Dynamic prediction of survival with clinical and genomic data

Hans C. van Houwelingen

Department of Medical Statistics and Bioinformatics (retired)
Leiden University Medical Center
The Netherlands

jcvanhouwelingen@lumc.nl

Based on joint work with
Jelle Goeman, Hein Putter

Summary:

An important clinical application of biostatistics is the development of statistical models for the prognosis of a patient at the moment of diagnosis. In cancer the usual way of giving a prognosis is by means of the x-year survival probability, with x=1, 5 or 10, for example. Traditionally, the prognosis is based on clinical information at the start of the treatment, like age, gender, size of the tumor, tumor stage etc. In the last decade new types of genomic information have become available like micro-array gene expression and proteomic mass spectrometry data. The problem with this new type of data is its abundance. Micro-arrays can measure the expression of tens of thousands of genes, for example.

The talk will address three issues:

1. How to obtain valid prognostic model based on high-dimensional genomic data.
2. How to assess the added value of the genomic information.
3. How to obtain robust dynamic predictions (predictions available later on in the follow-up)

Talk based on

van Houwelingen, HC; Bruinsma, T; Hart, AAM; van 't Veer, LJ; Wessels, LFA. 2006.
Cross-validated Cox regression on microarray gene expression data.
STATISTICS IN MEDICINE 25 (18): 3201-3216.

van Houwelingen, HC; Putter, H., 2011
Dynamic prediction in clinical survival analysis, CRC/Chapman & Hall
chapters 11 and 12. (Will appear on December 1, 2011)

Crash-course survival analysis.

Definitions

- Survival time T_{surv}
- Survival function $S(t) = P(T_{surv} > t)$
- Censoring time (end of follow-up) T_{cens}
- Censoring function $C(t) = P(T_{cens} > t)$
- Observed $T = \min(T_{cens}, T_{surv})$
- Event indicator $\delta = 1$ if $T = T_{surv}$, $\delta = 0$ if $T = T_{cens}$

Prediction model

- $$\boxed{\text{hazard } h(t)} = -\frac{S'(t)}{S(t)} = -\frac{d\ln(S(t))}{dt};\ h(t)dt = \frac{P(T \le t+dt)}{P(T \ge t)}$$
- Cox proportional hazard model $\boxed{h(t \mid X) = h_0(t)\exp(X'\beta)}$

marginal

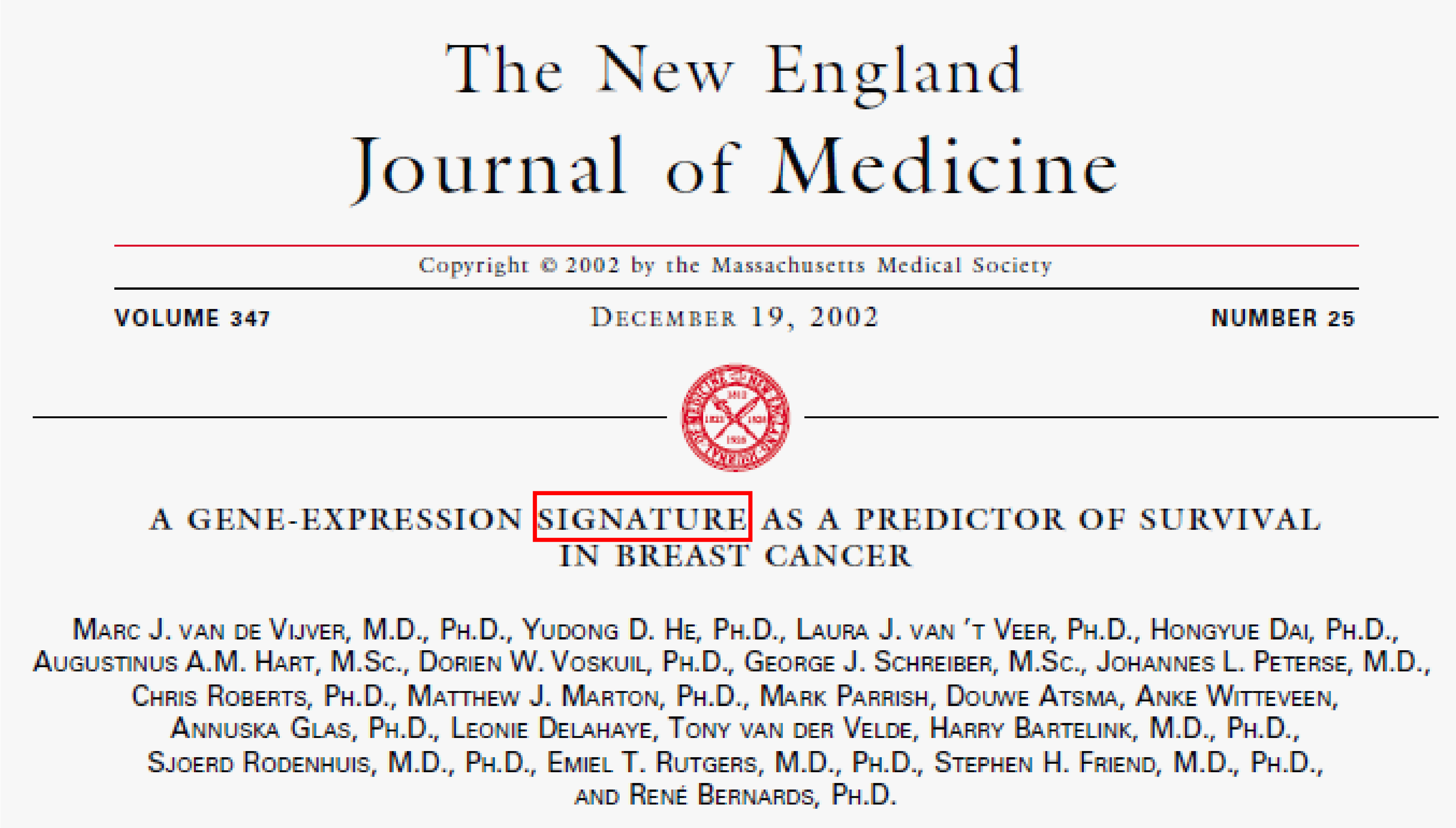
Estimation

- Survival and Censoring function estimated by Kaplan-Meier curves
- Likelihood of observation $(T,\delta) = (t,d)$: $\quad S(t)h(t)^d$
- Regression parameters β estimated by maximum partial likelihood
- Baseline hazard $h_0(t)$ estimated by Breslow estimator

