

Assessment and validation of risk prediction models

Martin Schumacher, Harald Binder and Thomas Gerds

`ms@imbi.uni-freiburg.de`

Department of Medical Biometry and Statistics
Institute of Medical Biometry and Medical Informatics
University Hospital Freiburg, Germany

Outline

- Risk prediction models for time-to-event data
- Measures of prediction error
- Bootstrap cross-validation and the $.632+$ estimator
- Illustration with breast cancer study
- Application in high-dimensional settings
- Discussion and conclusions

What is a risk prediction model?

- A rule $r(t|Z)$ that gives a predicted probability that an individual will be free of the event of interest up to time $t > 0$, given covariate information Z available at $t = 0$.
- Examples:
 - Gail model (Development of breast cancer)
 - SAPS II, APACHE (Mortality on intensive care unit)
 - Prognostic models for various cancer entities and other chronic diseases

How is a risk prediction model developed?

- Ideally, given externally
- Usually, by standard statistical methodology (e.g. Cox regression models)
- Often, by flexible statistical model building approaches (e.g. classification and regression trees, flexible regression models, artificial neural nets (ANNs), other machine learning techniques)
- Sometimes, by expert guesses

Assessment of risk prediction models

- In the same data in that it was developed (“Apparent-Error”-problem): potential overfitting leads to overoptimism of prediction error.
- In independent data
 - with different structure (external validation)
 - with identical or similar structure (internal validation).
- Dilemma: Use as much information for model development versus Obtain a reliable estimate of prediction error.
- Possible solutions: Data splitting (training and test set), Bootstrap cross-validation

Situation

- $\mathbf{X} = \{X_1, \dots, X_n\}$ right censored time-to-event data
with $X = (\tilde{T}, \Delta, Z) \quad iid \sim Q$
and $\tilde{T} = \min(T, C)$, $\Delta = 1\{T \leq C\}$, Z vector of covariates (of dimension p)
- \mathbf{X} is often called the "training set" represented through the empirical measure Q_n
- $r_n = r(X_1, \dots, X_n)$ risk prediction model developed ("trained") in \mathbf{X}

Example: Standard Cox regression model

Model: $\lambda(t | Z) = \lambda_0(t) \exp(\beta^T Z)$

Prediction: $r_n^{Cox}(t|Z) = \exp\left\{ -\hat{\Lambda}_0(t) \exp(\hat{\beta}^T Z) \right\}$

where $\hat{\beta}$ is the maximum partial likelihood estimators of the vector of regression coefficients β_1, \dots, β_p and $\hat{\Lambda}_0(t)$ denotes the Breslow estimator of the cumulative baseline hazard

