Cox mixture cure model and the standard Cox model.

Results described in detail:

- ACM shows a higher median value of IAUC (0.837) than CCM (0.786) and COX (0.726). The IQR is lower for COX (0.068) than for CCM (0.109) and ACM (0.133).

- ACM achieves a higher prediction accuracy (0.145) than CCM (0.169) and COX (0.195) regarding the median IBSC values. ACM models (IQR: 0.045) reveal a higher variability than CCM (IQR: 0.037) and COX (IQR: 0.013) models.

- ACM (0.289) has a higher median R2 value than CCM (0.151) and COX (0.044). The IQR of the R2 results is higher for ACM (0.367) than for CCM (0.148) and COX (0.053).

Time-related results

The second analysis is dedicated to the accuracy of the survival predictions at early and late terms. The performance is assessed with the area under the ROC curve, $AUC(t)$, the Brier Score, $BSC(t)$, and the explained variation, $R2(t)$, at time t, where t is referred to as years on account of simplicity.

Survival estimates of the I3V data are shown by Kaplan-Meier plots in figure 6-8. The survival estimates decrease heavily in the first year and the cure plateau is reached at year 6 to 7. The performance criteria are calculated for year $t=1,2,\dots,9$.

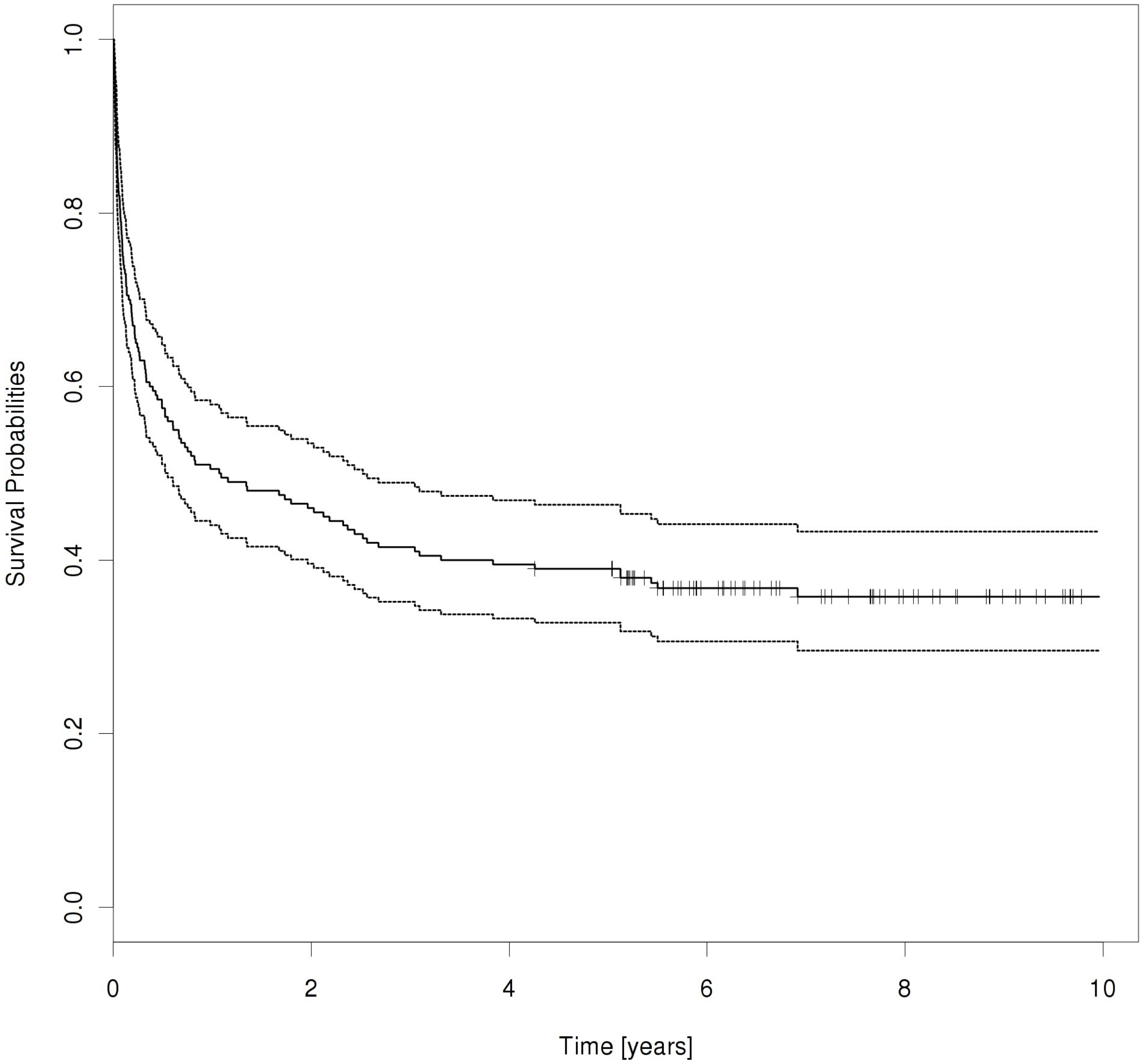

Figure 6-8: Product limit estimates and confidence intervals of the generated cure data (I3V). Vertical lines represent censored observations.

The values of the performance criteria $AUC(t)$, $BSC(t)$ and $R2(t)$ are shown in table 6-15. The survival models are presented in single columns and the time in the rows.

Year	AUC			BSC			R2		
	CCM	ACM	COX	CCM	ACM	COX	CCM	ACM	COX
Yr 1	0.692	**0.775**	0.635	0.227	**0.187**	0.244	0.136	**0.355**	0.034
Yr 2	0.723	**0.787**	0.663	0.211	**0.182**	0.234	0.184	**0.363**	0.068
Yr 3	0.748	**0.819**	0.699	0.195	**0.164**	0.219	0.223	**0.396**	0.104
Yr 4	0.777	**0.835**	0.728	0.181	**0.154**	0.208	0.266	**0.423**	0.136
Yr 5	0.779	**0.839**	0.725	0.178	**0.151**	0.207	0.270	**0.428**	0.134
Yr 6	0.804	**0.857**	0.749	0.170	**0.140**	0.200	0.281	**0.448**	0.143
Yr 7	0.838	**0.870**	0.779	0.142	**0.119**	0.181	0.390	**0.518**	0.213
Yr 8*	0.836	**0.871**	0.779	0.141	**0.112**	0.178	0.398	**0.551**	0.224
Yr 9*	0.836	**0.872**	0.779	0.108	**0.104**	0.148	0.537	**0.589**	0.355

Table 6-15: Performance of the survival models CCM, ACM and COX (single columns) obtained from 50 random splits of the I3V data. The prediction accuracy is measured by median values of $AUC(t)$, $BSC(t)$ and $R2(t)$ (global columns) at time $t=1,2,\ldots,9$. The highest performance is characterized by the highest AUC and R2 as well as the lowest BSC values. They are shown in bold numbers.
***At late time points the measurements might become slightly inaccurate as late observation times are not included in some test samples due to the random selection procedure.**

The results of the I3V and the CGE data are similar. The accelerated failure time model outperforms the Cox mixture cure model and the standard Cox model. The Cox mixture cure model achieves a higher level of performance than the standard Cox model.

Results described in detail:

- ACM reveals higher $AUC(t)$ values (between 0.775 and 0.872) than CCM (between 0.692 and 0.838) and COX (between 0.635 and 0.779). The differences of the $AUC(t)$ values between ACM and CCM (between 0.083 and 0.032) and between ACM and COX (between 0.140 and 0.091) slightly decrease over time. The values between CCM and COX remain constant over time (between 0.049 and 0.060).

- ACM (between 0.187 and 0.104) surpasses CCM (between 0.227 and 0.108) and COX (between 0.244 and 0.148) models with respect to $BSC(t)$ values. The differences between ACM and CCM remain constant over time (between 0.017 and 0.040) except for year 9, where the results are almost equal (difference 0.004). The difference between ACM and COX values are constant (between 0.044 and 0.066), whilst the differences between CCM and COX marginally increase over time (between 0.017 and 0.04).

- With respect to $R2(t)$ values: The ACM model (between 0.355 and 0.589) outperforms CCM (0.136 to 0.537) and COX (0.034 to 0.355). Differences regarding prediction performance decrease over time between ACM and CCM (year 1 to 9: from 0.219 to 0.052) as well as ACM and COX (year 1 to 9: from 0.321 to 0.234). Differences between CCM and COX (year 1 to 9: from 0.102 to 0.182) increase over time.

Model complexity and running time of the algorithms

In this section the runtime of the survival prediction procedure based on CCM, ACM and COX models and the number of model variables are examined. The number of variables included in the prediction models is given in table 6-16.

CCM models incorporate a median number of 3 variables (latency: 1 and incidence: 2), ACM includes even 10.5 variables (3 versus 7.5) and the standard Cox models incorporate only a single variable. 6 variables would be appropriate for mixture cure models. A single survival prediction with the standard Cox models takes almost 50 seconds, with the CCM models 4 minutes and with the ACM model 4 hours.

Effects on survival and opposite effects on cure cause a non-proportional situation. For Cox regression models survival and cure effects partially cancel each other out.

no. of variables	**CCM**	**ACM**	**COX**	**CORRECT**
survival	1.00 [1.00, 4.50]	7.50 [3.00, 9.75]	1.00 [1.00, 1.00]	3
cure	2.00 [1.00, 3.00]	3.00 [1.00, 5.00]	-	3
running time [min]*	3.98 [3.45, 4.32]	235.72 [164.95, 271.19]	0.78 [0.78, 0.79]	-

Table 6-16: Number of variables (median plus 25 % and 75 % percentiles) in the Cox (COX) and the incidence and latency parts of mixture cure models (CCM, ACM). The running time (median as well as 25 % and 75 % percentiles) of the algorithms for one prognosis based on cure and Cox models is shown in the last row of the table. CORRECT represents the number of correct variables.
***Server specifications: Intel Core i7 CPU 2.80 GHz, 8 GB RAM.**

6.6.3.2.2 Analysis of the exploratory objectives

The exploratory objective of this chapter examines whether the survival prediction procedure (based on the three survival models) selects the correct survival model (variables 1, 2, 3) or includes confounders.

Table 6-17 shows the frequency of correct variables (three variables 1, 2 and 3) and noise in the survival models. The frequency table is based on 50 prediction models. A gene can be included 50 times in each model part and 250 times in total (TOT). The correct variables (1, 2, 3) and the noise variables (OTHERS) are presented in the rows of the table. The cure (.INC) and survival parts (.LAT) of ACM and CCM and the results of the COX model are shown in the columns of table 6-17.

VARS	CCM.INC	CCM.LAT	ACM.INC	ACM.LAT	COX	TOT	TOT.INC	TOT.LAT	β	ζ
1	50	1	50	50	49	**200**	**100**	**100**	2	2
2	26	4	33	28	2	**93**	**59**	**34**	2	1.5
3	19	38	23	21	0	**101**	**42**	**59**	2	1
OTHERS	19	107	69	221	16	**432**	**88**	**344**	-	-

Table 6-17: Participation of single genes in survival prediction obtained from 50 random splits of the I3V data. The frequency matrix presents in how many cases the genes (rows) are involved in the prediction models. The analysis is based on COX models and mixture cure models (ACM and CCM with latency and incidence parts). VARS 1, 2 and 3 (rows) represent the variables associated with survival and cure. OTHERS describe the number of noise variables in the survival models. TOT.LAT and TOT.INC show in how many cases each variable is related to survival and to cure in total. The two columns on the right represent the true parameter values of the variables related to survival and cure.

Variables related to survival and cure are frequently identified by the prediction procedure based on the accelerated mixture cure model and by the incidence part of Cox mixture cure models (frequencies > 19). Variable 1 and 2 are seldom included in the survival part of Cox mixture cure models and variable 2 and 3 are rarely incorporated in standard Cox models. For Cox regression models effects on cure plateaus seem to be more important than effects on survival time.

The following findings can be reported:

- The first correct variable was incorporated in 80 % of all models (200/250), the second in only 37 % (93/250) of the models and the third variable in 40 % (101/250) of the models.

- ACM includes a higher number of correct variables (205/300 = 68 %) than CCM (138/300 = 46 %) and COX (51/150 = 34 %). Nevertheless ACM models involve 290 noise variables (59 % of all model variables), CCM 126 (48 %) and COX 16 (24 %).

- The incidence part of the survival models includes a higher number of correct variables (67 %: ACM 71 % and CCM 63 %) than the survival parts of the prediction models (43 %: ACM 66 %, COX 34 % and CCM 29 %).

The cure and survival part of ACM and the survival part of CCM incorporate a high number of correct variables but ACM included about 300 noise variables. The cure part of CCM showed a low failure rate but the survival part did not involve the correct variables. Signal variables seem to be only included in the survival part if they are not involved in the cure part. The standard Cox model includes a low number of noise variables but did only detect the second and third correct variable.

6.6.3.3 Three variables related to latency and the other three variables related to incidence

The research questions are investigated regarding the generated dataset with 200 observations and 1000 variables. Three variables are associated with cure (variable 4 to 6) and three variables are related to survival (variable 1 to 3).