negative is good
positive is bad

concentrated in
even times

The data are from



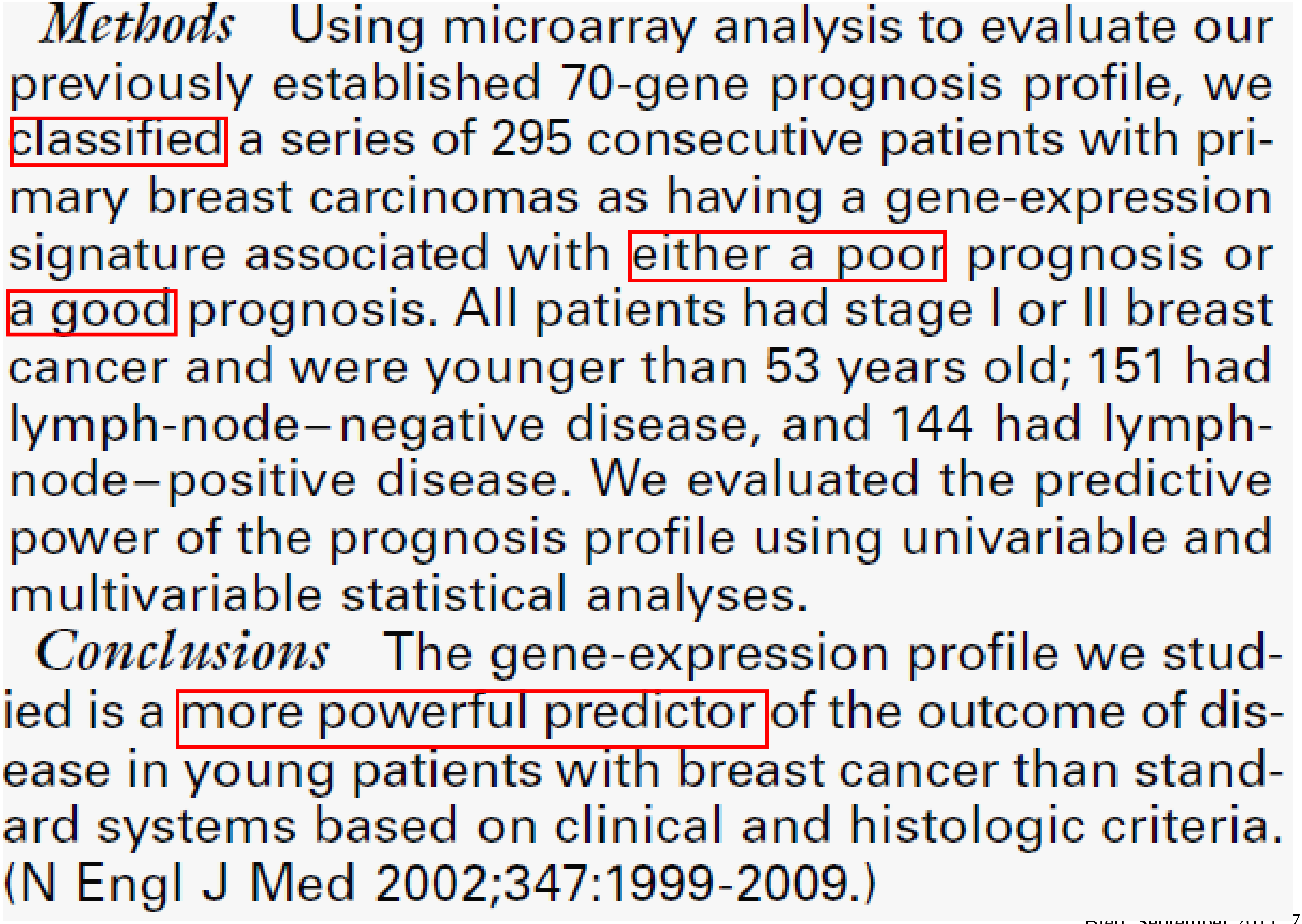
The New England Journal of Medicine

Copyright © 2002 by the Massachusetts Medical Society

VOLUME 347 December 19, 2002 NUMBER 25

A GENE-EXPRESSION SIGNATURE AS A PREDICTOR OF SURVIVAL IN BREAST CANCER

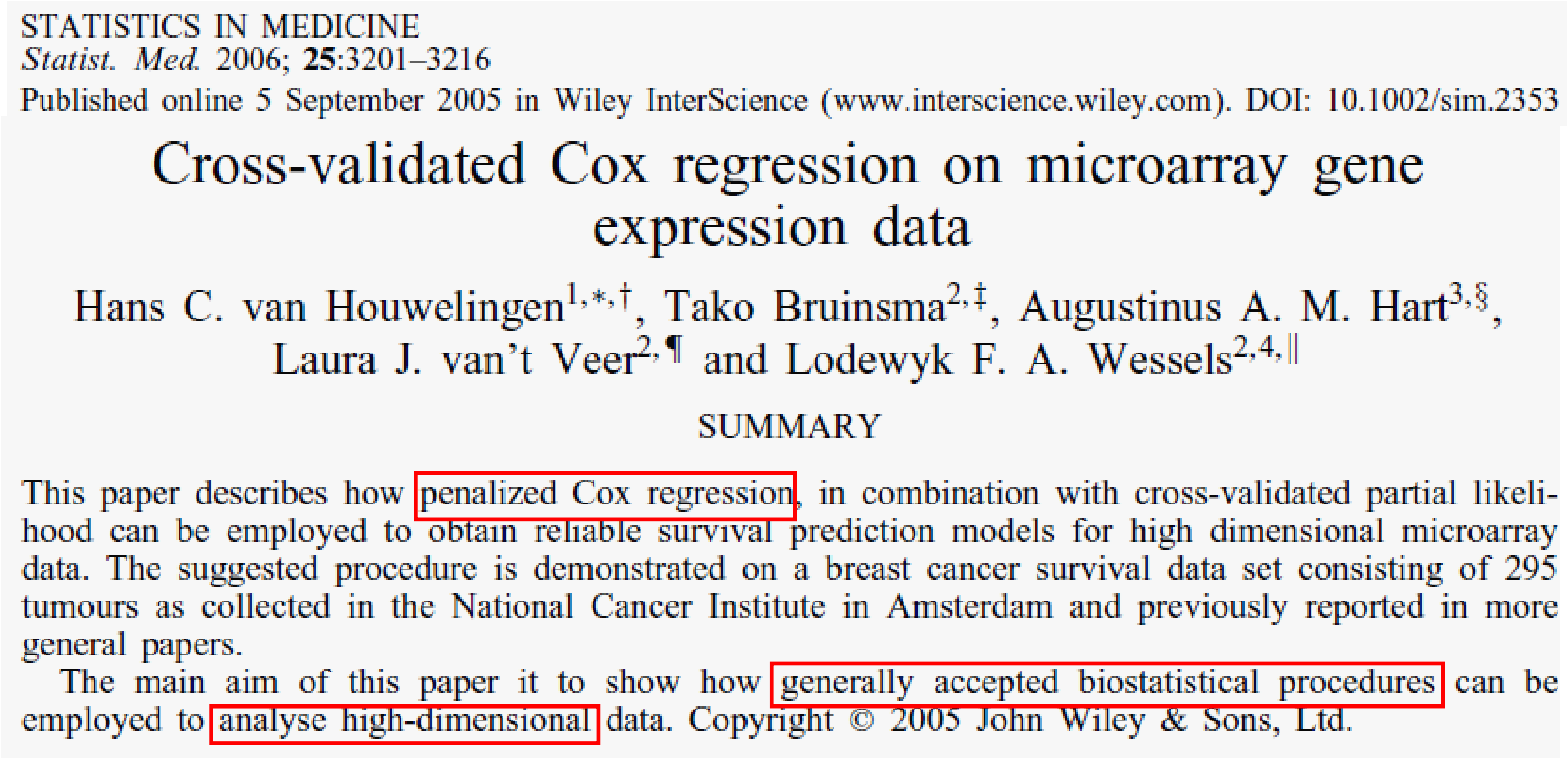
Marc J. van de Vijver, M.D., Ph.D., Yudong D. He, Ph.D., Laura J. van 't Veer, Ph.D., Hongyue Dai, Ph.D., Augustinus A.M. Hart, M.Sc., Dorien W. Voskuil, Ph.D., George J. Schreiber, M.Sc., Johannes L. Peterse, M.D., Chris Roberts, Ph.D., Matthew J. Marton, Ph.D., Mark Parrish, Douwe Atsma, Anke Witteveen, Annuska Glas, Ph.D., Leonie Delahaye, Tony van der Velde, Harry Bartelink, M.D., Ph.D., Sjoerd Rodenhuis, M.D., Ph.D., Emiel T. Rutgers, M.D., Ph.D., Stephen H. Friend, M.D., Ph.D., and René Bernards, Ph.D.



Methods Using microarray analysis to evaluate our previously established 70-gene prognosis profile, we classified a series of 295 consecutive patients with primary breast carcinomas as having a gene-expression signature associated with either a poor prognosis or a good prognosis. All patients had stage I or II breast cancer and were younger than 53 years old; 151 had lymph-node–negative disease, and 144 had lymph-node–positive disease. We evaluated the predictive power of the prognosis profile using univariable and multivariable statistical analyses.

Conclusions The gene-expression profile we studied is a more powerful predictor of the outcome of disease in young patients with breast cancer than standard systems based on clinical and histologic criteria. (N Engl J Med 2002;347:1999-2009.)

Re-analyzed in



STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:3201–3216
Published online 5 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2353

Cross-validated Cox regression on microarray gene expression data

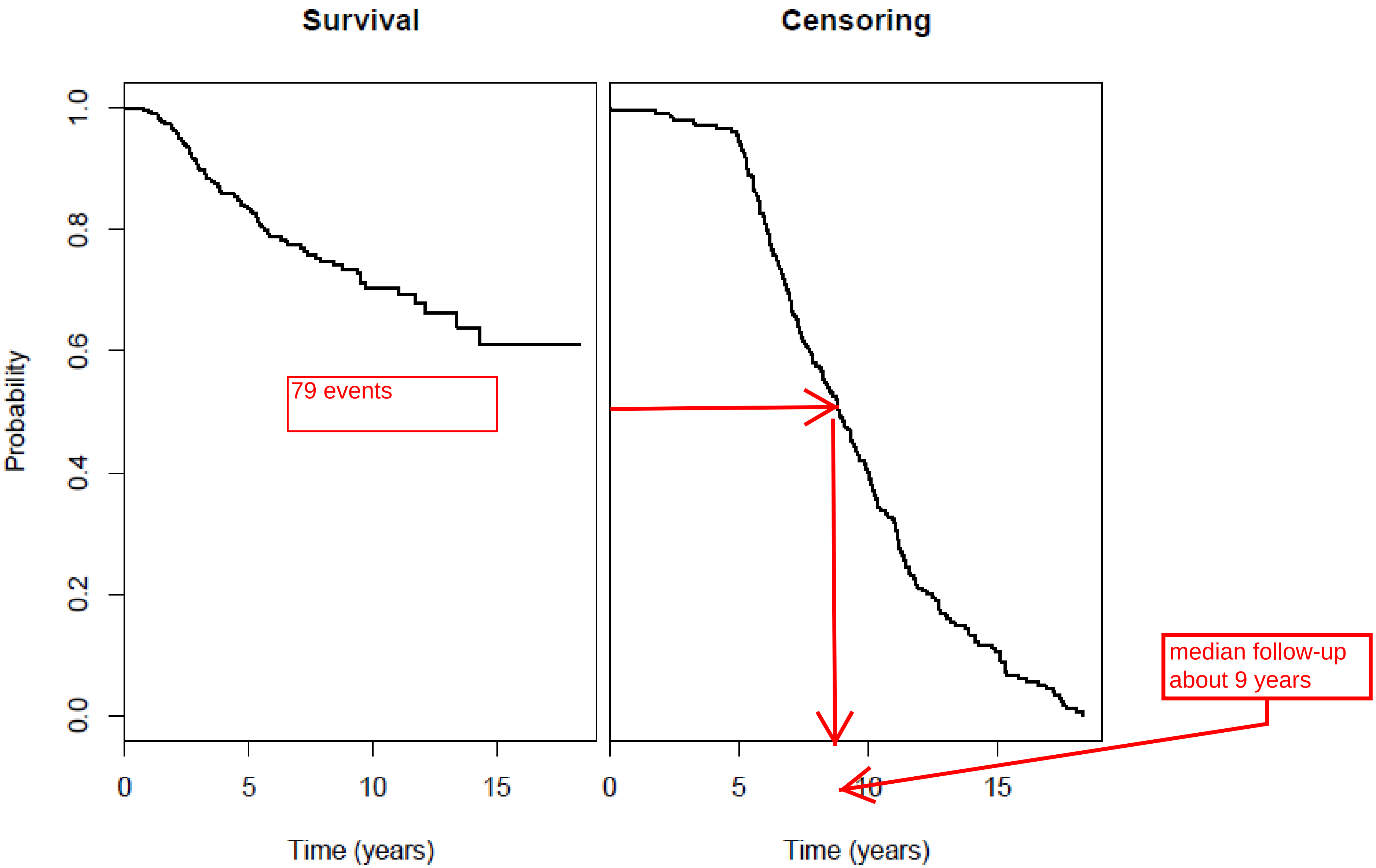
Hans C. van Houwelingen[1,*,†], Tako Bruinsma[2,‡], Augustinus A. M. Hart[3,§], Laura J. van't Veer[2,¶] and Lodewyk F. A. Wessels[2,4,‖]