$$\Lambda_0(t) = \int_0^t \lambda_0(s) ds, \text{ respectively.}$$

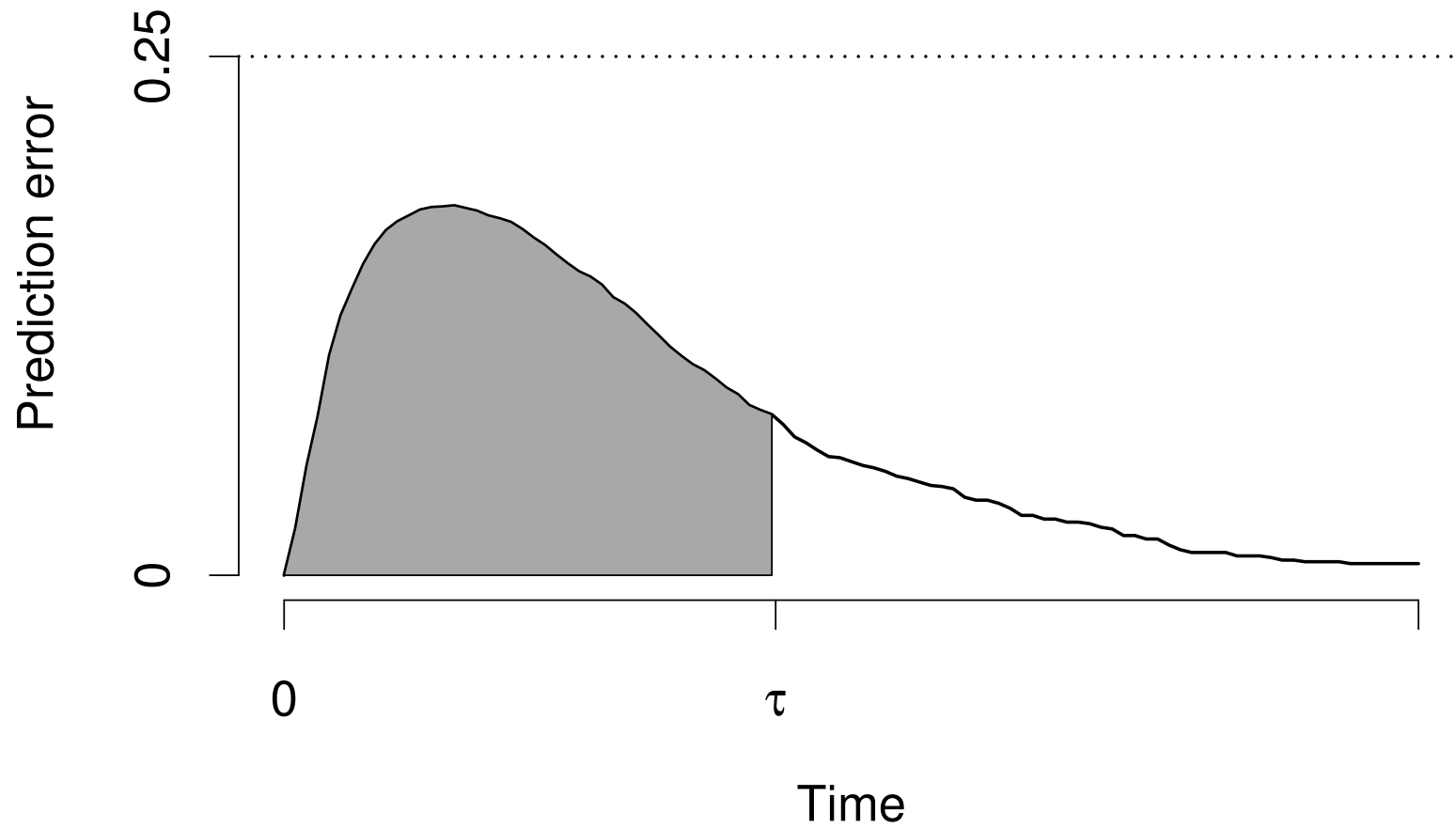
Prediction error (1):

- Event status $Y(t) = 1\{T > t\}$
- Squared residuals $\{Y(t) - r_n(t|Z)\}^2$
- "True prediction error curve" of r_n :

$$Err(t; r_n, Q_n) = E [\{Y(t) - r_n(t|Z)\}^2 | Q_n]$$

- $Err(\)$ is a random quantity ("conditional prediction error"); its expected value with regard to all possible training sets of size n is called "Expected true prediction error".

Prediction error curve



Prediction error (2):

- The prediction error curve can be estimated by

$$\overline{err}(t, r_n) = \frac{1}{n} \sum_{i=1}^n \left\{ Y_i(t) - r_n(t|Z_i) \right\}^2 W(t, \hat{G}_n, X_i)$$

- Since $Y(t)$ is observable up to \tilde{T} for censored observations inverse probability of censoring (IPC)-weighting scheme has to be used given by

$$W(t, \hat{G}_n, X_i) = \frac{1\{\tilde{T}_i \leq t\} \Delta_i}{\hat{G}_n(\tilde{T}_i - |Z_i)} + \frac{1\{\tilde{T}_i > t\}}{\hat{G}_n(t|Z_i)}$$

where $\hat{G}_n(t|Z)$ is a consistent estimate of $P(C > t|Z)$

Prediction error (3):

Marginal censoring model

$\hat{G}_n(t)$ Kaplan-Meier estimator of censoring survival function $P(C > t)$; ignores covariates (Graf et al., Stat Med 1999).

Conditional censoring model

$\hat{G}_n(t | Z)$ model-based estimator of censoring survival function $P(C > t | Z)$; models potential dependence of censoring on covariates. Possible choices are Cox or Aalen additive regression models or nonparametric regression (Gerds et al., Biom J 2006).

Resampling estimators of prediction error:

- Data splitting (training and test set)
- Leave-one-out cross-validation
- K-fold cross-validation
- Bootstrap cross-validation
- Improving on cross-validation ($.632$ and $.632+$ estimator)

Bootstrap-cross-validation (1):

- Draw B bootstrap samples \mathbf{X}_b^* each of size n from \mathbf{X} (with replacement)

- Define $\mathbf{X}_b^0 = \{X_i \notin \mathbf{X}_b^*\}$ and $b_0 = |\mathbf{X}_b^0|$

- Develop a risk prediction model r_b^* on \mathbf{X}_b^*

- Calculate

$$\widehat{Err}_{B0}(t, r_n) = \frac{1}{B} \sum_{b=1}^B \frac{1}{b_0} \sum_{i \in \mathbf{X}_b^0} \left\{ Y_i(t) - r_b^*(t|Z_i) \right\}^2 W(t, \hat{G}_n, X_i)$$