The parameter values of the variables related to cure are $\zeta_4 = 2$, $\zeta_5 = 1.5$, $\zeta_6 = 1$, the coefficients associated with survival are set to $\beta_1 = 2$, $\beta_2 = 2$ and $\beta_3 = 2$. $\zeta_1=\zeta_2=\zeta_3 = 0$ and $\beta_4=\beta_5=\beta_6 = 0$. The Kaplan-Meier curve of the data (I6V) is shown below.

6.6.3.3.1 Analysis of the primary objectives

The prediction performances of the COX, CCM and ACM model are assessed with the IAUC, IBSC and with median R2 values. Table 6-18 presents the median plus the first and third quartile of the performance values from 50 random splits of the I6V data.

	CCM	ACM	COX
IAUC	**0.837 [0.793, 0.872]**	0.824 [0.794, 0.854]	0.813 [0.776, 0.843]
IBSC	**0.153 [0.134, 0.178]**	0.156 [0.141, 0.171]	0.169 [0.155, 0.188]
R2	**0.380 [0.308, 0.448]**	0.164 [0.118, 0.213]	0.152 [0.088, 0.185]

Table 6-18: Performance of the survival models CCM, ACM and COX obtained from 50 random splits of the I6V data. The prediction accuracy is measured by IAUC, IBSC and R2. The models are assessed by the median plus the first and third quartile of the performance values. High prediction performance is characterized by high IAUC and R2 as well as low IBSC values.

A graphical representation of the results is shown in figure 6-9.

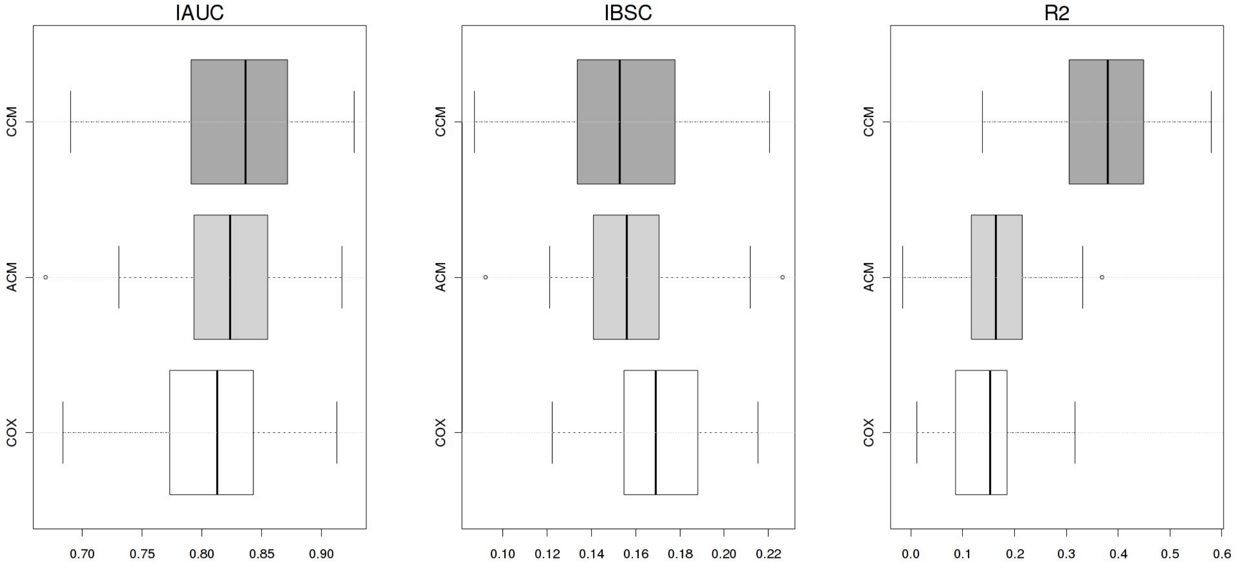

Figure 6-9: Prediction performance of the survival models CCM, ACM and COX measured by IAUC, IBSC and R2 (boxes). The values are obtained from 50 random splits of the I6V data. Models with high performance are characterized by high IAUC and R2 as well as low IBSC values.

The Cox mixture cure model outperforms the accelerated failure time mixture cure model. The latter achieves a higher level of performance than the standard Cox model. The results of the accelerated failure time mixture cure model and the Cox model show a lower variation than results from the Cox mixture cure model.

Results described in detail:

- CCM achieves a higher median value of IAUC (0.837) than ACM (0.824) and COX (0.813). The IQR is lower for ACM (0.060) and COX (0.067) compared to CCM (0.079).

- CCM shows a higher level of performance (0.153) than ACM (0.156) and COX (0.169) with respect to the median IBSC value. ACM (IQR: 0.030) and COX models (IQR: 0.033) provide a lower variability than CCM (IQR: 0.044) models.

- CCM (0.380) reveals a higher prediction accuracy than ACM (0.164) and COX (0.152) models related to the median R2 value. The IQR of the R2 results is lower for ACM (0.095) and COX (0.097) compared to CCM (0.140).

Time-related results

In a further analysis the primary research question is assessed with the area under the ROC curve, $AUC(t)$, the Brier Score, $BSC(t)$, and the explained variation, $R2(t)$, at time t, where t is referred to as years.

Survival probabilities of the generated data are presented by Kaplan-Meier plots shown in figure 6-10. The survival estimates decrease heavily in the first few months and the cure plateau is reached at year 5. The performance criteria are calculated for year $t=1,2,...,9$.

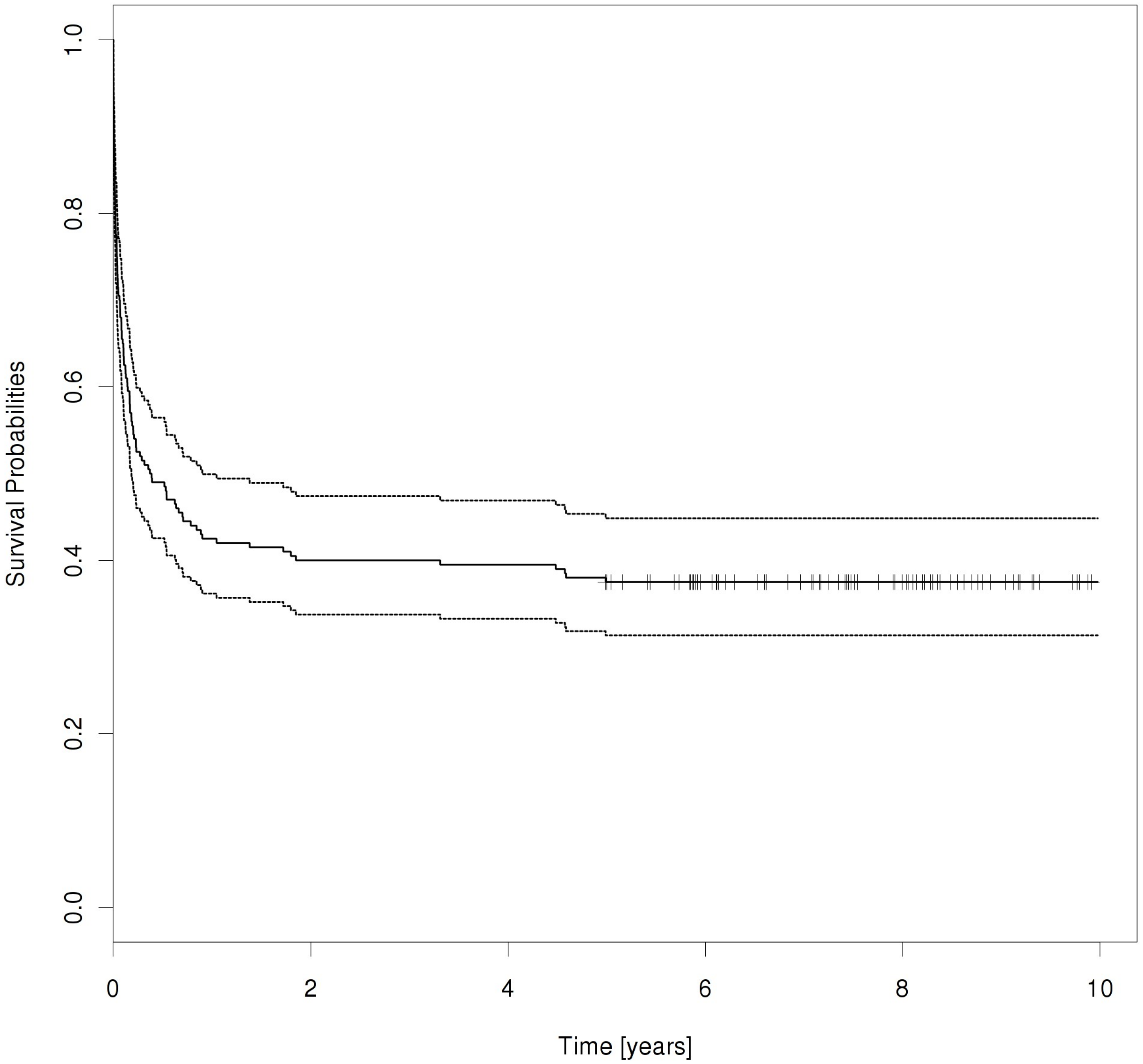

Figure 6-10: Product limit estimates and confidence intervals of the generated cure data (I6V). Vertical lines represent censored observations.

The values of the performance criteria $AUC(t)$, $BSC(t)$ and $R2(t)$ are shown in table 6-19. The survival models are presented in single columns and the time in the rows.

Year	**AUC**			**BSC**			**R2**		
	CCM	**ACM**	**COX**	**CCM**	**ACM**	**COX**	**CCM**	**ACM**	**COX**
Yr 1	**0.838**	0.793	0.774	**0.160**	0.178	0.191	**0.383**	0.287	0.223
Yr 2	**0.837**	0.828	0.809	**0.157**	0.160	0.174	**0.372**	0.355	0.277
Yr 3	**0.837**	0.828	0.809	**0.157**	0.159	0.174	**0.372**	0.366	0.277
Yr 4	**0.834**	0.826	0.807	**0.158**	0.159	0.174	0.363	**0.367**	0.276
Yr 5	**0.837**	0.833	0.817	**0.154**	**0.154**	0.166	0.364	**0.372**	0.293
Yr 6	0.823	**0.833**	0.817	0.164	**0.158**	0.169	0.322	**0.363**	0.283
Yr 7	0.823	**0.833**	0.817	0.165	**0.160**	0.171	0.318	**0.360**	0.274
Yr 8*	0.823	**0.834**	0.817	0.168	**0.161**	0.177	0.305	**0.356**	0.250
Yr 9*	0.823	**0.834**	0.817	0.168	**0.159**	0.178	0.305	**0.367**	0.243

Table 6-19: Performance of the survival models CCM, ACM and COX (single columns) obtained from 50 random splits of the I6V data. The prediction accuracy is measured by median values of $AUC(t)$, $BSC(t)$ and $R2(t)$ (global columns) at time $t=1,2,...,9$. The highest performance is characterized by the highest AUC and R2 as well as the lowest BSC values. They are shown in bold numbers.
***At late time points the measurements might become slightly inaccurate as late observation times are not included in some test samples due to the random selection procedure.**

The Cox mixture cure models outperform the accelerated failure time mixture cure models and the standard Cox models in the time span from year 1 to 5. Thereafter the accelerated failure time mixture cure models show a higher accuracy than the Cox mixture cure and the standard Cox models.

Results described in detail:

- CCM achieves higher $AUC(t)$ values (0.834 to 0.838) within the time period from year 1 to year 5 than ACM (0.793 to 0.833) and COX (0.774 to 0.817). The ACM (0.833 to 0.834) models outperform CCM (0.823) and COX (0.817) in the time span from year 6 to 9. Differences of $AUC(t)$ values increase over time between ACM and CCM (-0.045 to 0.011), remain constant between ACM and COX (between 0.016 and 0.019) and decrease between CCM and COX (year 1 to 9 from: 0.064 to 0.006).

- CCM (from 0.160 to 0.158) outperforms ACM (from 0.178 to 0.159) and COX (from 0.191 to 0.174) with respect to $BSC(t)$ values in the period from year 1 to 4. ACM and CCM reach the same level of performance in year 5 (0.154). ACM (between 0.161 and 0.158) surpasses CCM (between 0.168 and 0.164) and COX (between 0.178 and 0.169) in the time span from year 6 to 9. Differences of $BSC(t)$ values between ACM and CCM increase over time (year 1 to 9: from -0.018 to 0.009), remain constant between ACM and COX (between 0.011 and 0.019) and decrease between CCM and COX (0.031 to 0.005).

- Based on $R2(t)$ values the CCM model (between 0.372 and 0.383) achieves a higher level of performance than ACM (between 0.287 and 0.366) and COX (between 0.223 and 0.277) in the period between year 1 to 3. In the time span from year 4 to 9 ACM (between 0.356 and 0.372) outreaches CCM (between 0.305 and 0.364) and COX (between 0.243 and 0.293). The differences regarding prediction performance increase over time for ACM compared to CCM (year 1 to 9: from -0.096 to 0.062) and to COX (year 1 to 9: from 0.064 to 0.124) and decrease between CCM and COX (0.160 to 0.039).