SUMMARY

This paper describes how penalized Cox regression, in combination with cross-validated partial likelihood can be employed to obtain reliable survival prediction models for high dimensional microarray data. The suggested procedure is demonstrated on a breast cancer survival data set consisting of 295 tumours as collected in the National Cancer Institute in Amsterdam and previously reported in more general papers.

The main aim of this paper it to show how generally accepted biostatistical procedures can be employed to analyse high-dimensional data. Copyright © 2005 John Wiley & Sons, Ltd.

Information on survival and censoring



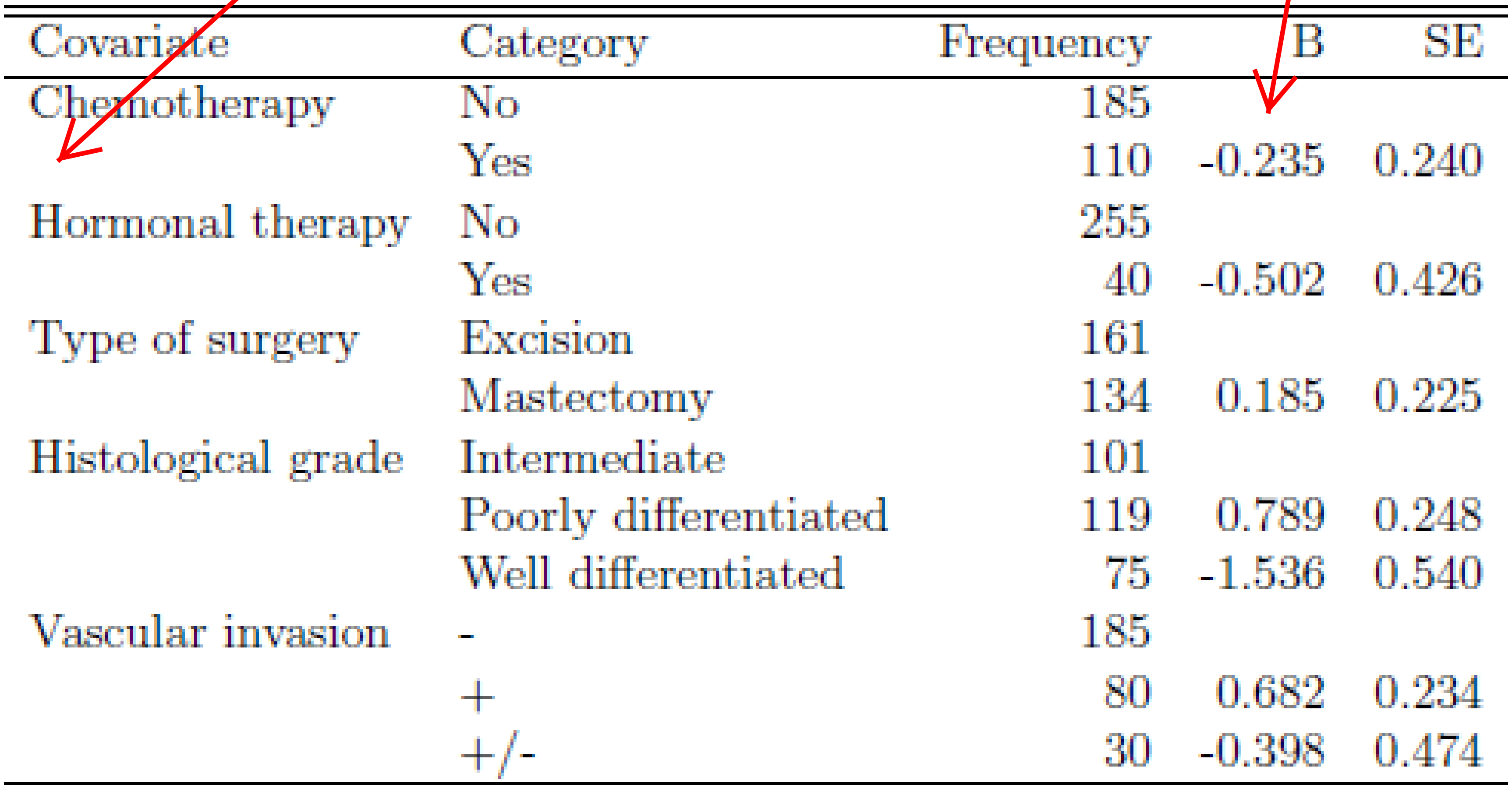
79 events

median follow-up about 9 years

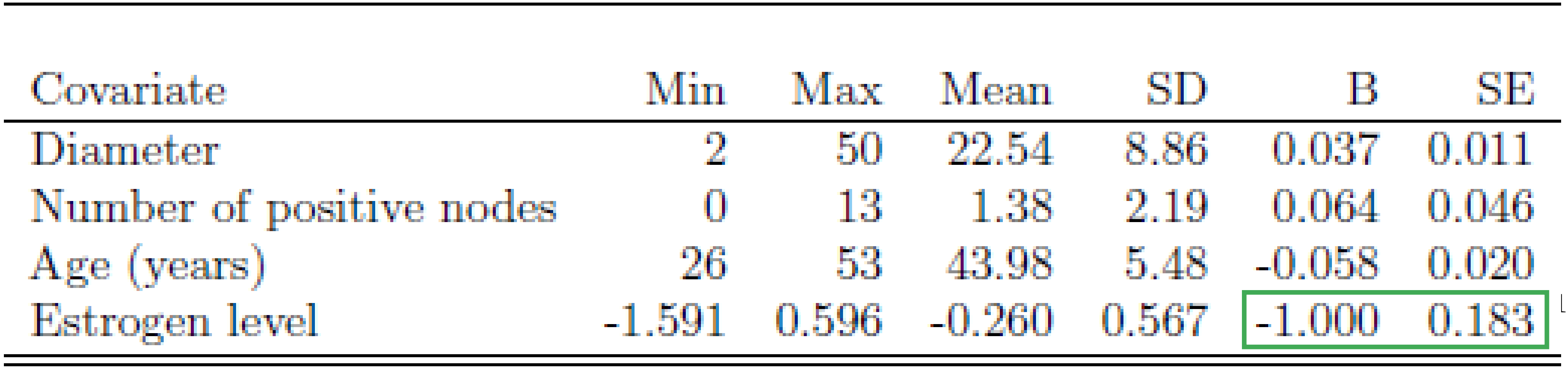
Therapy depends on situation of the patient. Model is not causal.

negative is good

Clinical information



Covariate	Category	Frequency	B	SE
Chemotherapy	No	185		
	Yes	110	-0.235	0.240
Hormonal therapy	No	255		
	Yes	40	-0.502	0.426
Type of surgery	Excision	161		
	Mastectomy	134	0.185	0.225
Histological grade	Intermediate	101		
	Poorly differentiated	119	0.789	0.248
	Well differentiated	75	-1.536	0.540
Vascular invasion	-	185		
	+	80	0.682	0.234
	+/-	30	-0.398	0.474