Bootstrap-cross-validation (2):

- $\overline{err}(\)$ tends to underestimate the true prediction error (“too optimistic”)
- $\widehat{Err}_{B0}(\)$ tends to overestimate the true prediction error (“too pessimistic”)

- A linear combination should do the job

$$\widehat{Err}_{\omega}(t, r_n) = \{1 - \omega(t)\} \overline{err}(t, r_n) + \omega(t) \widehat{Err}_{B0}(t, r_n)$$

- $\omega(t) = .632$ gives the famous .632 estimator (Efron, 1983)

The .632+ estimator (1)

- Rationale: $\omega(t)$ should reflect the amount of overfitting and should also be allowed to vary over time
- "Worst case" scenario: event status and covariates are independent

$$NoInf(t, r_n) = \frac{1}{n^2} \sum_{j=1}^n \sum_{i=1}^n \left\{ Y_i(t) - r_n(t|Z_j) \right\}^2 W(t, \hat{G}_n, X_i)$$

- Relative overfitting rate:

$$\hat{R}(t) = \frac{\widehat{Err}_{B0}(t, r_n) - \overline{err}(t, r_n)}{NoInf(t, r_n) - \overline{err}(t, r_n)}$$

The .632+ estimator (2)

- The time-dependent version of the .632+ estimator is defined by

$$\hat{\omega}(t) = .632 / (1 - .368 \hat{R}(t)) \quad (\text{Efron \& Tibshirani, 1997})$$

- Remember:

$$\widehat{Err}_{.632+}(t, r_n) = \{1 - \hat{\omega}(t)\} \overline{err}(t, r_n) + \hat{\omega}(t) \widehat{Err}_{B0}(t, r_n)$$

- $\hat{R}(t) \approx 0$, then $\widehat{Err}_{.632+}(\) \approx \widehat{Err}_{.632}(\)$
- $\hat{R}(t) \approx 1$, then $\widehat{Err}_{.632+}(\) \approx \widehat{Err}_{B0}(\)$

Illustration with breast cancer study (1)

- Prospective controlled clinical trial on treatment of primary node positive breast cancer (GBSG-2 study)
- Comprehensive cohort study where randomized and non-randomized patients were included and followed-up
- For 686 patients, complete data on six prognostic factors (age, tumor size, tumor grade, no. of involved lymph nodes, estrogen receptor, progesterone receptor) were available
- "Standard analysis" with a prespecified Cox regression model

Illustration with breast cancer study (2)