Model complexity and running time of the algorithms

The runtime of the survival prediction procedure and the number of variables included in CCM, ACM and COX models are presented in this section. The number of variables incorporated in the survival models is shown in table 6-20.

CCM models include 7 variables (latency: 5 and incidence: 2), ACM 7 variables (5 and 2) and the standard Cox models 3 variables. A single survival prediction with the standard Cox model takes almost 50 seconds, with the CCM model 3 minutes and with the ACM model almost 4 hours.

The correct mixture cure models include three variables related to survival and three variables related to cure. Thus the correct Cox model incorporates six survival variables describing differences in survival time and cure plateaus.

no. of variables	**CCM**	**ACM**	**COX**	**CORR(CM)**	**CORR(COX)**
survival	5.00 [4.00, 8.00]	5.00 [3.00, 9.00]	3.00 [2.00, 5.00]	3	6
cure	2.00 [2.00, 4.00]	2.00 [2.00, 3.00]	-	3	-
running time [min]*	3.04 [2.94, 3.08]	229.03 [208.31, 255.91]	0.78 [0.77, 0.78]	-	-

Table 6-20: Number of variables (median plus 25 % and 75 % percentiles) in the Cox (COX) and the incidence and latency parts of mixture cure models (CCM, ACM). The running time (median as well as 25 % and 75 % percentiles) of the algorithms for one prognosis based on cure and Cox models is shown in the last row of the table. CORR(CM) and CORR(COX) represent the number of correct variables for the mixture cure models and the Cox regression model.
***Server specifications: Intel Core i7 CPU 2.80 GHz, 8 GB RAM.**

6.6.3.3.2 Analysis of the exploratory objectives

The exploratory objective of this chapter examines whether the survival prediction procedure (based on the three survival models) selects the correct survival model (variables 1, 2, 3 related to survival and 4, 5, 6 associated with cure) or the prediction models include noise variables.

Table 6-21 presents the frequency of correct variables and noise variables in the survival models. The frequency table is based on 50 prediction models. Therefore a gene can be included 50 times in each model part and 250 times in total (TOT). The correct variables (1-6) and the noise variables (OTHERS) are shown in the rows of the table. The cure (.INC) and survival parts (.LAT) of ACM and CCM and the results of the COX model are shown in the columns of table 6-21.

VARS	CCM.INC	CCM.LAT	ACM.INC	ACM.LAT	COX	TOT	TOT.INC	TOT.LAT	β	ζ
1	0	50	0	0	0	**50**	**0**	**50**	2	0
2	0	47	0	0	0	**47**	**0**	**47**	2	0
3	0	48	0	0	0	**48**	**0**	**48**	2	0
4	50	1	50	50	50	**201**	**100**	**101**	0	2
5	41	0	38	41	40	**160**	**79**	**81**	0	1.5
6	0	0	0	0	0	**0**	**0**	**0**	0	1
OTHERS	54	144	41	186	99	**524**	**95**	**429**	-	-

Table 6-21: Participation of single genes in survival prediction obtained from 50 random splits of the I6V data. The frequency matrix presents in how many cases the genes (rows) are involved in the prediction models. The analysis is based on COX models and mixture cure models (ACM and CCM with latency and incidence parts). VARS 1, 2 and 3 (rows) represent the variables associated with survival. VARS 4, 5 and 6 are the variables associated with cure. OTHERS describe the number of noise variables in the survival models. TOT.LAT and TOT.INC show in how many cases each variable is related to survival and to cure in total. The two columns on the right represent the true parameter values of the variables related to survival and cure.

Variables associated with survival and cure are frequently (frequencies > 38) or almost never (frequency 0 or 1) incorporated in the mixture cure and the standard Cox model. One correct cure variable (6) was not even included in a single model, which is due to the weak effect of the variable 6 on cure. The Cox mixture cure models included a high number of the correct variables. In accelerated failure time mixture cure models and the standard Cox model only the cure related variables are included.

The following findings can be reported:

- The survival related variables (1, 2, 3) were only incorporated in the survival part of the Cox model. Not a single variable was included in the latency part of ACM and the standard Cox model. The cure related variables (4, 5) were included in the cure part of CCM, in Cox regression models and in both model parts of ACM. One cure variable (6) was not included in the survival models at all.

- COX (90/300 = 30 %) and ACM (88/300 = 29 %) include a lower number of relevant genes in the correct model part compared to CCM (236/300 = 79 %). ACM includes 318 incorrect variables (78 % of all model variables), COX 99 (52 %) and CCM 199 (46 %).

- The incidence part of the survival models had a higher detection rate of relevant variables (60 %: ACM 59 % and CCM 61 %) than the survival parts of the prediction models (32 %: ACM 0 %, CCM 97 % and COX 0 %).

The survival prediction procedure coupled with the Cox mixture cure model clearly distinguishes between variables exclusively related to survival and others solely associated with cure. The survival related variables were not detected by the survival prediction procedure based on the accelerated failure time models and Cox regression models. Two cure related variables (4 and 5) were included in the latency part of ACM and in COX models.

6.6.4 Conclusions about the comparison study with regard to the generated datasets

Three generated datasets are investigated in this research. The first two datasets include three variables all related to latency and incidence. In the second dataset the effects on survival and cure are opposite. The third dataset includes three variables related to survival and other three variables related to cure only.

Survival prediction based on mixture cure models was performed by a new survival prediction procedure coupled with univariate selection of survival and cure effects. The prediction procedure from chapter 3.4 combined with the standard univariate selection approach was applied to predict survival based on the Cox regression models. The algorithms were stable and the variation of the results was unremarkable. The accuracy of the prediction procedure and the ability to detect the correct survival model were examined for the three datasets.

The following findings can be reported:

- Predictions based on the mixture cure models achieve a higher performance level than predictions with the standard Cox models in all three scenarios. The accelerated failure time mixture cure model outperforms the Cox mixture cure model if latency and incidence effects are associated with the same variables. The Cox cure model surpasses the accelerated failure time mixture cure model if survival and cure effects refer to the same variables.

- The Cox regression model shows lower accuracy in early and late terms. The runtime of the algorithms takes around one minute. If survival and cure effects are related to the same variables the detection rate of the signal variables is high and a moderate number of noise variables is included in the data. If survival and cure effects are associated with different variables the standard Cox model incorporates cure variables and the correct survival variables are not included. Cure effects seem to have a stronger impact than differences in survival time.

- The accelerated failure time mixture cure model reaches a high performance level in late terms and is globally the best model if the same variables provide cure and survival effects. If latency and cure effects are related to different variables the accelerated failure time mixture cure model achieves a lower performance, especially in early terms. The reason is the high number of cure related variables which are included in both model parts. Not one single survival related variable is included in the models. This leads to a lower accuracy (compared to the Cox mixture cure model) in early terms. The performance in late terms is high. The accelerated failure time mixture cure model selects a high number of correct variables in the first two scenarios, but includes a high number of noise variables especially in the survival part. The runtime of the algorithms is enormous. Therefore, a few hours have to be planned for a single prediction procedure, even if powerful computer servers are used.

- The Cox mixture cure model achieves a higher level of performance than the pure Cox model but does not reach the accuracy of the accelerated failure time mixture cure model. Especially in late terms the Cox mixture cure model provides a lower precision than the accelerated failure time mixture cure model. If survival and cure effects are associated with different variables, the Cox mixture cure model has a higher performance than the accelerated failure time mixture cure model and correctly identifies effects on latency and incidence with a low failure rate. The prediction procedure only takes some minutes.

In practice the new survival prediction procedure coupled with the Cox mixture cure model may be of more use since pure survival and cure effects can be identified. The runtime of the procedure is moderate and the performance level is high for early and late terms. The survival prediction procedure based on the accelerated failure time mixture cure model has high performance at late terms but does not select the correct effects on survival. The procedure has a high runtime. The standard Cox model is fast but has lower performance than mixture cure models.

6.7 Conclusions about the comparison study based on cure data

In chapter 6 the prediction accuracy of the standard Cox regression model, the accelerated failure time mixture cure model and the Cox mixture cure model, was compared using survival data of long-term survivors. The survival models were fitted to microarray data and to generated cure data.

A new procedure was introduced applying survival prediction based on mixture cure models in a computationally efficient way. Modified versions of the univariate selection approach and the supervised principal components method were used to fit the survival models. The feasibility of the new model procedure was investigated by the prediction accuracy of the survival estimates, the stability of the survival models and the runtime of the survival prediction procedure.

The main objectives of this chapter were examined based on two real datasets, the Netherlands breast cancer data (van't Veer et al., 2002) and the Erasmus Medical Center breast cancer data (Wang et al., 2005), and three generated cure datasets.

The mixture cure models demonstrated a clear benefit over the standard Cox regression model as the overall performance of the AFT mixture cure model and the Cox mixture cure model was higher regarding all five datasets. The Cox proportional hazards model outperformed the mixture cure models only with respect to early time points of the Netherlands breast cancer data. The standard Cox regression model revealed high performance for early but low performance for late time points.

The main advantages of Cox regression models are accurate results for short terms, when cure is not obvious yet, fast algorithms for model fitting and an easy, straightforward and established implementation of model approaches. This topic was discussed in chapter 5. The main deficiencies of standard Cox regression models on cure data are low prediction accuracy for long term evaluation and a not explicit consideration of cure.

The mixture cure models provide a high performance at late terms and even early terms for the Erasmus Medical Center breast cancer data and the generated cure datasets. In direct comparison the Cox mixture cure model reached a higher or equal performance level regarding early terms. The accelerated failure time mixture cure model achieves a higher performance level regarding late terms. For the microarray datasets the AFT mixture cure model and Cox mixture cure model perform almost equally and small benefits for the Cox mixture cure model on early and for AFT mixture cure models on late terms can be observed. For the first and second generated dataset with three variables related to latency and incidence the AFT mixture cure model achieves a higher performance level than the Cox mixture cure model, which even increases at late time points. For the third generated dataset with three variables related to survival and other three variables related to cure the CCM surpasses ACM results in early terms, but ACM reaches a higher level of performance in late terms. This is due to the fact that the new model approach applied to CCM selects pure effects on survival and cure. The model technique applied to ACM identifies variables in the survival part that are related to survival and cure. This means that in early times ACM does not reach CCM but at late terms ACM outperforms CCM. Although the ACM does not follow the intention of the model approach, it is the best model with regard to global model performance.

The cure and survival part of the AFT mixture cure model, the cure part of the Cox mixture cure model and the standard Cox regression model incorporated the same genes. Only the survival part of Cox mixture cure models included genes not related to the other model parts. Hence only the Cox mixture cure model allows the identification of genes associated with survival exclusively and enables a clear distinction between genes that are related to survival and features associated with cure. This was demonstrated for the Erasmus Medical Center breast cancer data.

The main benefits of Cox mixture cure models are that they provide high prediction performance for long- and short-term survival and they include genes that are related to cure and to survival exclusively. In this work the Cox mixture cure models have a relatively low run time marginally higher compared to the standard Cox model.

One deficiency of Cox mixture cure models becomes apparent in small data samples with high cure rates. The survival effects seem to be dominated by effects on cure, resulting in difficulties in identifying genes with a high impact on latency. A further issue arises from the fact that model selection based on Cox mixture cure models needs special model approaches accounting for two model parts. The model techniques may have to be applied to each of the two model parts separately, since a direct model selection on mixture cure models demands high computational costs.

The main benefits of AFT mixture cure models are the highest model performance regarding long-term survival, highly accurate predictions for data with effects on survival and cure regarding the same variables and a high detection rate of truly significant variables. Whilst in Cox regression models the covariate effects are associated with the hazard, the effects of predictors are related to event times for AFT models.

AFT mixture cure models do not allow a clear distinction between genes solely related to survival and others only associated with cure. Furthermore model building in AFT mixture cure models needs approaches considering cure and survival parts since a direct selection of genes in mixture cure models would be very time-consuming. The AFT mixture cure models include a high number of noise variables for the generated data and the run time for model selection and assessment with respect to AFT mixture cure models is very high.

Compared to some results in the literature (Perperoglou, 2006) the mixture cure models in this work have a higher prediction accuracy than the standard Cox regression models. In addition the differences between long- and short-term performance of standard Cox and cure models could be demonstrated.

A few facts regarding the runtime of the survival prediction procedure: The comparison study consists of nearly 60 million single regression models, univariate logistic regression models, Cox and AFT regression models and mixture cure models. Model selection regarding AFT mixture cure models are extremely time-consuming compared to Cox mixture cure (by a factor of up to 130) and standard Cox models (factor up to 750). In order to grant a practical use for survival prediction using cure models the model algorithms for the selection of AFT mixture models need to be much faster.

Chapter 7

7 Summary and Outlook

This thesis was dedicated to survival prediction from microarray data. The primary aim was to compare survival models, model approaches and tuning strategies with regard to the performance of the survival predictions.

This work is divided into two sections. In the first section model approaches, resampling techniques and tuning criteria were examined on the basis of Cox regression models. In the second part the performances of the standard Cox model and mixture cure models were assessed in case of long-term survivors. Additionally the effects of genes on survival and cure were determined.