Covariate	Min	Max	Mean	SD	B	SE
Diameter	2	50	22.54	8.86	0.037	0.011
Number of positive nodes	0	13	1.38	2.19	0.064	0.046
Age (years)	26	53	43.98	5.48	-0.058	0.020
Estrogen level	-1.591	0.596	-0.260	0.567	-1.000	0.183

the other ones contained no information or were too noisy

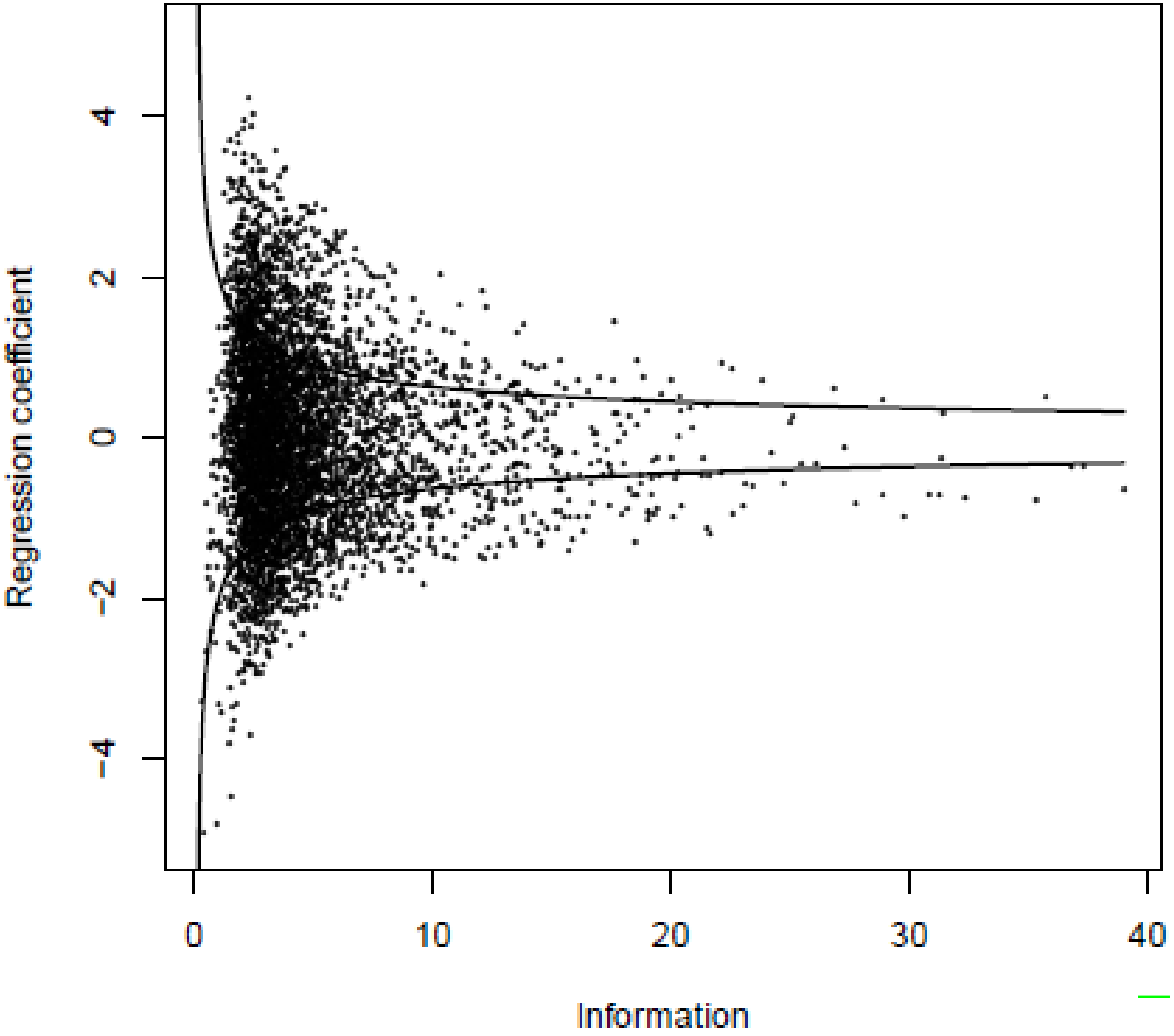
Genomic information

- Gene expression (ln(ratio)) on 4919 genes (out of 24885 genes)
- Internally normalized at geometric mean=0.

Many genes show some effect on survival.

Funnel plot of regression coefficient b versus $Information = 1/se(b)^2$.

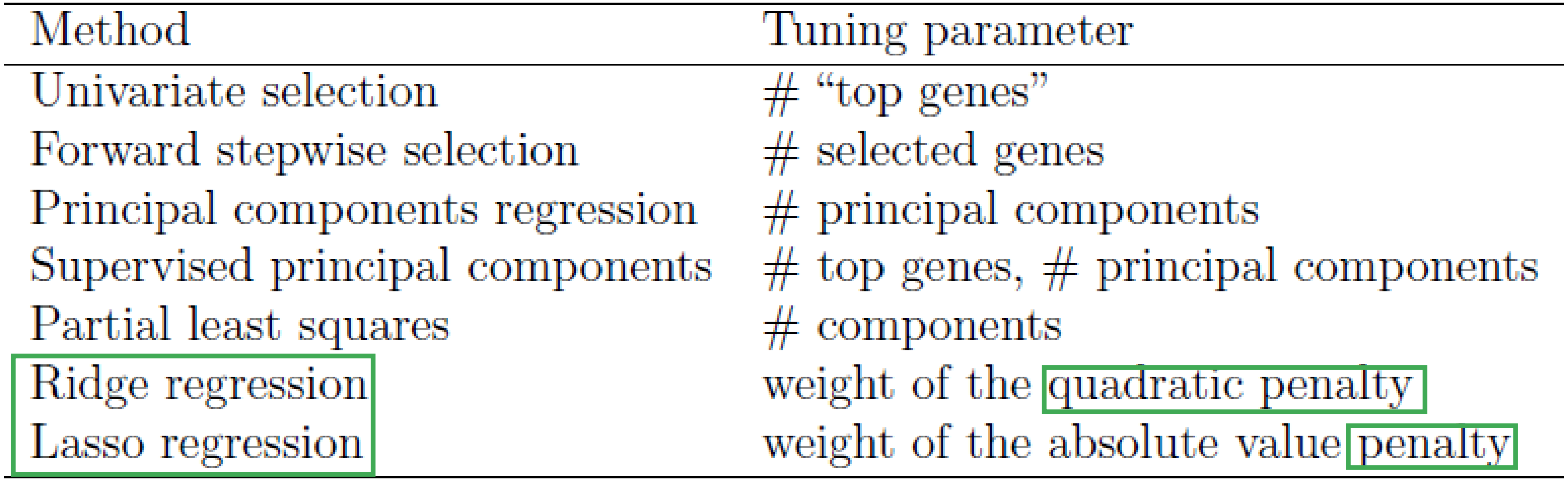
Band corresponds with $\pm 2 \cdot se(b)$



Major problem: How to handle so many predictors?

Using them all in Cox regression does not make sense. Some form of tuning is needed.

Possible approaches



Method	Tuning parameter
Univariate selection	# "top genes"
Forward stepwise selection	# selected genes
Principal components regression	# principal components
Supervised principal components	# top genes, # principal components
Partial least squares	# components
Ridge regression	weight of the quadratic penalty
Lasso regression	weight of the absolute value penalty

Methods compared in Bøvelstad et al., Bioinformatics . 2007

Conclusion: Ridge regression (as used in my paper) performs best on this type of data.

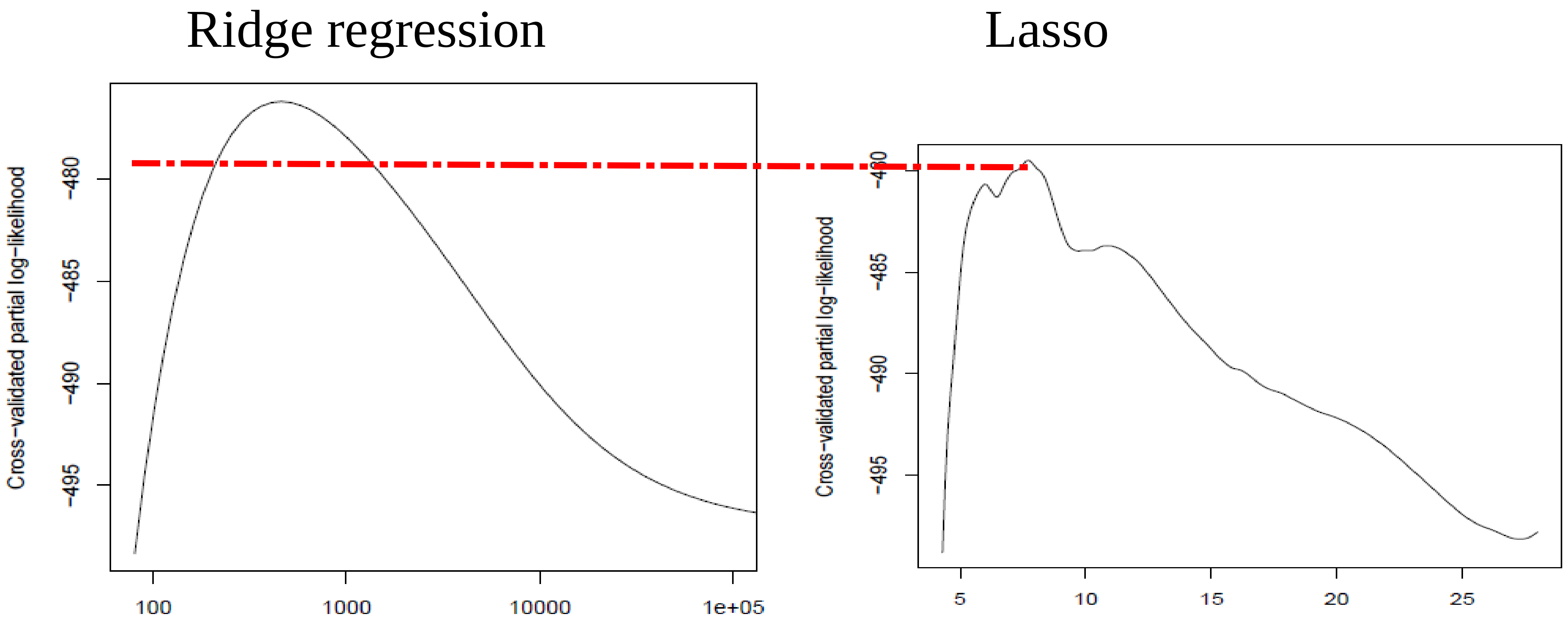
Penalized Cox regression using genomic data