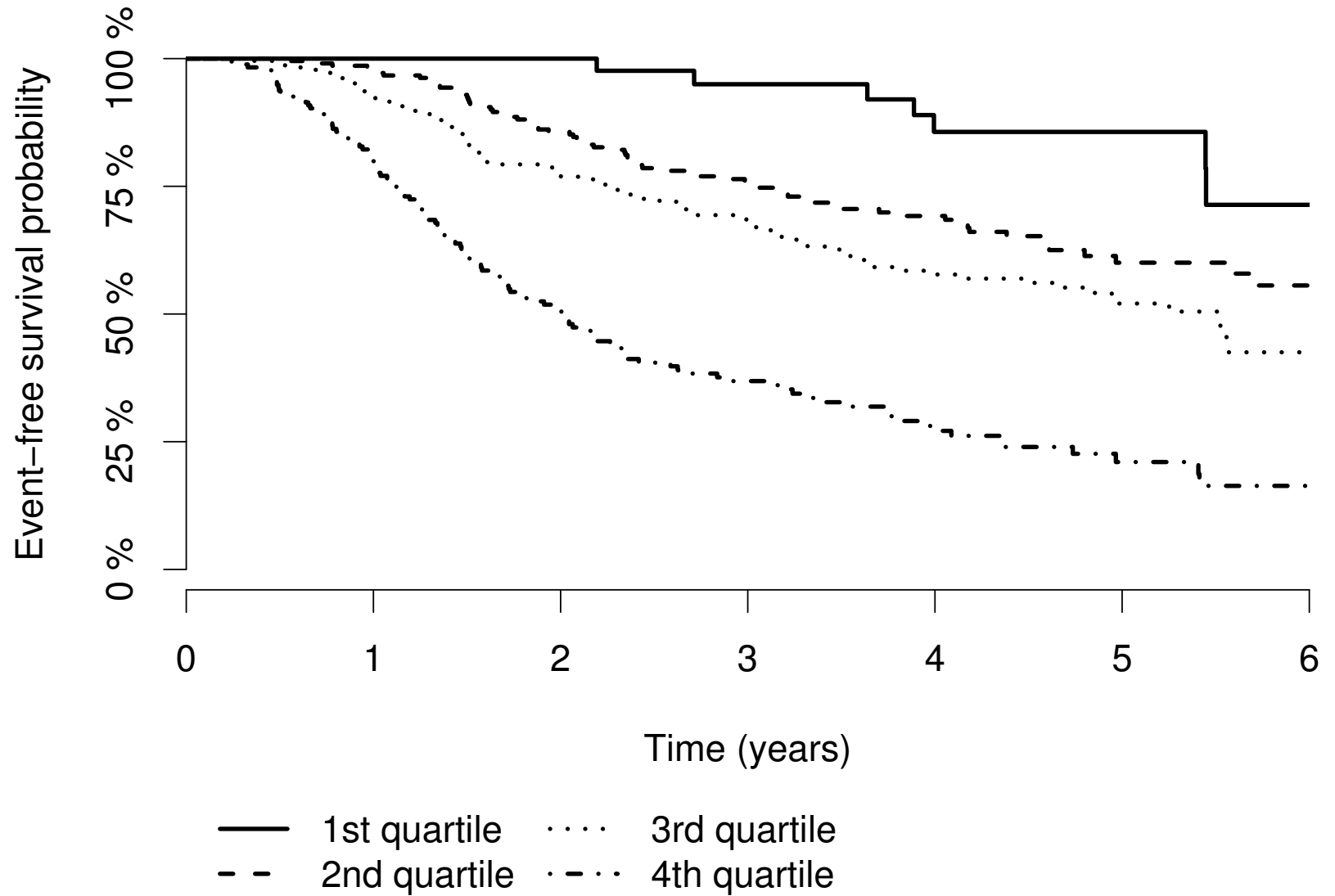


Illustration with breast cancer study (3)

- A variety risk of prediction models is created that differ in their degree of flexibility
 - "Standard analysis" with a prespecified Cox regression model
 - Multivariate fractional polynomials (MFP) with selection of functional form
 - Faraggi-Simon type artificial neural nets (ANNs) for censored survival data with and without weight decay
 - (Classification and regression trees with stringent and less stringent stopping criteria, leading to few or many terminal nodes, respectively)

Illustration with breast cancer study (4)