Review of model approaches

Model approaches were investigated in some publications in a systematic way (Bovelstad et al., 2007, van Wieringen et al., 2009, Haibe-Kains et al. 2008, and Witten and Tibshirani, 2010). The first analysis in this work reevaluates model approaches and compares new model fitting techniques.

The results of the comparison study confirmed a high prediction accuracy of the partial least squares, the ridge regression (Bovelstad et al., 2007, and van Wieringen et al., 2009) and the supervised principal components approach (Witten and Tibshirani, 2010). For gene expression data the feature extraction techniques performed well. For datasets with high signal variables feature selection approaches like forward stepwise selection, the univariate selection and the supervised principal components approach demonstrated high prediction accuracy.

An additional result of the comparison study was that combined feature selection techniques (data are preselected using univariate selection) and extraction procedures achieve the highest performance. All model approaches exhibited a higher prediction performance in conjunction with the univariate selection technique. Hence hybrid methods may be a worthwhile research object.

Therefore combined procedures or shrinkage techniques, e.g. the supervised principal components, the ridge regression and the lasso are recommended to be used for survival prediction regarding gene expression data.

New issues of investigations

In the statistical literature the elastic net approach and survival ensembles were seldom compared to established techniques like partial least squares and the supervised principal components technique. In this work the elastic net approach was benchmarked against subset, variance and other shrinkage-based techniques. Elastic net and the boosting approach achieved average prediction performance. Survival trees revealed a low performance.

The **optimal resampling technique** was a further objective of this work. The prediction performances of survival models, validated by 5-, 10- and 20-fold cross-validation, were compared. This issue was examined by Subramanian and Simon (2011), who reported that leaving-one-out cross-validation exhibits high performance for survival models applied to high-signal data with a low sample size. 5- and 10-fold (or even 2-fold) cross-validation performed well for low-signal data and a low sample size (N ~ 40 to 80). 5-, 10- and leaving-one-out cross-validation achieved the same level of performance for data with a high sample size (N = 160) or with a low sample size and high-signal variables.

The results of this work and the suggestions from Subramanian and Simon (2011) are similar. Hence the following recommendations can be made:

- For microarray data (moderate or low-signal data): If the sample size is N < 100 5- and 10-fold cross-validation achieve a marginally higher performance than 20-fold cross-validation or leaving-one-out cross-validation. For data with sample sizes N > 100 differences regarding prediction performance between 5-, 10-fold and 20-fold cross-validation are negligible. Hence 5- and 10-fold cross-validation are recommended as they save computational time. However in this study 20-fold cross-validation marginally outperformed 5- and 10-fold cross-validation using the Rosenwald dataset (N = 240, learning sample N = 160).

- Using high-signal data: For sample sizes N < 100 20-fold cross-validation (or even leaving-one-out cross-validation based on Subramanian and Simon, 2011) achieved a high performance level. If the sample size is N >= 100 the prediction accuracy from 5-, 10- and 20-fold cross-validation achieves the same level of performance. 5- and 10-fold cross-validation are recommended to reduce the running time of the model fitting procedure.

The impact of the **tuning criterion** on model performance is often not taken into consideration. In this work the size of the survival models is selected by the integrated Brier score and the cross-validation partial log-likelihood method. The cross-validation partial log-likelihood metric is an established criterion to select the complexity of Cox models based on high-dimensional data. The cross-validation partial log-likelihood method can be applied to (semi)parametric models, the Brier score to parametric and non-parametric survival models (Porzelius et al., 2010). A key result of this thesis is that survival models tuned by the integrated Brier score outperformed models validated by the cross-validation partial log-likelihood criterion. Nevertheless it must be taken into account that two of the four model performance criteria used to assess the tuning criteria were derived from the Brier score.

Hence the integrated Brier score is an alternative to the cross-validated partial log-likelihood criterion for high-dimensional Cox regression models and can be taken into consideration for non-parametric models.

Survival models for cure data

The main objective of the second part of this work is the survival prediction from gene expression data in case of long-term survivors. The prediction performances of **mixture cure models** and standard Cox regression models as well as the feasibility of the new survival prediction procedure were the main research questions. Effects on survival and cure were further subjects of investigation in this thesis.

Mixture cure models achieved a much higher performance than Cox proportional hazards models. The standard Cox model reveals high to average performance in the short term. Mixture cure models reach a high level of overall performance. The accelerated failure time mixture cure models marginally surpass Cox mixture models in the long term. Cox mixture cure models minimally outreach the AFT mixture cure models in the short term.

Model selection for the pure Cox regression models can be performed by fast and established algorithms, but the survival models do not account for cure. Mixture cure models reach a higher performance level, especially in the long term. The Cox mixture cure model allows an interpretation of survival and cure effects. For situations with non-proportional hazards, cure effects and survival time may not be adequately modeled with pure Cox.

Based on the results of chapter 6 the Cox mixture cure model is recommended for mixed populations of cured and uncured subjects. The survival model shows a high performance regarding early and late terms, the algorithms for model selection run relatively fast and Cox mixture cure models allow an interpretation of effects on survival and cure. Nevertheless new approaches based on shrinkage techniques may improve the model performance. The new survival prediction procedure is feasible in conjunction with the Cox mixture cure model but reveals deficiencies in combination with the accelerated failure time model.

Features related to survival and cure

The exploratory objective of this work was the detection of **features showing a high impact on survival and cure**.

In the first part of this work the survival models incorporated a high number of genes since the data consist of a high proportion of low-signal genes. Models fitted from feature extraction approaches including a high number of the genes outperformed feature selection methods incorporating only a few genes. Feature selection approaches surpassed extraction techniques for data with high-signal variables due to the selection of a high number of correct variables. They only incorporated a low number of noise variables.

Based on Witten and Tibshirani (2010) and the results of this work the following model approaches are recommended: Feature extraction approaches achieve a high level of performance for microarray data. Model-based approaches like univariate selection, forward stepwise selection, the lasso or the approach by Goeman et al. (2005) are recommended if single genes with a high impact on survival are of primary interest.

In the second part of this work survival and cure effects were examined. AFT mixture cure models, standard Cox models and the incidence part of Cox mixture cure models included the same effects. Only the Cox mixture cure models incorporated genes solely related to incidence or latency. The application of Cox mixture cure models is recommended for samples including long-term survivors.

Further research issues

This thesis gives new insights into the model selection process like the performance of new model approaches, the validation process and the choice of the tuning criteria. In the last chapter a new model approach to fit mixture cure models was described and examined. Based on the results of chapter 5 and 6 new research questions are raised:

1) **Resampling techniques:** Cross-validation is the most common technique to tune high-dimensional models. Further investigations regarding the performance of resampling techniques should take the sample size of the training data and the data structure (high- and low-signal data, mixed populations) as well as model approaches into account.

2) **Model approaches for Cox models:** Hybrid methods combining feature selection and extraction approaches achieve a high level of performance. A well known approach is the supervised principal components technique combining the univariate selection and the principal components technique. Another one is the elastic net approach connecting ridge regression and the lasso. The implementation of new hybrid methods may be an interesting research question. The main disadvantage of hybrid methods is the costly complexity selection since two tuning parameters have to be chosen and the algorithms are very time-consuming.

3) **Model techniques for cure data:** Newer and faster algorithms for model fitting are needed for mixture cure models to become more attractive regarding survival prediction with gene expression data. Further research issues could be the implementation of shrinkage techniques (Liu et al., 2012), partial least squares approaches or the application of new survival models for cure data. Tsodikov and Garibotti implemented new survival models for cure data via the R package *nltm* (non-linear transformation models) including functions to fit bounded cumulative hazards models. The functions are not aligned for high-dimensional data, but Tsodikov worked on the R package *rpNLTM* that contains recursive partitioning and regression tree algorithms for non-linear transformation models.

4) **Further developments** to improve the prediction performance of survival models from gene expression data may combine molecular and clinical predictors. A systematic comparison of survival models from gene expression data and clinical covariates was applied by Bovelstad et al. (2009). Suitable model approaches have to give higher priority to clinical variables, which would otherwise “disappear” in the mass of molecular variables.

The implementation of powerful algorithms for model selection in cure models like the development of boosting or shrinkage-based approaches are further research projects. For Cox regression models the implementation of hybrid approaches are worthwhile research topics.

References

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. The Annals of Statistics, 6: 701-726.

Access Excellence @ the National Health Museum (2013). Microarrary Technology. Available at: *http://www.accessexcellence.org/RC/VL/GG/microArray.php.* (Accessed: March 20, 2013).

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19: 716-723.

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. The Annals of Statistics, 22: 1299-1327.

American Cancer Society Cancer Information Database. (2011). Oncogenes, tumor suppressor genes, and cancer. Available at:
www.cancer.org/acs/groups/cid/documents/webcontent/002550-pdf.pdf. (Accessed: July 10, 2012).

Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. PLOS Biology, 2: 511-522.

Bastien, P. (2004). PLS-Cox model: application to gene expression. COMPSTAT 2004 - Proceedings in Computational Statistics, 655-662.

Bastien, P., Vinzi, E. and Tenenhaus, M. (2005). PLS generalised linear regression. Computational Statistics and Data Analysis, 48: 17-46.

Benner, A., Zucknick, M., Hielscher, T., Ittrich, C. and Mansmann, U. (2010). High-dimensional Cox models: the choice of penalty as part of the model building process. Biometrical Journal, 52: 50-69.

Berkson, J. and Gage, R. P. (1952). Survival curves for cancer patients following treatment. Journal of the American Statistical Association, 47: 501-515.

Binder, H. and Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinformatics, 9: 10-19.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy (with discussion). Journal of the Royal Statistical Society, Series B, 11: 15-44.

Boulesteix, A. L., Strobl, C., Augustin, T. and Daumer, M. (2008). Evaluating microarray-based classifiers: an overview. Cancer Informatics, 6: 77-97.

Bovelstad, H. M., Nygard, S. and Borgan, O. (2009). Survival prediction from clinico-genomic models - a comparative study. BMC Bioinformatics, 10: 413.

Bovelstad, H. M., Nygard, S., Storvold, H. L., Aldrin, M., Borgan, O., Frigessi, A. and Lingjaerde, O. C. (2007). Predicting survival from microarray data - a comparative study. Bioinformatics, 23: 2080-2087.

Braga-Neto, U. M. and Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? Bioinformatics, 20: 374-380.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24: 123-140.

Breiman, L. (2001). Random forests. Machine Learning, 45: 5-32.

Bühlmann, P. (2004). Bagging, boosting and ensemble methods. Papers / Humboldt-Universität Berlin, Center for Applied Statistics and Economics (CASE), No.2004: 31.

Bullinger, L., Döhner, K., Bair, E., Fröhling, S., Schlenk, R. F., Tibshirani, R., Döhner, H. and Pollack, J. R. (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. New England Journal of Medicine, 350: 1605-1616.

Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., Saghatchian d'Assignies, M., Bergh, J., Lidereau, R., Ellis, P., Harris, A., Bogaerts, J., Therasse, P., Floore, A., Amakrane, M., Piette, F., Rutgers, E., Sotiriou, C., Cardoso, F. and Piccart, M. J. (2006). Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. Journal of the National Cancer Institute, 98: 1183-1192.

Cai, C., Zou, Y., Peng, Y. and Zhang, J. (2012). smcure: An R-package for estimating semiparametric mixture cure models. Computer Methods and Programs in Biomedicine, 108: 1255-1260.

Campbell, N. A. (1997). Biologie. Spektrum Akademischer Verlag, Heidelberg.

Chambless, L. E. and Diao, G. (2006). Estimation of time-dependent area under the ROC curve for long-term risk prediction. Statistics in Medicine, 25: 3474-3486.

Chambless, L. E., Cummiskey, C. P. and Cui, G. (2011). Several methods to assess improvement in risk prediction models: extension to survival analysis. Statistics in Medicine, 30: 22-38.

Clancy, S. and Brown, W. (2008). Translation: DNA to mRNA to protein. Nature Education 1(1).

Collett, D. (2003). Modelling survival data in medical research. 2nd edition, Chapman & Hall/CRC, London, UK.

Cook, N. R. and Ridker, P. M. (2009). The use and magnitude of reclassification measures for individual predictors of global cardiovascular risk. The Annals of Internal Medicine, 150: 795-802.

Corbiere F. and Joly P. (2007). A SAS macro for parametric and semiparametric mixture cure models. Computer Methods and Programs in Biomedicine, 85:173-180.

Cox, D. R. (1972). Regression models and life tables. Journal of the Royal Statistical Society, Series B, 34: 187-220.

Cox, D. R. and Snell, E. J. (1989). The analysis of binary data. 2nd edition, Chapman & Hall/CRC, London, UK.

Datta, S., Le-Rademacher, J. and Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. Biometrics, 63: 259-271.

De Castro, M., Cancho, V. G. and Rodrigues, J. (2010). A hands-on approach for fitting long-term survival models under the GAMLSS framework. Computer Methods and Programs in Biomedicine, 97: 168-177.

Debouck, C. and Goodfellow, P. N. (1999). DNA microarrays in drug discovery and development. Nature Genetics, 21: 48-50.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. The Annals of Statistics, 7: 1-26.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American Statistical Association, 78: 316-331.