baseline hazard

- penalized log-likelihood: $l_{pen}(\beta, h_0) = l(\beta, h_0) - \lambda \cdot pen(\beta)$
- Ridge regression $pen(\beta) = 0.5\sum_j \beta_j^2$
- LASSO $pen(\beta) = \sum_j |\beta_j|$
- Both implemented in Goeman's *R*-package **"penalized"**
- Optimal λ, λ_{opt}, obtained through cross-validation (using the cross-validated partial log-likelihood CVPL)
- Big difference
  - o Ridge regression $\hat{\beta}_j \neq 0$ for all j no "feature" selection
  - o Lasso $\hat{\beta}_j = 0$ for most j strong feature selection

Results

Crossvalidated partial log-likelihoods



Ridge regression performs better and is “smoother”.
Optimal Lasso uses only 16 genes in the model.

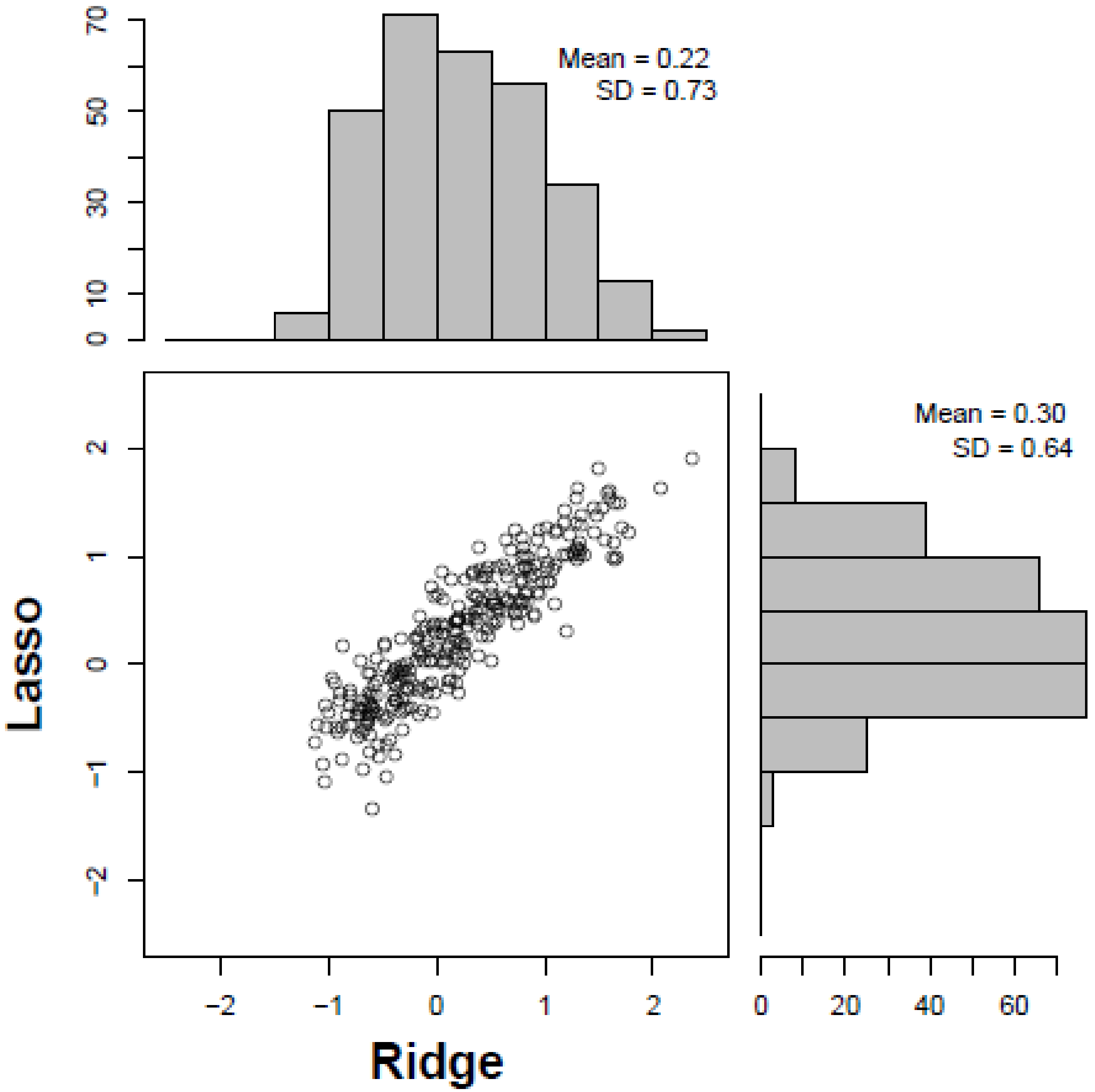
Relation between Ridge and Lasso can be studied by comparing

$$PI_{Ridge} = X'\hat{\beta}_{Ridge} \text{ and } PI_{Lasso} = X'\hat{\beta}_{Lasso}$$

Correlation =0.90

$$SD(PI_{Ridge}) > SD(PI_{Lasso})$$

Although the regression coefficients differ very much.



It is hard to see the difference between the two predictors.

Proportional hazards may not be true

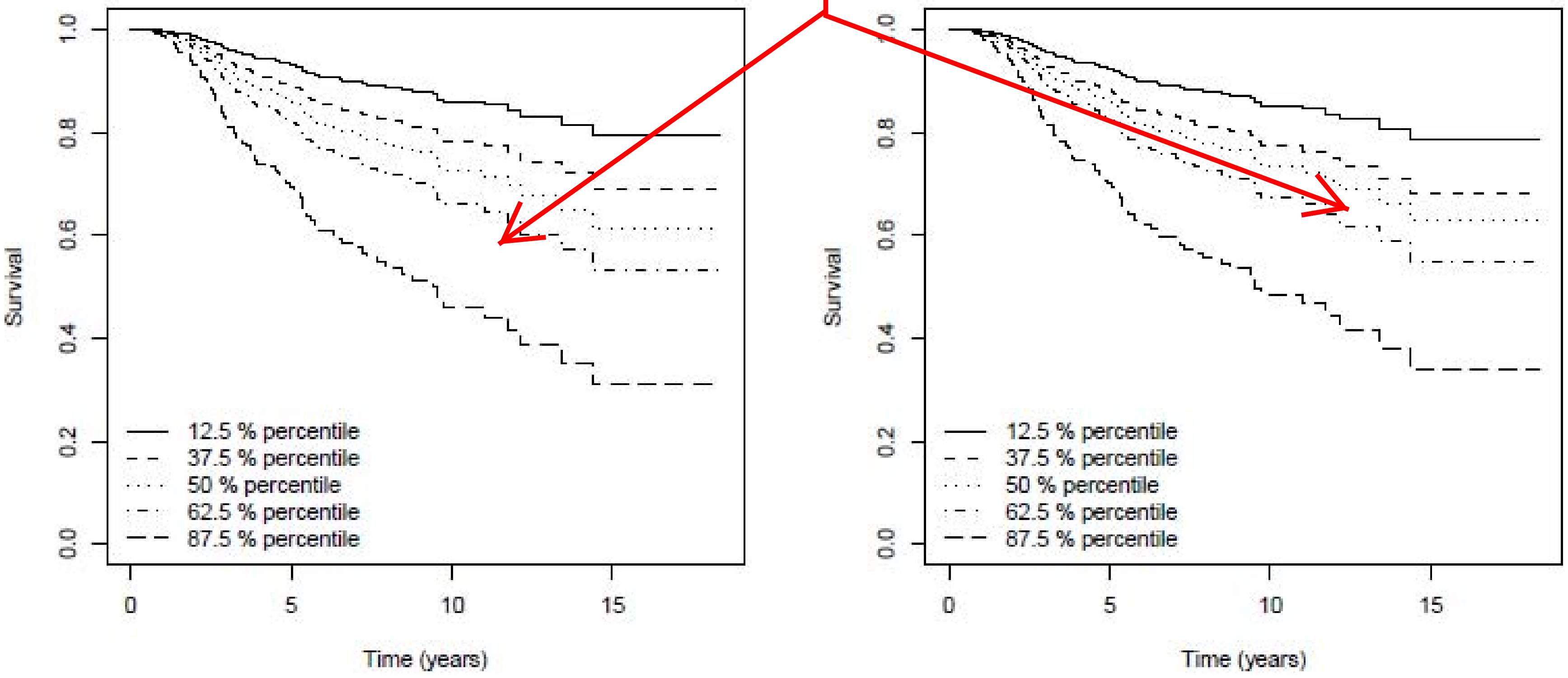
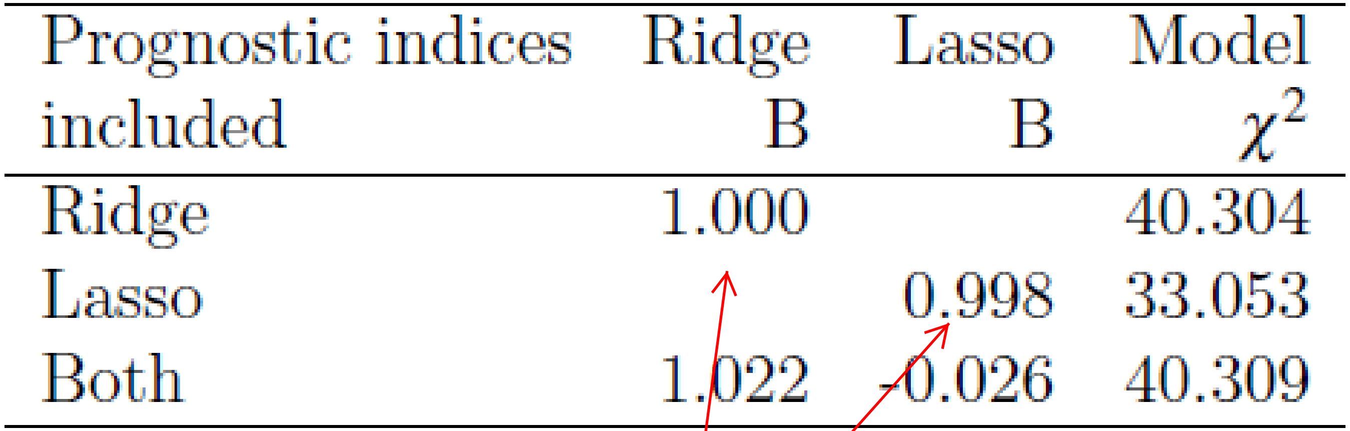