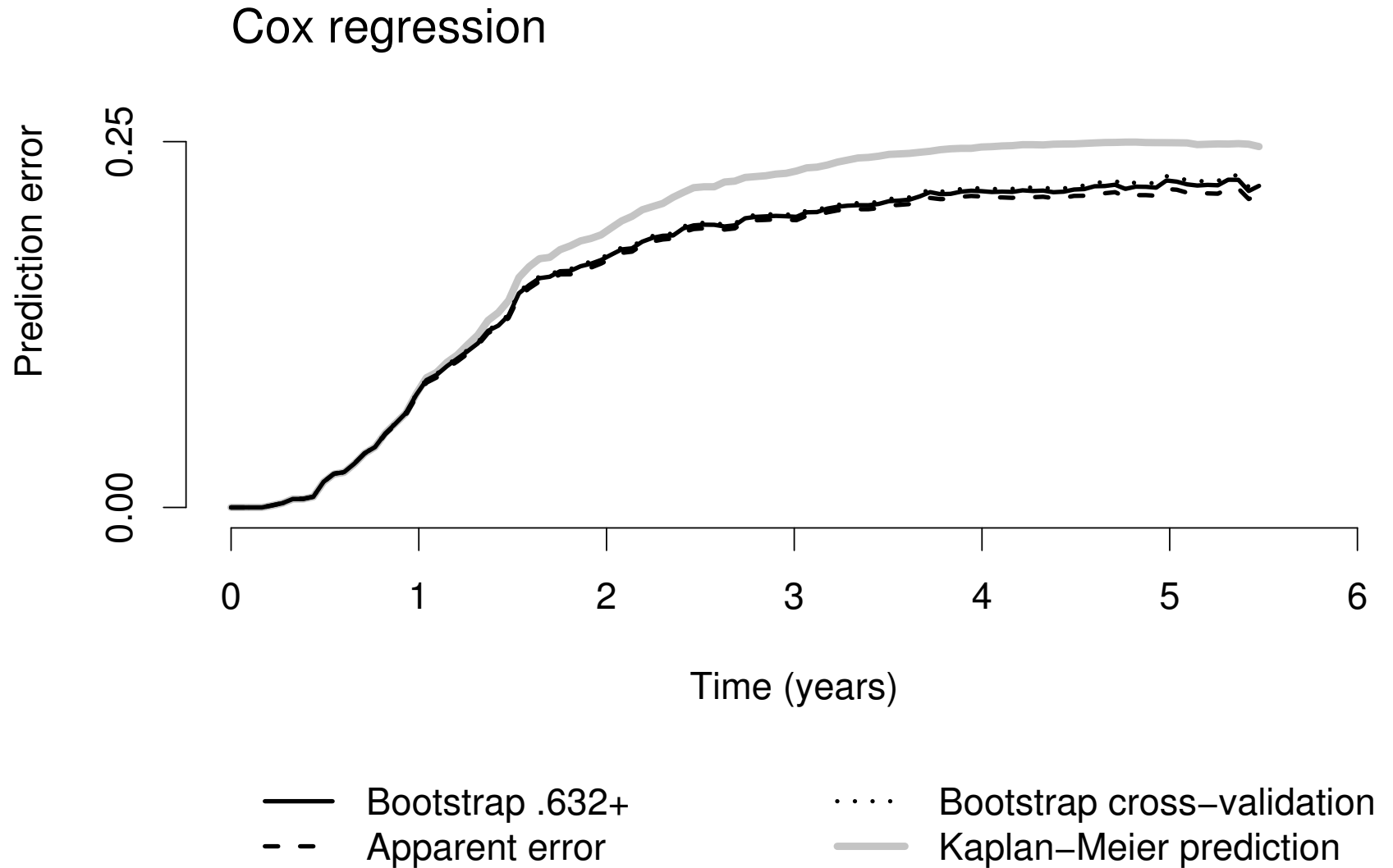


Illustration with breast cancer study (5)