Efron, B. and Tibshirani, R. (1998). An introduction to the bootstrap. Chapman and Hall, London, UK.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. The Annals of Statistics, 32: 407-499.

Engler, D. and Li, Y. (2009). Survival analysis with high-dimensional covariates: an application in microarray studies. Statistical Applications in Genetics and Molecular Biology, 8: 1-22.

Evers, L. and Messow, C. M. (2008). Sparse kernel methods for high-dimensional survival data. Bioinformatics, 24: 1632-1638.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. The Annals of Statistics, 36: 2605-2637.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96: 1348-1360.

Farewell, V. T. (1977). A model for a binary variable with time-censored observations. Biometrika, 64: 43-46.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. Biometrics, 38: 1041-1046.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. Information and Computation, 121: 256-285.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference, 148–156.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics and Data Analysis, 38: 367-378.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biology, 5: R80.

Gentleman, R. C., Carey, V. J., Huber, W., Irizarry, R. and Dudoit, S. (2005). Bioinformatics and computational biology solutions using R and Bioconductor. Series: Statistics for Biology and Health, Springer-Verlag, New York, USA.

Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. Biometrical Journal, 52: 70-84.

Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., Anninga, J. K. and van Houwelingen, H. C. (2005). Testing association of a pathway with survival using gene expression data. Bioinformatics, 21: 1950-1957.

Goeman, J. J., van de Geer, S. A., de Kort, F. and van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. Bioinformatics, 20: 93-99.

Gönen, M. and Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. Biometrika, 92: 965-970.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine, 18: 2529-2545.

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

Gupta, V., Cherkassky, A., Chatis, P., Joseph, R., Johnson, A. L., Broadbent, J., Erickson, T. and DiMeo, J. (2003). Directly labeled mRNA produces highly precise and unbiased differential gene expression data. Nucleic Acids Research, 31: e13.

Haibe-Kains, B. (2009). Identification and assessment of gene signatures in human breast cancer. PhD thesis. Universite Libre de Bruxelles, Belgium.

Haibe-Kains, B., Desmedt, C., Sotiriou, C. and Bontempi, G. (2008). A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? Bioinformatics, 24: 2200-2208.

Hansen, L. K. and Salamon, P. (1990). Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12: 993-1001.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine, 15: 361-387.

Hastie, T. and Tibshirani, R. (1990a). Exploring the nature of covariate effects in the proportional hazards model. Biometrics, 46: 1005-1016.

Hastie, T. and Tibshirani, R. (1990b). Generalized additive models. Chapman and Hall, London, UK.

Hastie, T. and Tibshirani, R. (2004). Efficient quadratic regularization for expression arrays. Biostatistics, 5: 329-340.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). Elements of statistical learning: data mining, inference, and prediction. 2nd edition, Springer, New York, USA.

Hastie, T., Tibshirani, R., Botstein, D. and Brown, P. (2001). Supervised harvesting of expression trees. Genome Biology, 2: 1-12.

Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D. and Brown, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biology, 1: RESEARCH0003.

Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P. and Botstein, D. (1999). Imputing missing data for gene expression arrays. Technical report. Stanford University, Division of Biostatistics.

Heagerty, P. J., Lumley, T. and Pepe, M. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics, 56: 337-344.

Hielscher, T., Zucknick, M., Werft, W. and Benner, A. (2010). On the prognostic value of survival models with application to gene expression signatures. Statistics in Medicine, 29: 818-829.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12: 55-67.

Hofner, B., Mayer, A., Robinzonov, N. and Schmid, M. (2012). Model-based boosting in R - a hands-on tutorial using the R package *mboost*. Computational Statistics.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. and van der Laan, M. (2006). Survival ensembles. Biostatistics, 7: 355-373.

Hothorn, T., Lausen, B., Benner, A. and Radespiel-Troeger, M. (2004). Bagging survival trees. Statistics in Medicine, 23: 77-91.

Huang, J. and Harrington, D. (2002). Penalized partial likelihood regression for right censored data with bootstrap selection of the penalty parameter. Biometrics, 58: 781-791.

Jain, K. K. (2001). Biochips for Gene Spotting. Science, 294: 621-623.

Kaderali, L., Zander, T., Faigle, U., Wolf, J., Schultze, J. L. and Schrade, R. (2006). CASPAR: a hierarchical bayesian approach to predict survival times in cancer from gene expression data. Bioinformatics, 22: 1495-1502.

Kalbfleisch, J. D. (1978). Non-parameteric bayesian analysis of survival time data. Journal of the Royal Statistical Society, Series B, 40: 214-221.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. Biometrika, 60: 267-278.

Kalbfleisch, J. D. and Prentice, R. L. (2002). The statistical analysis of failure time data. 2nd edition. Wiley. New York, USA.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53: 457-481.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90: 773-795.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. Journal of Mathematical Analysis and Applications, 33: 82-95.

Klein, J. P. (1991). Small sample moments of some estimators of the variance of the Kaplan-Meier and Nelson-Aalen estimators. Scandinavian Journal of Statistics, 18: 333-340.

Knippers, R. (2006). Molekulare Genetik. 9. Aufl., Thieme, Stuttgart.

Lawless, J. F. (1982). Statistical models and methods for lifetime data. Wiley, New York, USA.

Leamer, E. E. (1978). Specification searches. Wiley, New York, USA.

Li, C. S. and Taylor, J. M. (2002). A semi-parametric accelerated failure time cure model. Statistics in Medicine, 21: 3235-3247.

Li, H. (2006). Censored data regression in high-dimension and low-sample size settings for genomic applications. Working Paper. University of Pennsylvania, Collection of Biostatistics Research Archive.

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

Li, H. and Luan, Y. (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. Pacific Symposium of Biocomputing, 8: 65-76.

Li, H. and Luan, Y. (2005). Boosting proportional hazards models using smoothing splines with applications to high-dimensional microarray data. Bioinformatics, 21: 2403-2409.

Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. Bioinformatics, 20: 3406-3412.

Liu, H., Li, J. and Wong, L. (2005). Use of extreme patient samples for outcome prediction from gene expression data. Bioinformatics, 21: 3377-3384.

Liu, X., Peng, Y., Tu, D. and Liang, H. (2012). Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. Statistics in Medicine, 31: 2882-2891.

Liu, Z., Chen, D., Tan, M., Jiang, F. and Gartenhaus, R. B. (2010). Kernel based methods for accelerated failure time model with ultra-high dimensional data. BMC Bioinformatics, 11: 606.

Magee, L. (1990). R2 measures based on Wald and likelihood ratio joint significance tests. The American Statistician, 44: 250-253.

Malone J. H. and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biology, 9: 34.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports, 50: 163-170.

Meinshausen, N. (2007). Relaxed lasso. Computational Statistics and Data Analysis, 52: 374-393.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. The Annals of Statistics, 34: 1436-1462.

Michiels, S., Koscielny, S. and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet, 365: 488-492.

Molinaro, A. M., Simon, R. and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21: 3301-3307.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. Biometrika, 78: 691-692.

National Cancer Institute. (2009). Understanding Cancer Series. Available at: *http://www.cancer.gov/cancertopics/understandingcancer/cancer* (Accessed: July 30, 2012).

Nguyen, D. V. and Rocke, D. M. (2002). Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics, 18: 1625-1632.

Nobel Media AB. (2012). "DNA-RNA-Protein". Available at: *http://www.nobelprize.org/educational/medicine/dna/index.html.*, Copyright © Nobel Media AB, (Accessed: July 10, 2012).

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society, Series B, 69: 659-677.

Park, P. J., Tian, L. and Kohane, I. S. (2002). Linking expression data with patient survival times using partial least squares. Bioinformatics, 18: S120-127.

Pawitan, Y., Bjöhle, J., Wedren, S., Humphreys, K., Skoog, L., Huang, F., Amler, L., Shaw, P., Hall, P. and Bergh, J. (2004). Gene expression profiling for prognosis using Cox regression. Statistics in Medicine, 23: 1767-1780.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 2: 559-572.

Pencina, M. J. and D'Agostino, R. B. (2004). Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in Medicine, 23: 2109-2123.

Peng, Y. (2003). Fitting semiparametric cure models. Computational Statistics and Data Analysis, 41: 481-490.

Peng, Y. and Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. Biometrics, 56: 237-243.

Perperoglou, A. (2006). Modelling long term survival with non-proportional hazards. Doctoral thesis. Leiden University, Netherlands.

Perperoglou, A., Keramopoullos, A. and van Houwelingen, H. C. (2007). Approaches in modelling long-term survival: an application to breast cancer. Statistics in Medicine, 26: 2666-2685.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. Journal of the Royal Statistical Society, Series A, 135: 185-207.

Porzelius, C., Schumacher, M. and Binder, H. (2010). A general, prediction error-based criterion for selecting model complexity for high-dimensional survival models. Statistics in Medicine, 29: 830-838.

R Core Team. (2012). R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Rama, R., Swaminathan, R. and Venkatesan, P. (2010). Cure models for estimating hospital-based breast cancer survival. Asian Pacific Journal of Cancer Prevention, 11: 387-391.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M, Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B. and Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-Cell lymphoma. The New England Journal of Medicine, 346, 1937-1947.

Schapire, R. E. (1990). The strength of weak learnability. Machine Learning, 5: 197-227.

Schapire, R. E. (2003). The boosting approach to machine learning an overview. Nonlinear Estimation and Classification. Springer.

Schemper, M. and Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression. Biometrics, 56: 249-255.

Schmid, M. and Hothorn, T. (2008). Flexible boosting of accelerated failure time models. BMC Bioinformatics, 9: 269.

Schober D. (2002). Microarrays, Genexpressionsanalyse und Bioinformatik. BIOspektrum 3/02, 8. Jahrgang.

Schumacher, M., Binder, H. and Gerds, T. (2007). Assessment of survival prediction models based on microarray data. Bioinformatics, 23: 1768-1774.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6: 461-464.

Segal, M. R. (2006). Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. Biostatistics, 7: 268:285.

Shao, J. (1993). Linear model selection by cross-validation. Journal of the American Statistical Association, 88: 486-494.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of Statistical Software, 39: 1-13.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lonning, P. E. and Borresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences, 98: 10869-10874.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich R., Geisler, S., Demeter, J., Perou, C. M., Lonning, P. E., Brown, P. O., Borresen-Dale, A. L. and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences, 100: 8418-8423.

Sposto, R. (2002). Cure model analysis in cancer: an application to data from the Children’s Cancer Group. Statistics in Medicine, 21: 293-312.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software, 23: 1-46.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J. and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology, 21: 128-138.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Series B, 36: 111-147.

Subramanian, J. and Simon, R. (2011). An evaluation of resampling methods for assessment of survival risk prediction in high-dimensional settings. Statistics in Medicine, 30: 642-653.

Sy, J. P. and Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. Biometrics, 56: 227-236.

Taylor, J. M. G. (1995). Semi-parametric estimation in failure time mixture models. Biometrics, 51: 899-907.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58: 267-288.

Tibshirani, R. (1997): The lasso method for variable selection in the Cox model. Statistics in Medicine, 16: 385-395.

Tibshirani, R. (2009). Univariate shrinkage in the Cox model for high dimensional data. Statistical Applications in Genetics and Molecular Biology, 8: 1-18.

Tsodikov, A. (2002). Semi-parametric models of long- and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage. Statistics in Medicine, 21: 895-920.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine, 30: 1105-1117.

Van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskouil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. The New England Journal of Medicine, 347: 1999-2009.

Van der Laan, M. J. and Robins, J. M. (2003). Unified methods for censored longitudinal data and causality. Springer, New York, USA.

Van Houwelingen, H. C. (2007). High-dimensional model building for survival data: How to measure prognostic performance and to regulate the complexity. Presentation. Munich, Germany.

Van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van’t Veer, L. J. and Wessels, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. Statistics in Medicine, 25: 3201-3216.

Van Wieringen, W. N., Kun, D., Hampel, R. and Boulesteix, A. L. (2009). Survival prediction using gene expression data: a review and comparison. Computational Statistics and Data Analysis, 53: 1590-1603.

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

Venables, W. N. and Ripley, B. D. (2002). Modern applied statistics with S. 4th edition, Springer, New York, USA.

Verweij, P. J. and van Houwelingen, H. C. (1993). Cross-validation in survival analysis. Statistics in Medicine, 12: 2305-2314.

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

Witten, D. M. and Tibshirani, R. (2010). Survival analysis with high-dimensional covariates. Statistical Methods in Medical Research, 19: 29-51.

Wold, H. (1966). Estimation of principal components and related methods by iterative least squares. In: Krishnaiah, P. R., editor. Multivariate Analysis. Academic Press, New York, USA, 391-420.

Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5: 241-259.

Xiao, Y., Hua, J. and Dougherty, E. R. (2007). Quantification of the impact of feature selection on the variance of cross-validation error estimation. EURASIP Journal on Bioinformatics and Systems Biology, 16354-16364.

Xu, J., Yang, Y. and Ott, J. (2005). Survival analysis of microarray expression data by transformation models. Computational Biology and Chemistry, 29: 91-94.