Figure 11.3: *Survival curves for ridge (left) and lasso (right)*

However:

Term due to Mark van der Laan

Lasso, does not contain any “additional information” as can be seen from a “super learner” model on the cross-validated predictors.



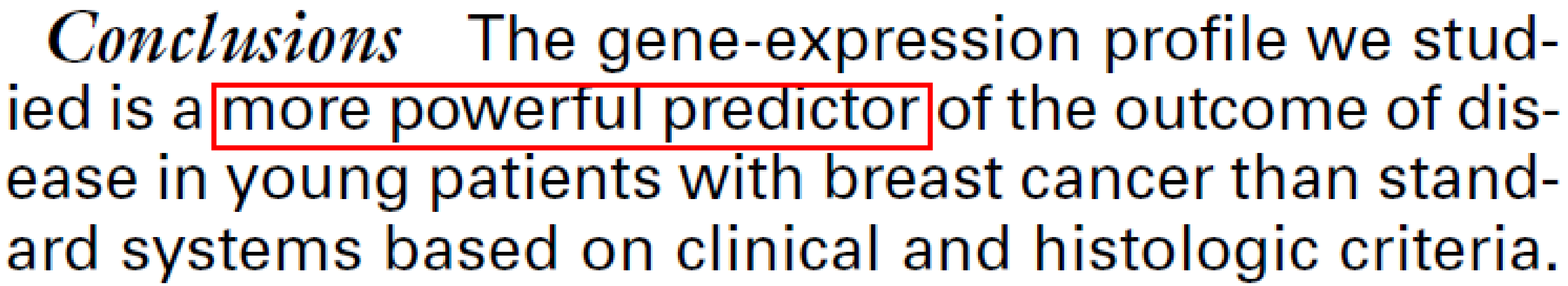
Prognostic indices included	Ridge B	Lasso B	Model χ^2
Ridge	1.000		40.304
Lasso		0.998	33.053
Both	1.022	-0.026	40.309

Table 11.2: *Cox regression on cross-validated prognostic indices*

Calibration is OK

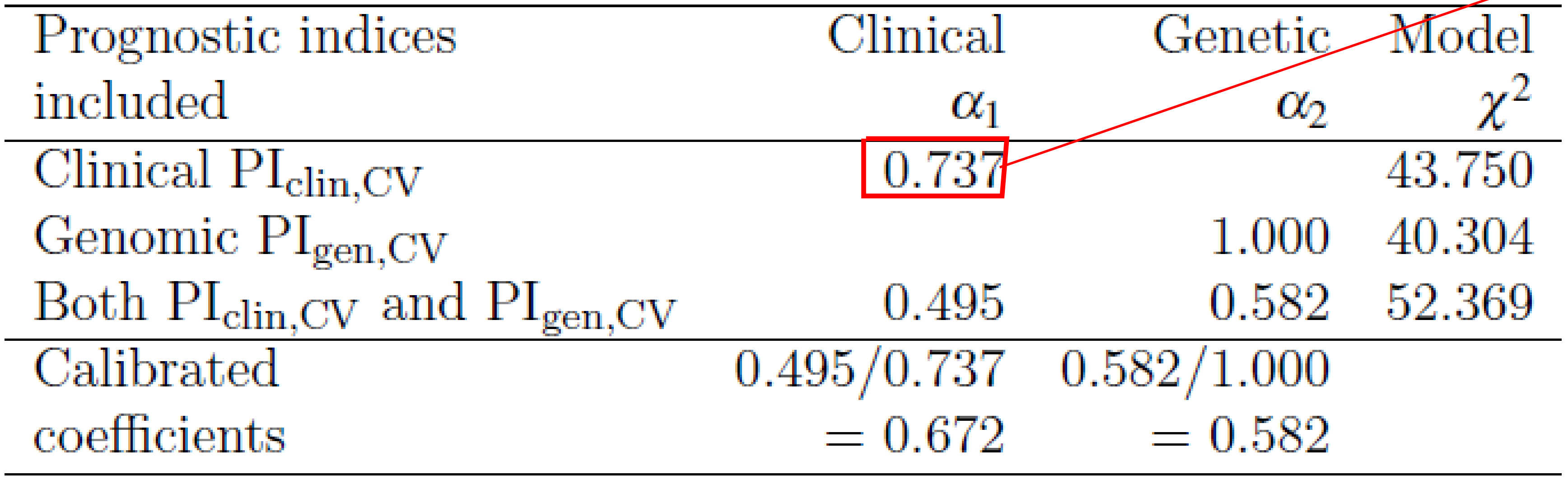
Genomic versus clinical predictor.

Remember the Van de Vijver paper



Conclusions The gene-expression profile we studied is a more powerful predictor of the outcome of disease in young patients with breast cancer than standard systems based on clinical and histologic criteria.

However, Clinical performs slightly better and Genomics does not add vary much (correlation r=0.652)



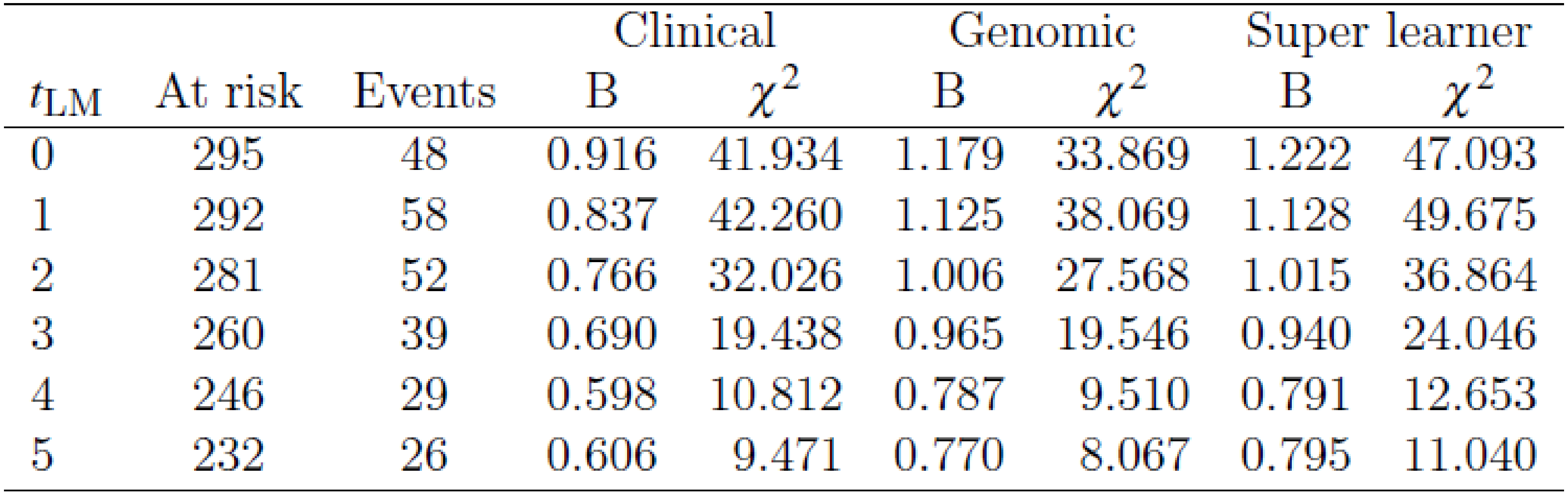
Prognostic indices included	Clinical α_1	Genetic α_2	Model χ^2
Clinical $\mathrm{PI}_{\mathrm{clin,CV}}$	0.737		43.750
Genomic $\mathrm{PI}_{\mathrm{gen,CV}}$		1.000	40.304
Both $\mathrm{PI}_{\mathrm{clin,CV}}$ and $\mathrm{PI}_{\mathrm{gen,CV}}$	0.495	0.582	52.369
Calibrated coefficients	$0.495/0.737$ $= 0.672$	$0.582/1.000$ $= 0.582$	

Too small

Table 11.3: *Super model Cox regression*

Dynamic prediction based on Landmarking.