Multivariate fractional polynomials

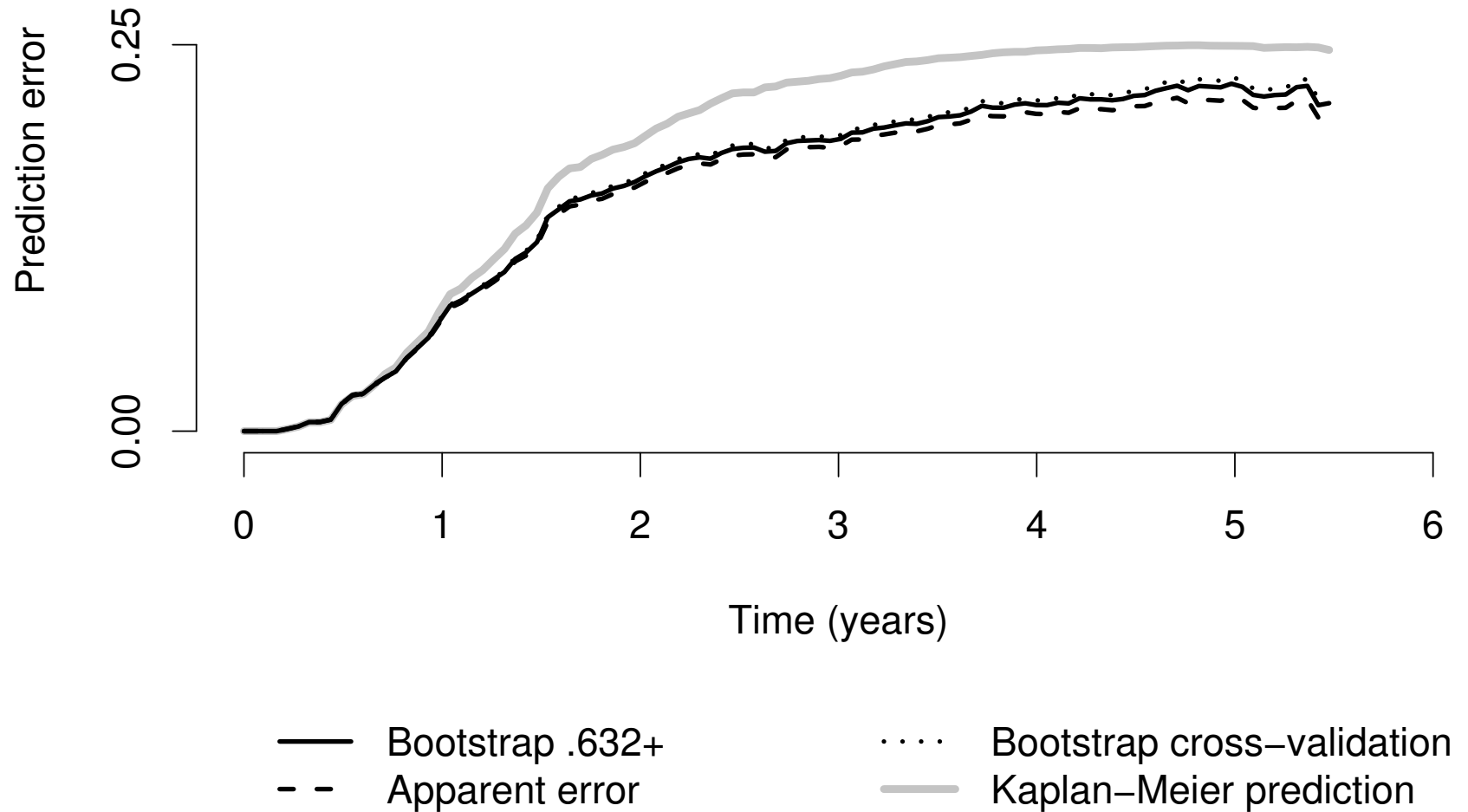


Illustration with breast cancer study (6)

Artificial neural net decay=0

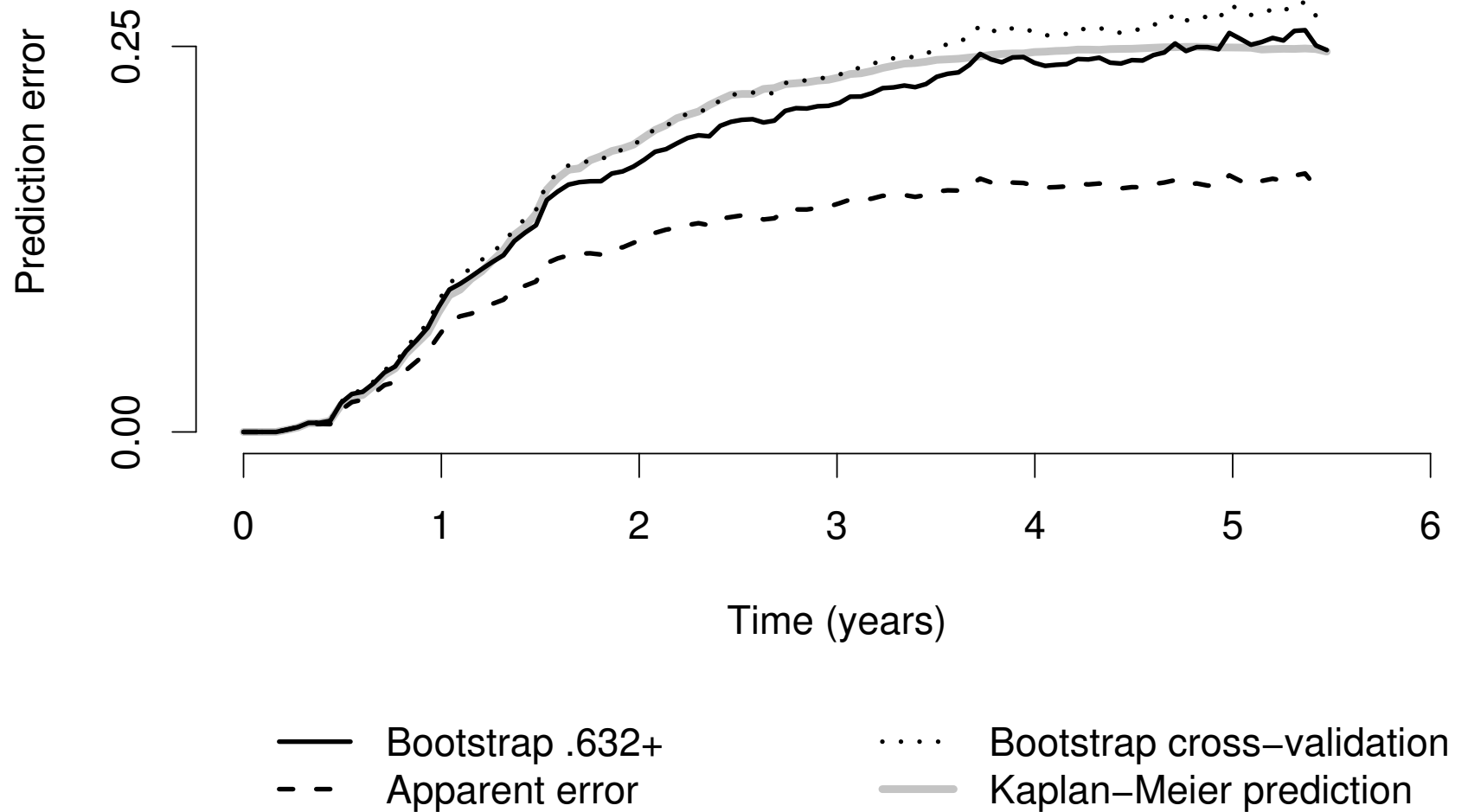


Illustration with breast cancer study (7)