Yakovlev, A. Y. and Tsodikov, A. D. (1996). Stochastic models of tumor latency and their biostatistical applications. World Scientific. Singapore.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for Cox’s proportional hazards model. Biometrika, 94: 691-703.

Zhang, J. and Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. Statistics in Medicine, 26: 3157-3171.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101: 1418-1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B, 67: 301-320.

Appendix A

A.1 Additional analyses based on generated data

Prediction performance of nine model approaches measured by median IAUC, IBSC, deviance and R2 results. The tuning parameter was selected by 5-fold CV (table A-1) and 20-fold CV (table A-2):

	CVPL				IBSC			
	IAUC	IBSC	DEV	R2B	IAUC	IBSC	DEV	R2B
UPV	**0.819**	**0.096**	**-41.87**	**0.435**	**0.828**	**0.092**	**-44.52**	**0.455**
FSS	**0.826**	**0.093**	**-41.15**	**0.451**	**0.835**	**0.088**	**-45.99**	**0.481**
PCR	0.667	0.151	-5.00	0.123	0.696	0.142	-8.43	0.160
SPC	**0.839**	**0.086**	**-53.23**	**0.492**	**0.841**	**0.085**	**-53.93**	**0.498**
LAS	0.811	0.100	-35.24	0.396	0.812	0.100	-32.14	0.406
RID	0.684	0.146	-1.42	0.140	0.678	0.148	-0.01	0.126
NET	0.779	0.110	-31.43	0.349	0.725	0.128	-6.60	0.233
RSF	0.616	0.159	-10.56	0.053	0.735	0.137	-2.01	0.189
GBS	0.803	0.105	-31.08	0.386	0.809	0.100	-34.17	0.416

Table A-1: Median performance of nine model approaches (rows) obtained from 50 random splits of the generated data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 5-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

	CVPL				IBSC			
	IAUC	IBSC	DEV	R2B	IAUC	IBSC	DEV	R2B
UPV	**0.828**	**0.093**	**-46.31**	**0.460**	**0.836**	**0.090**	**-52.12**	**0.483**
FSS	**0.828**	**0.093**	**-48.26**	**0.462**	**0.836**	**0.089**	**-52.61**	**0.484**
PCR	0.680	0.149	-6.35	0.126	0.699	0.145	-8.74	0.167
SPC	**0.841**	**0.087**	**-54.95**	**0.491**	**0.842**	**0.086**	**-55.98**	**0.496**
LAS	0.815	0.100	-35.74	0.430	0.817	0.098	-32.53	0.436
RID	0.683	0.148	-1.44	0.136	0.680	0.149	-0.01	0.125
NET	0.787	0.109	-6.97	0.370	0.734	0.127	-4.96	0.235
RSF	0.624	0.160	-4.88	0.057	0.742	0.135	-3.47	0.194
GBS	0.811	0.100	-37.96	0.419	0.823	0.095	-41.37	0.455

Table A-2: Median performance of nine model approaches (rows) obtained from 50 random splits of the generated data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 20-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Tuning parameter values for the model approaches

Table A-3, A-4 and A-5 summarize the tuning parameter values of the ten model approaches from 50 survival models (using the full dataset). With regard to the UPV and FSS approach the tuning parameter represents the number of features included in the prediction models. Using the PCR and PLS approaches the number of components, regarding LAS and RID the amount of parameter shrinkage are validated. With respect to RSF the number of survival trees and regarding the GBS the boosting step width is tuned.

The supervised principal components regression and the elastic net approach include two tuning parameters. The number of preselected features and the number of components are selected regarding SPC as well as the amount of L1 and L2 shrinkage with respect to NET.

	CVPL					**IBSC**				
	Q0	**Q1**	**Q2**	**Q3**	**Q4**	**Q0**	**Q1**	**Q2**	**Q3**	**Q4**
UPV	4	7.25	9	10	14	4	8	10	11	15
FSS	6	7	8	10	12	6	8	9	10	12
PCR	1	4	7	10	20	4	14	18	20	20
SPC-C	1	1	1	2	10	1	1	1	1	12
SPC-P	10	10	10	10	10	5	10	10	10	50
LAS	12	15	16	18	21	11	16.25	19	22.75	35
RID	1000	3000	5000	8750	1000000	1000	1000000	1000000	1000000	1000000
NET-L2	1	1	5	11	50	1	100	1000	1000000	1000000
NET-L1	10	15	15	15	20	15	20	20	25	30
RSF	10	10	10	10	25	10	25	50	75	100
GBS	0.01	0.03	0.03	0.04	0.08	0.01	0.04	0.055	0.09	0.1

Table A-3: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of nine model approaches (rows) obtained from 50 random data splits of the GEN data. The survival models are validated by 10-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

	CVPL					IBSC				
	Q0	Q1	Q2	Q3	Q4	Q0	Q1	Q2	Q3	Q4
UPV	2	8	9	10	14	5	9	10	10	13
FSS	3	7.25	8	10	12	6	9	10	12	12
PCR	1	3	6.5	11	17	3	14.25	18	20	20
SPC-C	1	1	1	2	10	1	1	1	1	10
SPC-P	10	10	10	10	10	10	10	10	10	10
LAS	10	14	15.5	17	20	13	16	18	20.75	25
RID	1000	3000	6000	20000	1000000	1000	1000000	1000000	1000000	1000000
NET-L2	1	1	3	11	50	5	100	1000	1000000	1000000
NET-L1	10	15	15	15	20	15	20	20	25	25
RSF	10	10	10	10	50	10	25	50	75	100
GBS	0.01	0.02	0.03	0.04	0.09	0.02	0.04	0.045	0.07	0.1

Table A-4: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of nine model approaches (rows) obtained from 50 random data splits of the GEN data. The survival models are validated by 5-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

Tuning parameters for 20-fold CV:

	CVPL					IBSC				
	Q0	Q1	Q2	Q3	Q4	Q0	Q1	Q2	Q3	Q4
UPV	4	8	9	10	15	5	8	9	10	15
FSS	5	7	8	9.75	12	6	8.25	10	12	12
PCR	1	3	5	9.25	19	1	8.25	17	20	20
SPC-C	1	1	1	2	10	1	1	2	5	10
SPC-P	10	10	10	10	10	5	10	10	10	50
LAS	13	15.25	18	19	23	14	18	19	23	34
RID	2000	4250	7000	20000	1000000	1000	1000000	1000000	1000000	1000000
NET-L2	1	1	5	11	50	3	100	1000	1000000	1000000
NET-L1	15	15	15	20	20	10	20	22.5	25	30
RSF	10	10	10	10	25	10	25	50	75	100
GBS	0.01	0.02	0.03	0.04	0.07	0.01	0.04	0.05	0.07	0.1

Table A-5: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of nine model approaches (rows) obtained from 50 random data splits of the GEN data. The survival models are validated by 20-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

A.2 Additional analyses using the AML data

Prediction performance of eight model approaches measured by median IAUC, IBSC, deviance and R2 values. The tuning parameter was selected by 5-fold CV (table A-6) and 20-fold CV (table A-7):

	CVPL				IBSC			
	IAUC	IBSC	DEV	R2B	IAUC	IBSC	DEV	R2B
UPV	0.561	0.207	6.32	0.037	0.593	0.196	5.46	0.078
FSS	0.537	0.216	6.69	0.008	0.545	0.209	6.89	0.023
PCR	0.580	0.203	**-0.33**	0.039	**0.618**	**0.180**	**-2.75**	**0.156**
SPC	0.579	0.201	1.62	0.070	**0.618**	**0.182**	**-0.80**	**0.137**
LAS	**0.611**	**0.194**	**-2.65**	**0.086**	0.609	0.193	**-2.49**	0.082
RID	**0.587**	**0.190**	**-2.33**	**0.094**	0.593	0.190	-0.34	0.094
NET	0.581	**0.191**	50.48	**0.090**	0.584	0.194	13.56	0.098
GBS	**0.583**	**0.194**	4.67	0.084	**0.625**	**0.186**	6.17	**0.113**

Table A-6: Median performance of eight model approaches (rows) obtained from 50 random splits of the AML data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 5-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

	CVPL				IBSC			
	IAUC	IBSC	DEV	R2B	IAUC	IBSC	DEV	R2B
UPV	0.565	0.205	7.52	0.045	0.594	0.191	3.73	0.099
FSS	0.569	0.205	5.20	0.039	0.593	0.196	5.05	0.055
PCR	0.558	0.200	**-0.02**	0.047	0.600	**0.181**	**-1.79**	**0.137**
SPC	**0.575**	0.198	3.35	0.059	**0.603**	**0.182**	0.93	**0.145**
LAS	**0.600**	**0.190**	**-1.43**	**0.075**	**0.633**	**0.181**	**-3.04**	**0.122**
RID	0.571	**0.193**	**-1.91**	**0.082**	0.577	0.188	**-0.14**	0.085
NET	0.551	**0.195**	74.69	0.068	0.575	0.195	23.83	0.065
GBS	**0.590**	**0.195**	7.91	**0.070**	**0.618**	0.188	9.00	0.103

Table A-7: Median performance of eight model approaches (rows) obtained from 50 random splits of the AML data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 20-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Tuning parameter values for the model approaches

Table A-8, A-9 and A-10 summarize the tuning parameter values of the eight model approaches (of the full dataset).

	CVPL					IBSC				
	Q0	Q1	Q2	Q3	Q4	Q0	Q1	Q2	Q3	Q4
UPV	1	1	2	4	13	1	4	7	9.75	15
FSS	1	1	1	1	3	1	1	2	3	5
PCR	1	2	3	7	20	1	2	15	17	20
SPC-C	1	1	1.5	3.75	17	1	2	6	14	20
SPC-P	1	6.25	550	1000	1000	5	50	100	1000	1000
LAS	9	12	13	15	28	5	12	15	23.75	31
RID	3000	6250	9000	20000	70000	2000	8250	250000	1000000	1000000
NET-L2	1	1	11	100	10000	1	3	50	10000	1000000
NET-L1	10	10	10	10	30	10	10	15	15	30
GBS	0.01	0.01	0.01	0.0175	0.05	0.01	0.02	0.035	0.05	0.1

Table A-8: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of eight model approaches (rows) obtained from 50 random data splits of the AML data. The survival models are validated by 5-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

	CVPL					IBSC				
	Q0	Q1	Q2	Q3	Q4	Q0	Q1	Q2	Q3	Q4
UPV	1	1	2	5	14	1	2	5.5	8	15
FSS	1	1	1	2	3	1	1	2	2.75	5
PCR	1	1.25	3	12.75	18	1	3	12.5	18	20
SPC-C	1	1	2	5	16	1	2	4.5	9.75	20
SPC-P	1	6.25	75	1000	1000	2	50	100	1000	1000
LAS	8	12	13	15	29	8	13	16	26	30
RID	1000	6000	8000	17500	1000000	1000	7750	55000	1000000	1000000
NET-L2	1	1	3	50	1000	1	3	1000	10000	1000000
NET-L1	10	10	10	15	20	10	10	15	20	30
GBS	0.01	0.01	0.01	0.02	0.05	0.01	0.02	0.04	0.06	0.1

Table A-9: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of eight model approaches (rows) obtained from 50 random data splits of the AML data. The survival models are validated by 10-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

	CVPL					IBSC				
	Q0	Q1	Q2	Q3	Q4	Q0	Q1	Q2	Q3	Q4
UPV	1	1	3	5	12	1	3	6.5	9.75	15
FSS	1	1	1.5	2	3	1	1	2	3	5
PCR	1	3	4	12.75	19	1	3.25	14	18	20
SPC-C	1	1	2	5.75	15	1	2.25	7	11.75	20
SPC-P	1	5	30	100	1000	1	20	100	1000	1000
LAS	6	10.25	13	16.75	27	6	11	14	21.25	33
RID	3000	6000	7500	10000	1000000	1000	8000	550000	1000000	1000000
NET-L2	1	1	11	100	10000	1	1	30.5	775000	1000000
NET-L1	10	10	10	10	15	10	10	10	15	35
GBS	0.01	0.01	0.01	0.02	0.06	0.01	0.02	0.045	0.07	0.1

Table A-10: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of eight model approaches (rows) obtained from 50 random data splits of the AML data. The survival models are validated by 20-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

A.3 Additional analyses using the DLBCL data

Prediction performance of eight approaches measured by median IAUC, IBSC, deviance and R2 values. The tuning parameter was selected by 5-fold CV (table A-11) and 20-fold CV (table A-12):

	CVPL				**IBSC**			
	IAUC	**IBSC**	**DEV**	**R2B**	**IAUC**	**IBSC**	**DEV**	**R2B**
UPV	0.545	0.220	7.84	0.012	0.550	**0.219**	7.58	0.010
FSS	0.555	**0.218**	3.21	0.016	0.555	**0.219**	4.91	0.016
PCR	0.566	0.222	**-0.08**	0.013	0.535	0.222	**-0.19**	0.018
SPC	0.548	0.223	7.14	0.010	0.549	**0.219**	0.41	**0.030**
LAS	**0.582**	0.221	**-0.75**	0.014	**0.592**	0.220	**-0.21**	0.005
RID	**0.588**	**0.217**	**-2.42**	**0.027**	**0.574**	**0.219**	**-0.15**	**0.033**
NET	**0.574**	**0.216**	19.90	**0.029**	0.554	0.220	88.20	0.015
GBS	0.573	0.219	8.60	**0.038**	**0.572**	**0.219**	8.20	**0.038**

Table A-11: Median performance of eight model approaches (rows) obtained from 50 random splits of the DLBCL data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 5-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

	CVPL				**IBSC**			
	IAUC	**IBSC**	**DEV**	**R2B**	**IAUC**	**IBSC**	**DEV**	**R2B**
UPV	0.566	0.214	9.44	0.024	0.573	0.210	0.57	0.041
FSS	**0.587**	**0.208**	**-1.93**	**0.053**	**0.587**	**0.208**	**-0.87**	**0.064**
PCR	0.556	0.220	0.04	0.009	0.546	0.216	-0.51	0.020
SPC	0.548	0.219	5.97	0.021	0.585	**0.205**	**-0.58**	**0.064**
LAS	0.579	0.216	**-1.27**	0.020	**0.587**	**0.209**	**-0.99**	**0.063**
RID	**0.594**	**0.212**	**-3.25**	**0.041**	**0.597**	0.212	-0.34	0.043
NET	**0.585**	0.214	34.15	0.035	0.564	0.220	75.99	0.016
GBS	0.583	**0.213**	6.19	**0.049**	0.582	0.211	8.02	0.048

Table A-12: Median performance of eight model approaches (rows) obtained from 50 random splits of the DLBCL data. The prediction accuracy is assessed with IAUC, IBSC, DEV and R2 (columns). The survival models are validated by 20-fold CV and either the CVPL (left four columns) or IBSC tuning criterion (right four columns). High prediction performance (top three approaches of each column are shown in bold numbers) is characterized by high IAUC and R2 as well as low IBSC and DEV values.