Predict from t_{LM} to $t_{LM} + w$



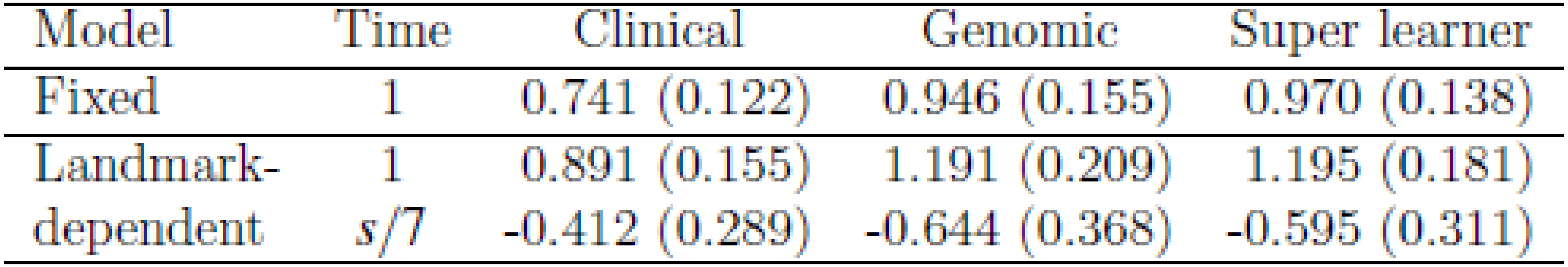
			Clinical		Genomic		Super learner	
t_{LM}	At risk	Events	B	χ^2	B	χ^2	B	χ^2
0	295	48	0.916	41.934	1.179	33.869	1.222	47.093
1	292	58	0.837	42.260	1.125	38.069	1.128	49.675
2	281	52	0.766	32.026	1.006	27.568	1.015	36.864
3	260	39	0.690	19.438	0.965	19.546	0.940	24.046
4	246	29	0.598	10.812	0.787	9.510	0.791	12.653
5	232	26	0.606	9.471	0.770	8.067	0.795	11.040

Prediction window $w = 5$ years.

Prediction based on existing (cross-validated) predictors.

Supermodel smoothes the landmark effect

Stacking the landmark data sets (s)



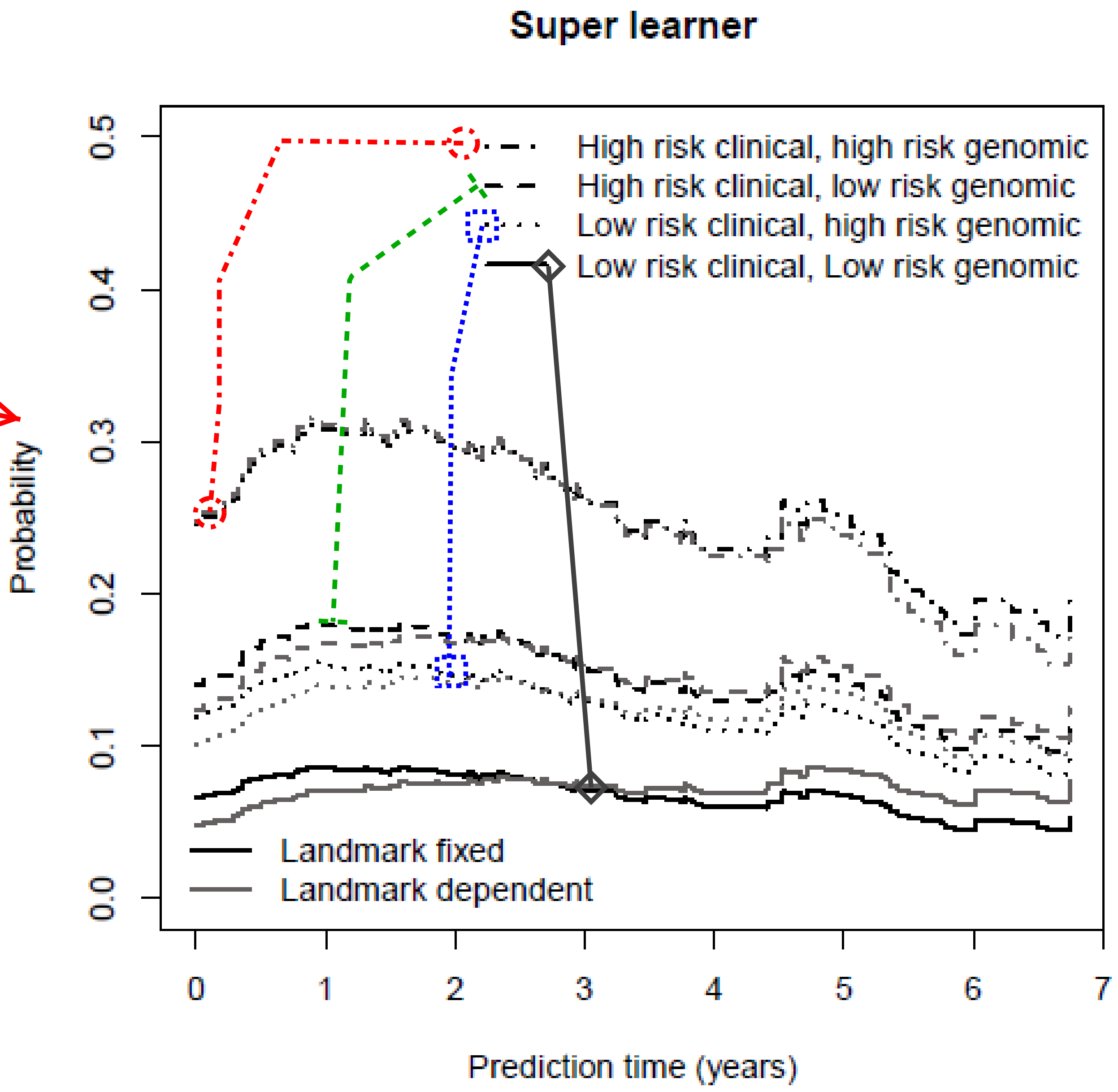
Model	Time	Clinical	Genomic	Super learner
Fixed	1	0.741 (0.122)	0.946 (0.155)	0.970 (0.138)
Landmark-	1	0.891 (0.155)	1.191 (0.209)	1.195 (0.181)
dependent	$s/7$	-0.412 (0.289)	-0.644 (0.368)	-0.595 (0.311)

5-year prediction.

“High” =mean+st.dev.

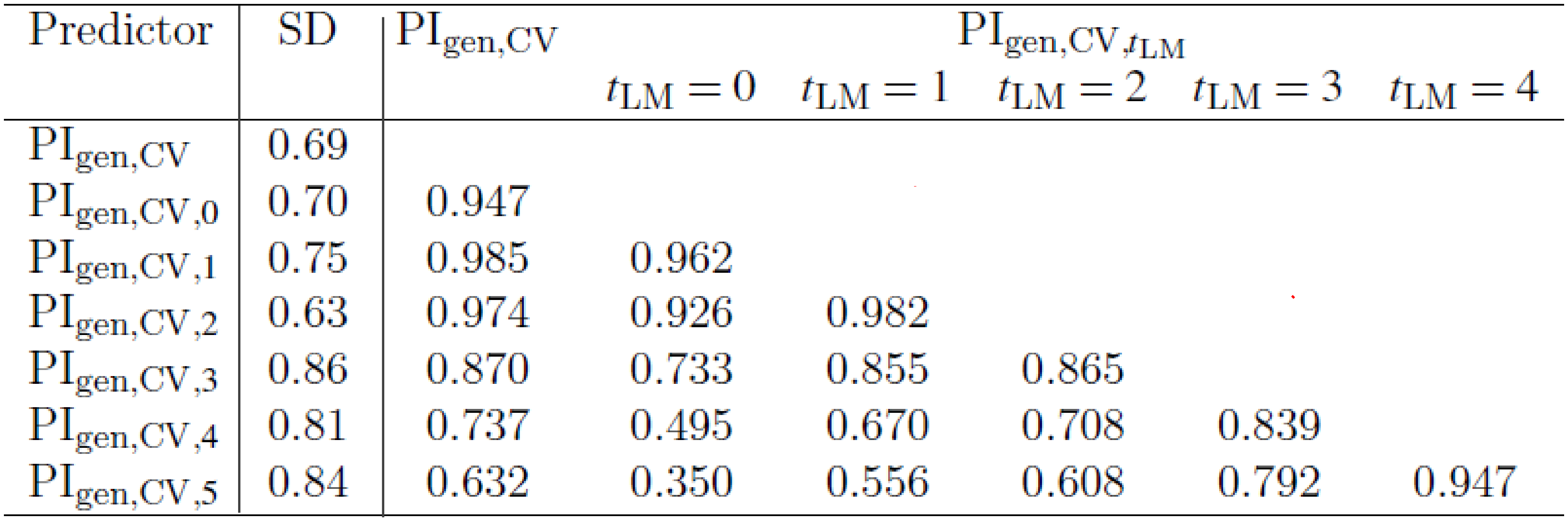
“Low” = mean-st.dev.