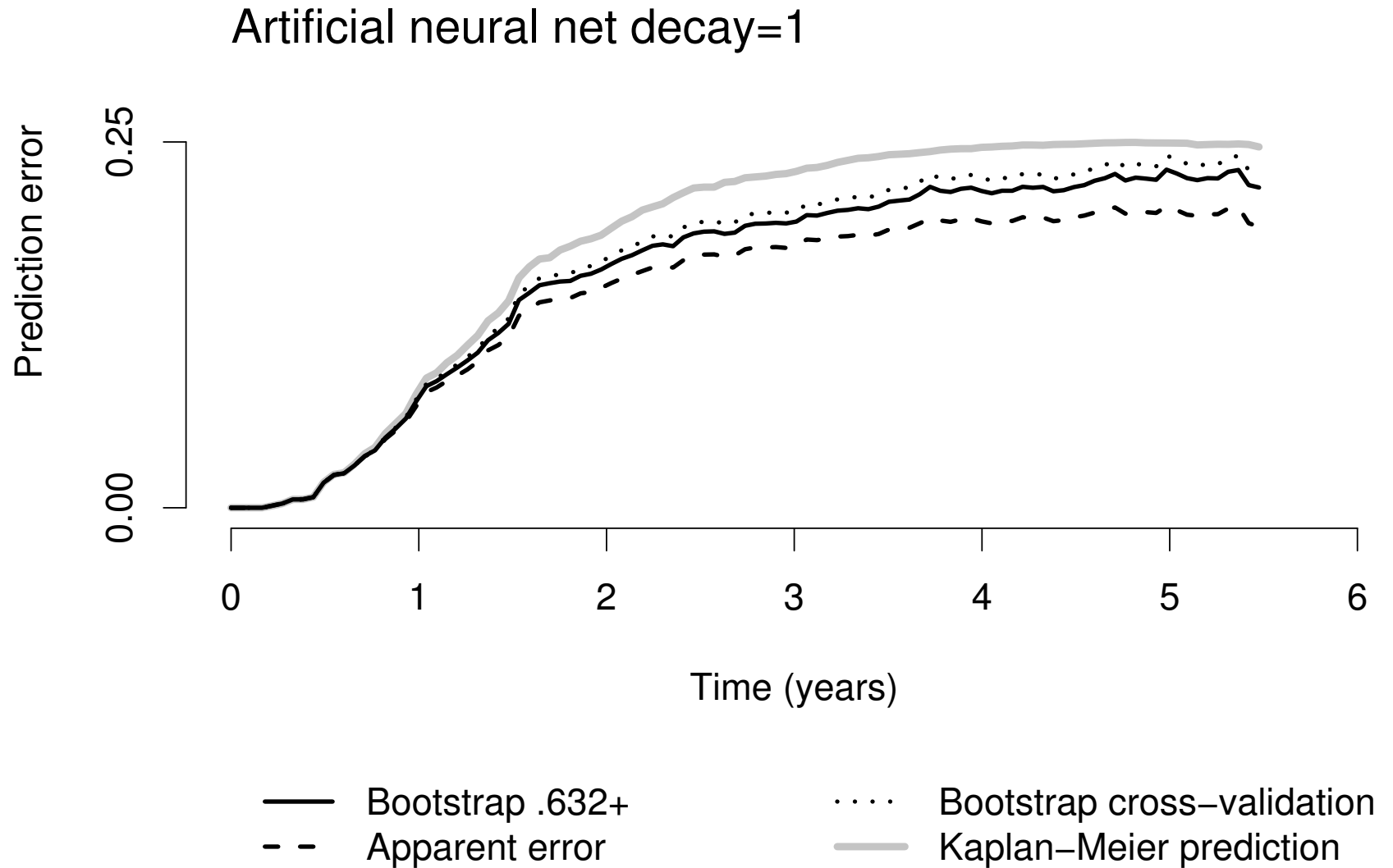


Illustration with breast cancer study (8)