Tuning parameter values for the model approaches

Table A-13, A-14 and A-15 summarize the tuning parameter values of the eight model approaches (using the full dataset).

	CVPL					IBSC				
	Q0	Q1	Q2	Q3	Q4	Q0	Q1	Q2	Q3	Q4
UPV	1	1	2	2	15	1	1	1	2	15
FSS	1	1	1	1	4	1	1	1	1	3
PCR	1	6	8	11.75	20	1	1	2	5	20
SPC-C	1	1	1	2	20	1	1	1	1	9
SPC-P	1	2	50	1000	1000	1	5	1000	1000	1000
LAS	13	20	24.5	33.5	39	15	32	35	37	40
RID	2000	6000	8500	20000	1000000	1000	5250	130000	1000000	1000000
NET-L2	1	7	100	1000	100000	1	1	1	1	1000000
NET-L1	10	10	15	20	40	10	35	35	40	40
GBS	0.01	0.01	0.01	0.01	0.08	0.01	0.01	0.01	0.01	0.1

Table A-13: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of eight model approaches (rows) obtained from 50 random data splits of the DLBCL data. The survival models are validated by 5-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

	CVPL					IBSC				
	Q0	Q1	Q2	Q3	Q4	Q0	Q1	Q2	Q3	Q4
UPV	1	1	2	4	15	1	1	1	1.75	13
FSS	1	1	1	2	5	1	1	1	1	4
PCR	1	6	8	10.5	20	1	1	3	7	20
SPC-C	1	1	1	2	9	1	1	1	1	7
SPC-P	1	2	10	1000	1000	1	1	100	1000	1000
LAS	13	21	22	27	39	24	34	37.5	39	40
RID	3000	6000	8500	10000	1000000	1000	9000	1000000	1000000	1000000
NET-L2	1	7	100	1000	10000	1	1	1	1	1000000
NET-L1	10	10	15	20	35	20	35	40	40	40
GBS	0.01	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.01	0.1

Table A-14: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of eight model approaches (rows) obtained from 50 random data splits of the DLBCL data. The survival models are validated by 10-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

	CVPL					IBSC				
	Q0	**Q1**	**Q2**	**Q3**	**Q4**	**Q0**	**Q1**	**Q2**	**Q3**	**Q4**
UPV	1	1	2	5	13	1	1	1	2	8
FSS	1	1	1	1	5	1	1	1	1	4
PCR	1	1	7	10	20	1	1	5	7	20
SPC-C	1	1	1	2.75	10	1	1	1.5	2	20
SPC-P	1	1.25	30	1000	1000	1	2	100	1000	1000
LAS	15	20.25	23.5	29.75	40	19	30.25	34	38	40
RID	4000	6250	9000	17500	1000000	1000	3250	35000	1000000	1000000
NET-L2	1	11	100	1000	10000	1	1	1	11	1000000
NET-L1	10	10	15	20	40	10	30	35	40	40
GBS	0.01	0.01	0.01	0.01	0.08	0.01	0.01	0.01	0.01	0.1

Table A-15: Tuning parameter values (minimum as well as 25 % percentiles, median, 75 % percentiles, maximum) of eight model approaches (rows) obtained from 50 random data splits of the DLBCL data. The survival models are validated by 20-fold CV and either the CVPL (left column) or IBSC tuning criterion (right column).

Appendix B

B.1 Summary

This thesis is dedicated to survival prediction from gene expression data. The key topics of the text are the assessment of model approaches and tuning strategies for model fitting as well as the prediction performance of survival models for populations of mixed frail and immune patients. A further issue is the impact of single genes on the prediction models.

This work is divided in two parts. In the first section the ten most popular approaches for model fitting based on high-dimensional data are introduced. These can be classified into parameter shrinkage, subset selection, ensemble methods and techniques based on derived input directions. Differences and similarities between the model approaches are discussed from various theoretical perspectives e.g. if model selection is related to the outcome variable, if interdependences between features are considered or if single or aggregated genes are selected. The strengths and weaknesses of the techniques are discussed and systematically evaluated.

The secondary topics of this work are tuning strategies, which significantly affect the prediction performance of survival models from microarray data. These consist of many components like the resampling technique, the choice of the tuning parameter and the tuning criterion in order to select the complexity of the survival models.

This work examines resampling techniques and tuning criteria. A comparison between the prediction performances of survival models validated by 5-, 10- and 20-fold cross-validation is made in this thesis. 5- and 10-fold cross-validation is a commonly used resampling technique for model fitting. A lower number of cross-validation samples (e.g. leaving-one-out cross-validation) leads to a higher sample bias and lower performance. This work examines the impact of 20-fold cross-validation on the prediction performance of survival models.

This work investigates the influence of the tuning criterion regarding the accuracy of the prediction model. Survival models tuned by the cross-validation partial log-likelihood criterion are compared to models validated by the integrated Brier score. The assets and deficiencies of the tuning criteria, e.g. the areas of application, are discussed and the prediction performances of the models are examined.

The impact of single genes on survival is an experimental research question of this work. Heuristic algorithms are used to identify significant features selected by the model approaches. Genes related to survival are compared between the model approaches.

The model techniques, tuning strategies and survival effects are compared regarding a population of frail patients.

The second part of this work refers to the comparison of survival models applied to a population of mixed susceptible and insusceptible patients. The class of semiparametric mixture cure models is introduced. The Cox and AFT mixture cure model including a latency and an incidence model part are described.

A new survival prediction procedure is presented that takes both components of the survival models into account. Genes related to survival and cure are selected separately. The mixture cure models are fitted by procedures that are based on single effects on latency and incidence. Benefits and deficiencies of mixture models and the standard Cox proportional hazards model are presented in this work. The prediction accuracy of the pure Cox model, the mixture cure model and the AFT mixture cure model are compared. The impact of genes on survival and cure are discussed in the text.

Further issues of this thesis are survival prediction procedures to tune and fit the survival models, criteria to measure the prediction performance of survival models and steps to acquire gene expression data from tissue samples.

The main objectives of this work are summarized in primary, secondary and exploratory hypotheses. The research questions are examined using four cancer and four generated datasets. The simulated data contain high-signal variables. Some recommendations are addressed for the selection of the model fitting technique and the tuning strategy. Further research fields are outlined.

B.2 Zusammenfassung

Diese Arbeit ist der Überlebensvorhersage aus Genexpressionsdaten gewidmet. Es werden unterschiedliche Ansätze zur Modellselektion, Tuningstrategien und der Einfluss einzelner Gene auf das Überleben untersucht. Des Weiteren wird die Vorhersagegüte von Überlebensmodellen für eine gemischte Patienten-Population geprüft, die teils anfällig für Rezidive ist und teils geheilt ist.

Diese Arbeit besteht aus zwei Teilen. Im ersten Teil werden die zehn bekanntesten Ansätze zur Modellierung hochdimensionaler Daten vorgestellt. Diese kann man in Parameter Shrinkage-, Best-Subset-, Ensemblemethoden und Ansätze basierend auf achsentransformierten Variablenräumen einteilen. Unterschiede und Gemeinsamkeiten zwischen den Modellansätzen werden unter verschiedenen Gesichtspunkten diskutiert wie beispielsweise der Frage, ob die erklärte Variable in die Modellselektion einbezogen wird, ob Wechselwirkungen zwischen den Variablen berücksichtigt werden und ob Einzel- oder aggregierte Gene selektiert werden. Die Stärken und Schwächen der verschiedenen Techniken werden beschrieben.

Die sekundäre Fragestellung dieser Arbeit betrifft die Tuningstrategie, die einen starken Einfluss auf die Vorhersagegüte eines Überlebensmodells aus Microarraydaten hat. Sie umfasst viele Einzelaspekte wie die Resampling-Technik und die Wahl des Tuning-Parameters und eines Maßes zur Bestimmung der vorhersageoptimierten Modellgröße (Tuningkriterium).

Diese Arbeit befasst sich mit der Anzahl der Validierungsstichproben und mit dem Tuningkriterium. Es wird die Vorhersagequalität der Überlebensmodelle, die mit 5-, 10- und 20-facher Kreuzvalidierung bestimmt werden, verglichen. Da 5- und 10-fache Kreuzvalidierung eine übliche Wahl für die Abstimmung von Modellen mit Microarraydaten ist und eine niedrigere Anzahl von Kreuzvalidierungs-Stichproben (beispielsweise eine Leaving-One-Out Kreuzvalidierung) zu einem höheren Stichprobenbias und zu einer niedrigeren Vorhersagequalität führt, untersucht diese Arbeit zusätzlich den Einfluss der 20-fachen Kreuzvalidierung auf die Güte der Lebensdauermodelle.

Das zweite Thema, das im Rahmen dieser Arbeit behandelt wird, ist der Einfluss des Tuningkriteriums auf die Präzision der Modelle. Es werden Überlebensmodelle verglichen, die durch das Cross-Validation Partial Log-Likelihood Kriterium und den integrierten Brier Score selektiert wurden. Die Vor- und Nachteile der Tuningkriterien, wie beispielsweise Anwendungsbereiche für die Maße, werden beschrieben und die Güte der Modelle wird untersucht.

Die experimentelle Fragestellung dieser Arbeit betrifft den Einfluss einzelner Gene auf das Überleben. Heuristische Algorithmen werden verwendet, um die Effekte der Gene auf die Lebensdauer zu bestimmen. In dieser Arbeit wird geprüft, ob einflussreiche Gene existieren bzw. welche Modellierungsansätze signifikante Gene entdecken können.

Modellierungstechniken, Tuningstrategien und die Untersuchung von Effekten einzelner Gene auf das Überleben werden auf eine homogene Population von nicht geheilten Patienten angewendet.

Der zweite Teil dieser Arbeit befasst sich mit dem Vergleich von Lebensdauermodellen, die auf einer gemischten Population von rückfälligen und geheilten Patienten entwickelt werden. Die Klasse der semiparametrischen Cure-Modelle, insbesondere die Cox und AFT Misch-Cure-Modelle, die aus einem Latenz- und Inzidenzteil bestehen, wird beschrieben.

Ein neuer Modellierungsansatz wird vorgestellt, der die zwei Modellteile der Cure-Modelle berücksichtigt, indem Gene mit einem Einfluss auf das Überleben und auf Heilung separat bestimmt werden und das Cure-Modell aus den Einzeleffekten entwickelt wird.

Die Vor- und Nachteile der Anwendung von Cure-Modellen auf Curedaten im Vergleich mit dem Standard Coxmodell werden in dieser Arbeit gezeigt. Die Vorhersagegüte der Cox und AFT Misch-Modelle und des Standard Coxmodells sowie der Einfluss von Genen auf Überleben und den Cure werden beschrieben.

Weitere Themen dieser Arbeit sind Survial-Vorhersage Prozeduren, die Modelltuning und Modellentwicklung beinhalten, Maße zur Vorhersagequalität der Überlebensmodelle und Prozesse, um Genexpressionsdaten aus Gewebeproben zu gewinnen.

Die Ziele dieser Arbeit werden in primären, sekundären und explorativen Hypothesen zusammengefasst. Die Forschungsfragen werden auf Basis von vier Tumordatensätzen und vier generierten Datensätzen geprüft, wobei die generierten Daten hochsignifikante Variablen enthalten. Zum Abschluss werden geeignete Modellierungsstrategien für hochdimensionale Lebensdauermodelle empfohlen und zukünftige Forschungsthemen vorgestellt.