of death within next 5 years



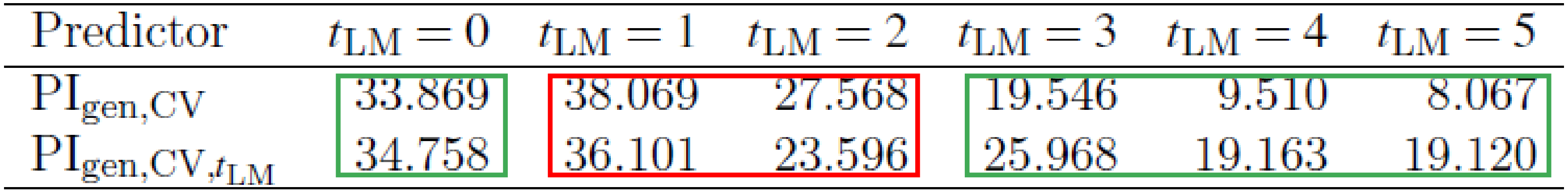
Could we do better?

Genetic predictors per landmark data set



Predictor	SD	$\mathrm{PI}_{\mathrm{gen,CV}}$	$\mathrm{PI}_{\mathrm{gen,CV},t_{\mathrm{LM}}}$				
			$t_{\mathrm{LM}}=0$	$t_{\mathrm{LM}}=1$	$t_{\mathrm{LM}}=2$	$t_{\mathrm{LM}}=3$	$t_{\mathrm{LM}}=4$
$\mathrm{PI}_{\mathrm{gen,CV}}$	0.69						
$\mathrm{PI}_{\mathrm{gen,CV},0}$	0.70	0.947					
$\mathrm{PI}_{\mathrm{gen,CV},1}$	0.75	0.985	0.962				
$\mathrm{PI}_{\mathrm{gen,CV},2}$	0.63	0.974	0.926	0.982			
$\mathrm{PI}_{\mathrm{gen,CV},3}$	0.86	0.870	0.733	0.855	0.865		
$\mathrm{PI}_{\mathrm{gen,CV},4}$	0.81	0.737	0.495	0.670	0.708	0.839	
$\mathrm{PI}_{\mathrm{gen,CV},5}$	0.84	0.632	0.350	0.556	0.608	0.792	0.947

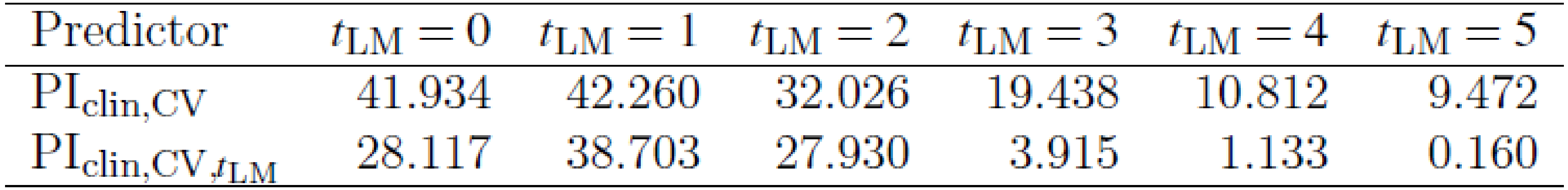
Table 12.4 *Standard deviations and correlations of cross-validated landmark specific genomic ridge predictors*



Predictor	$t_{\mathrm{LM}}=0$	$t_{\mathrm{LM}}=1$	$t_{\mathrm{LM}}=2$	$t_{\mathrm{LM}}=3$	$t_{\mathrm{LM}}=4$	$t_{\mathrm{LM}}=5$
$\mathrm{PI}_{\mathrm{gen,CV}}$	33.869	38.069	27.568	19.546	9.510	8.067
$\mathrm{PI}_{\mathrm{gen,CV},t_{\mathrm{LM}}}$	34.758	36.101	23.596	25.968	19.163	19.120

Table 12.5 *Comparison of model χ^2 for different approaches, using the genomic data*

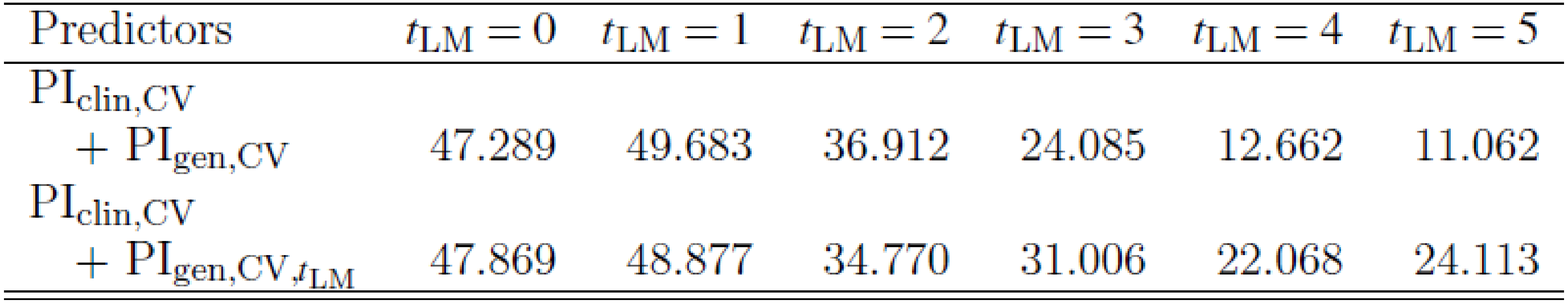
Degenerates for clinical predictor



Predictor	$t_{\mathrm{LM}}=0$	$t_{\mathrm{LM}}=1$	$t_{\mathrm{LM}}=2$	$t_{\mathrm{LM}}=3$	$t_{\mathrm{LM}}=4$	$t_{\mathrm{LM}}=5$
$\mathrm{PI}_{\mathrm{clin,CV}}$	41.934	42.260	32.026	19.438	10.812	9.472
$\mathrm{PI}_{\mathrm{clin,CV},t_{\mathrm{LM}}}$	28.117	38.703	27.930	3.915	1.133	0.160

Table 12.6 *Comparison of model χ^2 for different approaches, using the clinical data*

Combination of “adaptive genomic” and fixed “clinical” is fine.



Predictors	$t_{\mathrm{LM}}=0$	$t_{\mathrm{LM}}=1$	$t_{\mathrm{LM}}=2$	$t_{\mathrm{LM}}=3$	$t_{\mathrm{LM}}=4$	$t_{\mathrm{LM}}=5$
$\mathrm{PI}_{\mathrm{clin,CV}}$ $+\ \mathrm{PI}_{\mathrm{gen,CV}}$	47.289	49.683	36.912	24.085	12.662	11.062
$\mathrm{PI}_{\mathrm{clin,CV}}$ $+\ \mathrm{PI}_{\mathrm{gen,CV},t_{\mathrm{LM}}}$	47.869	48.877	34.770	31.006	22.068	24.113

Conclusion/discussion

Fixed model

- High-dimensional genomic data can be useful for prediction
- Lasso-versus-Ridge regression: pro’s and con’s
- Genomic does not beat Clinical

Dynamic model

- Effect of predictors changes over time
- Landmarking versus time-varying effects
- Genomic beats clinical later on in the follow up.
- Need for update clinical data (relapse, metastasis, etcetera)

References

- van de Vijver, M. J., He, Y. D., van `t Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, 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. & Bernards, R. (2002), **`A gene-expression signature as a predictor of survival in breast cancer'**, New England Journal of Medicine 347, 1999-2009
- Goeman, J. J., Oosting, J., Cleton-Jansen, A.-M., Anninga, J. K. & van Houwelingen, H. C. (2005), **`Testing association of a pathway with survival using gene expression data'**, Bioinformatics 21, 1950-1957.
- van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van't Veer, L. J. & Wessels, L. F. A. (2006), **`Cross-validated Cox regression on microarray gene expression data'**, Statistics in Medicine 25, 3201-3216.
- Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A. & Lingjærde, O. C. (2007), **`Predicting survival from microarray data -a comparative study'**, Bioinformatics 23, 2080-2087.

- van Houwelingen, H. C. (2007), **`Dynamic prediction by landmarking in event history analysis'**, Scandinavian Journal of Statistics 34, 70-85.
- Goeman, J. J. (2010), **`L_1 penalized estimation in the Cox proportional hazards model'**, Biometrical Journal 52, 70-84.
- Bøvelstad, H. M., Nygård, S. & Borgan, Ø., (2009), **`Survival prediction from clinico-genomic models - a comparative study'**, BMC Bioinformatics 10, art. no. 413.
- van Houwelingen, H.C. & Putter, H., (2011), **‘Dynamic prediction in clinical survival analysis’**, CRC-Chapman & Hall