B.3 Curriculum vitae

Personal Data

Peter Wohlmuth, Mag.
Appener Weg 7
20251 Hamburg, Germany.

01577 47 499 59
peter.wohlmuth@gmx.at

10th July 1977
Austrian citizen, married.

Education

1987-1995
BG, BRG und BORG Eisenstadt, Austria.

1995-1996
Studies of Law, University of Vienna, Austria.

1996-2003
Studies of Statistics, University of Vienna, Austria.

Master Thesis:
Methods for automatic model selection and usage for economic time series.

2003-2004
Military duty, Martinskaserne Eisenstadt, Austria.

Since 2007
Doctoral program of Statistics, University of Vienna, Austria.

Since 2009
Doctoral Thesis:
Survival prediction with microarray data

Work experience

August 1996
Internship, Provincial government of Burgenland, Eisenstadt, Austria.

July 1997
Internship, municipality Müllendorf, Austria.

August 1998
Internship, Provincial government of Burgenland, Eisenstadt, Austria.

July 1999
Internship, municipality Müllendorf, Austria.

August 2000
Internship, municipality Müllendorf, Austria.

September – December 2004
Statistician: Kreutzer, Fischer & Partner Consulting GmbH, Wimbergergasse 14-16, 1070 Vienna, Austria.

Tasks:
Calculation of the consumption of refrigerated food, analysis of economic and social factors influencing consumption, setup of a consumer database, algorithm for optimal weighting of households, identification of clusters of consumers and products and correlations between them using multivariate techniques, model for the influence of advertisements on turnover within these clusters.

January – July 2005
Statistician: Statistics Austria, Guglgasse 13, 1110 Vienna, Austria, population census.

Tasks:
Preparation, clearance and analysis of huge registry datasets, data merge of the databases, anonymisation of the data, presentation of the project, evaluation of the results by comparison with a short survey.

August 2005 – March 2006
Contract for work and labour: Fischer & Partner Consulting GmbH.

Tasks:
Economic analysis of the development of the labour market in Austria, of the purchasing power and of the economic development of the Burgenland after the EU enlargement, development of education, income, age structure, structure of labour force, and the size of households.

March 2006 – June 2008
Biostatistician: ABCSG, Boltzmanngasse 24-26, 1090 Vienna, Austria.

Tasks:
Analysis of clinical trials: phase II-III, descriptive, uni- and multivariate analysis for scientific presentations and publications, design of trials, sample size and power

calculations.

Writing statistical analysis plans, generating case report forms of clinical trials, cooperation with project- and data management, implementation and evaluation of data management procedures, coordination of tasks for preparing and analyzing data.

Cooperation with physicians in writing scientific articles, formulating statistical hypotheses and interpretation of the outcome, preparing the output for publications in journals.

Programming of status reports for sponsors and payment lists for physicians recruiting patients, backup of the patient database, and support for SAS in the ABCSG.

Since June 2008
Biostatistician and department head of biostastics and data mangement: Asklepios proresearch, Lohmühlenstrasse 5, Haus J, 20099 Hamburg, Germany.

Tasks:
Design and analysis of clinical trials, sample size and power calculation, setup of a data registry for health services research and medical research.

Implementing medical hypothesis, writing statistical analysis plans, delegation of data management tasks, preparation and analysis of data, and statistical summaries.

Skills

English

Computer literacy:

Windows 95/98/2000/XP/Vista and Linux (SUSE, Ubuntu, Mint), MS Office 2000/XP/2003, OpenOffice, LibreOffice, GoLive, SPSS, SAS, Java (elementary), R and S-Plus, database management systems (Infermed Macro).

February 2005
Seminar: **Access: Basics**
Introduction of MS Access, characteristics of databases, data import/export, generating reports and forms, design of databases considering standards in databases, reports using SQL queries.

March 2005
Seminar: **Host: Basics**
Introduction to the tasks of the host, connection to the host, up- and download, connectivity with local software applications like SAS, host applications.

April 2005
Seminar: **SAS: Introduction**
Introduction to the SAS Software, programming with data and proc step, generating reports and lists, formats, import/export of/to other data sources and –formats, improving the SAS output.

April 2005
Seminar: **Data design**
Databases and SQL, setup and design of databases, implementing database projects, rules for databases, clearing databases and database management systems.

May 2005
Seminar: **Excel: Macro-Programming**
Visual Basic programming in MS Office, automation of tasks, recording and editing of macro codes, macro programming and implementation in MS Office applications.

May 2005
Seminar: **SAS/EG and DB2**
Introduction to SAS Enterprise Guide, advantages in comparison to SAS, graphical routines in SAS/EG, connectivity to data bases, descriptive analysis and statistical methods in SAS/EG.

June 2005
Seminar: **SAS: Programming**
Data manipulation, data merge, loops, conditions, Macro programming in SAS, data- and numerical formats, installation of SAS tasks.

May/June 2010
Seminar: Medical seminar for non-medics.

Other skills:

Driving license.

Schippinger, W., Samonigg, H., Schaberl-Moser, R., Greil, R., Thödtmann, R., Tschmelitsch, J., Jagoditsch, M., Steger, G. G., Jakesz, R., Herbst, F., Hofbauer, F., Rabl, H., Wohlmuth, P., Gnant, M., Thaler, J. (2007): A prospective randomised phase III trial of adjuvant chemotherapy with 5-fluorouracil and leucovorin in patients with stage II colon cancer. *British Journal of Cancer* (2007) 97, 1021-1027.

Mlineritsch, B., Tausch, C., Singer, C., Luschin-Ebengreuth, G., Jakesz, R., Ploner, F., Stierer, M., Melbinger, E., Menzel, C., Urbania, A., Fridrik, M., Steger, G. G., Wohlmuth, P., Gnant, M., Greil, R. (2007): Exemestane as primary systemic treatment for hormone receptor positive post-menopausal breast cancer patients: a phase II trial of the Austrian Breast and Colorectal Cancer Study Group (ABCSG-17). *Breast Cancer Res. Treat.* (2007).

Gnant, M., Mlineritsch, B., Luschin-Ebengreuth, G., Kainberger, F., Kaessmann, H., Piswanger-Soelkner, C., Seifert, M., Schippinger, W., Menzel, C., Dubsky, P., Fitzal, F., Steger, G., Wohlmuth, P., Mittlboeck, M., Greil, R., Marth, C., Kubista, E., Samonigg, H., Jakesz, R., on behalf of the ABCSG (2007): Bone mineral density (BMD) at 5 years after diagnosis in premenopausal patients with endocrine-responsive breast cancer, after 3 years of adjuvant endocrine treatment with goserelin and tamoxifen or anastrozole or both treatments in combination with zoledronic acid - new results from ABCSG-12 . *San Antonio Breast Cancer Symposium 12, 2007*. abstract #26.

Mittlböck, M., Gnant, M., Greil, R., Fridrik, M. and Wohlmuth, P. (2007): Evaluation of Surrogate Markers When Surrogate and True Endpoints Are Survival Times. 28th Meeting of the Inter- national Society of Clinical Biostatistics, Alexandroupolis, Greece, 29.7.-2.8.2007.

Gnant, M., Mlineritsch, B., Luschin-Ebengreuth, G., Kainberger, F., Kässmann, H., Piswanger-Sölkner, J. C., Seifert, M., Ploner, F., Menzel, C., Dubsky, P., Fitzal, F., Bjelic-Radisic, V., Steger, G., Greil, R., Marth, C., Kubista, E., Samonigg, H., Wohlmuth, P., Mittlböck, M., Jakesz, R., on behalf of the Austrian Breast and Colorectal Cancer Study Group (ABCSG), Vienna, Austria: Adjuvant endocrine therapy plus zoledronic acid in premenopausal women with early-stage breast cancer: 5-year follow-up of the ABCSG-12 bone-mineral density substudy . *The Lancet Oncology* - Vol. 9, Issue 9, September 2008, Pages 840-849

Fürnkranz, A., Julian, J. K., Schmidt, B., Wohlmuth, P., Tilz, R., Kuck, K. H., Ouyang, F. (2011). Ipsilateral pulmonary vein isolation performed by a single continuous circular lesion: role of pulmonary vein mapping during ablation. Europace. 2011 Jul;13(7):935-41. doi: 10.1093/europace/eur067. Epub 2011 Mar 31.
Hager, A., Wohlmuth, P., Otte, B. (2011). Inzidenz der feuchten altersabhängigen Makuladegeneration (AMD) nach pars-plana Vitrektomie bei Makulaforamen und Makula pucker. Poster and Abstract. Deutsche Ophthalmologische Gesellschaft.

Chun, K. R., Fürnkranz, A., Köster, I., Metzner, A., Tönnis, T., Wohlmuth, P., Wissner, E., Schmidt, B., Ouyang, F., Kuck, K. H. (2012). Two versus one repeat freeze-thaw cycle(s) after cryoballoon pulmonary vein isolation: the alster extra pilot study. J Cardiovasc Electrophysiol. 2012 Aug;23(8):814-9. doi: 10.1111/j.1540-8167.2012.02315.x. Epub 2012 Apr 4.

Hager, A., Wohlmuth, P., Otte, B. (2012). Prävalenz der feuchten AMD nach PPV mit Lösung vitreoretinaler Adhärenzen. Klin Monatsbl Augenheilkd 2012; 229 - KV44. DOI: 10.1055/s-0032-1331557.

Hellmeyer, L., Herz, K., Liedtke, B., Wohlmuth, P., Schmidt, S., Hackeloeer, B. J. (2012). The underestimation of immaturity in late preterm infants. Arch Gynecol Obstet. 2012 Sep;286(3):619-26. doi: 10.1007/s00404-012-2366-7. Epub 2012 May 5.

Herz, K., Wohlmuth, P., Liedtke, B., Schmidt, S., Hackelöer, B. J., Hellmeyer, L. (2012). Late preterms: the influence of foetal gender on neonatal outcome. Z Geburtshilfe Neonatol. 2012 Jun;216(3):141-6. doi: 10.1055/s-0032-1309050. Epub 2012 Jun 21.

Konstantinidou, M., Wissner, E., Chun, J. K., Koektuerk, B., Metzner, A., Tilz, R. R., Rillig, A., Fuernkranz, A., Wohlmuth, P., Ouyang, F., Kuck, K.H. (2012). Luminal esophageal temperature rise and esophageal lesion formation following remote-controlled magnetic pulmonary vein isolation. Heart Rhythm. 2011 Dec;8(12):1875-80. doi: 10.1016/j.hrthm.2011.07.031. Epub 2011 Jul 28.

Rillig, A., Meyerfeldt, U., Tilz, R. R., Talazko, J., Arya, A., Zvereva, V., Birkemeyer, R., Miljak, T., Hajredini, B., Wohlmuth, P., Fink, U., Jung, W. (2012). Incidence and long-term follow-up of silent cerebral lesions after pulmonary vein isolation using a remote robotic navigation system as compared with manual ablation. Circ Arrhythm Electrophysiol. 2012 Feb;5(1):15-21. doi: 10.1161/CIRCEP.111.967497. Epub 2012 Jan 13.

Rillig, A., Schmidt, B., Feige, B., Wißner, E., Metzner, A., Mathew, S., Wohlmuth, P., Ouyang, F., Kuck, K. H., Tilz, R. R. (2012). Novel visually guided left atrial isthmus ablation using a robotic navigation system: safety, feasibility and clinical outcome. Clin Res Cardiol 101, Suppl 1, April 2012. 78. Jahrestagung der Deutschen Gesellschaft für Kardiologie.13. April 2012

Tilz, R. R., Rillig, A., Thum, A. M., Arya, A., Wohlmuth, P., Metzner, A., Mathew, S., Yoshiga, Y., Wissner, E., Kuck, K. H., Ouyang, F. (2012). Catheter ablation of long-standing persistent atrial fibrillation: 5-year outcomes of the Hamburg Sequential Ablation Strategy. J Am Coll Cardiol. 2012 Nov 6;60(19):1921-9. doi: 10.1016/j.jacc.2012.04.060. Epub 2012 Oct 10.

Rillig, A., Schmidt, B., Steven, D., Meyerfeldt, U., Di Biase, L., Wissner, E., Becker, R., Thomas, D., Wohlmuth, P., Gallinghouse, G. J., Scholz, E., Jung, W., Willems, S., Natale, A., Ouyang, F., Kuck, K. H., Tilz, R. (2013). Study design of the man and machine trial: a prospective international controlled noninferiority trial comparing manual with robotic catheter ablation for treatment of atrial fibrillation. J Cardiovasc Electrophysiol. 2013 Jan;24(1):40-6. doi: 10.1111/j.1540-8167.2012.02418.x. Epub 2012 Nov 6.

Metzner, A., Burchard, A., Wohlmuth, P., Rausch, P., Bardyszewski, A., Gienapp, C., Tilz, R. R., Rillig, A., Mathew, S., Deiss, S., Makimoto, H., Ouyang, F., Kuck, K. H., Wissner, E. (2013). Increased Incidence of Esophageal Thermal Lesions using the Second-Generation 28mm Cryoballoon. Circ Arrhythm Electrophysiol. 2013 Jun 7. [Epub ahead of print] doi: 10.1161/CIRCEP.113.000228