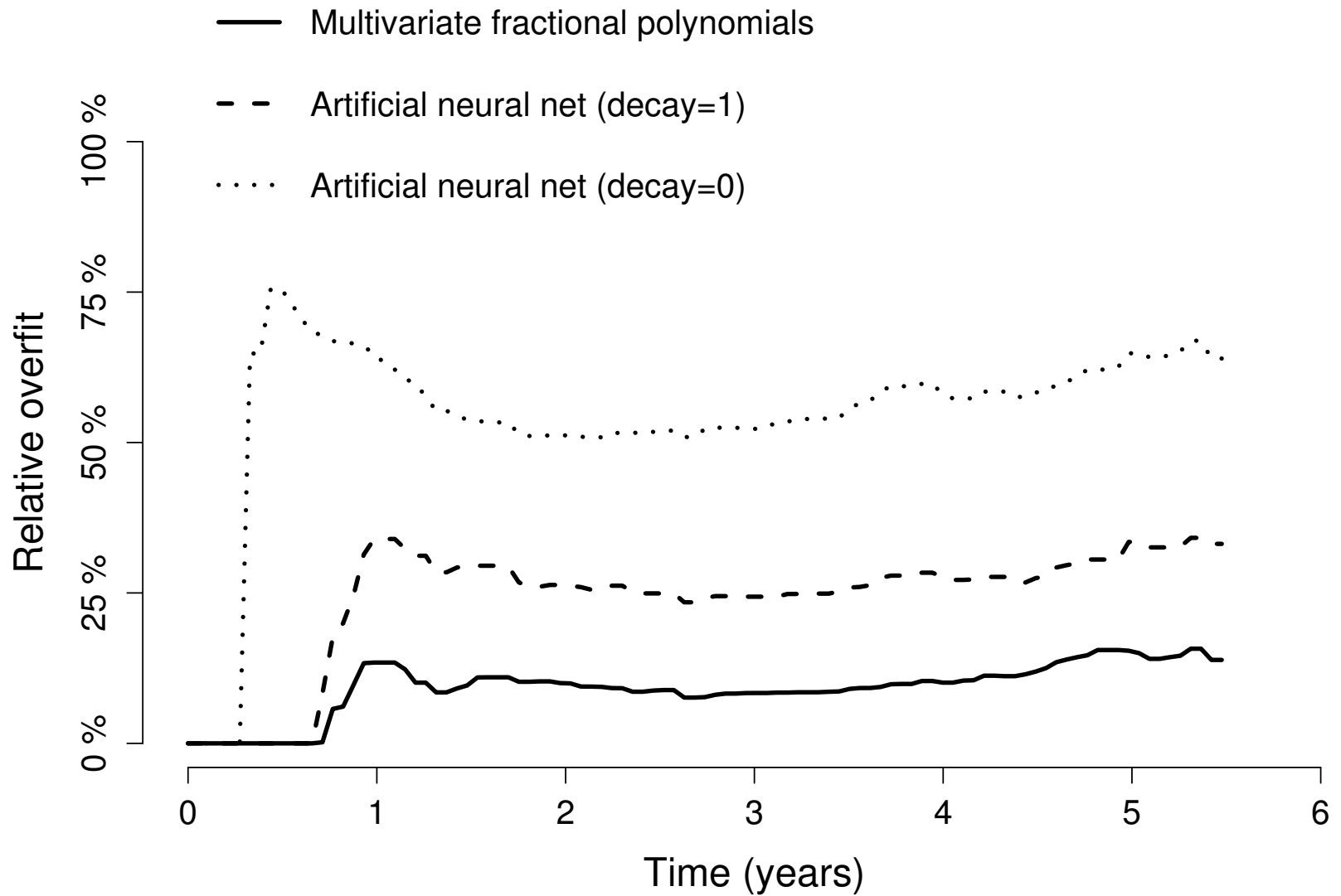


Illustration with breast cancer study (9)

Estimates of prediction error in the GBSG-2 study (Area under the prediction error curve in the interval 0 – 5.5 years)

Risk prediction model	\overline{err}	\widehat{Err}_{B0}	$\widehat{Err}_{.632+}$
Kaplan Meier	.179	.180	.179
Cox regression	.157	.161	.159
Multivariate fractional polynomial	.150	.157	.154
ANN, decay = 0	.118	.187	.173
ANN, decay = 1	.144	.164	.158

Application in high-dimensional settings (1)

- Generalization of $.632+$ estimator seems to work in a well-behaved "classical" situation ($n = 686$, 299 events, $p = 6$)
- But what's about high dimensions $p \gg n$?
- Serious overfitting may be overwhelming!
- Usual approach: Splitting the small (with respect to n) data set into training and test set!

Rosenwald DLBCL study (1)

- Prognostic study (Rosenwald et al., 2002) with retrospective collection of tumor-biopsy specimens and clinical data in 240 patients with untreated diffuse large-B-cell lymphoma (DLBCL); 138 deaths, 5-year OS: 48%
- Lymphochip cDNA microarray technology with $p = 7399$ "genes" was applied
- Data set was splitted in training set ($n = 160$) and test set ($n = 80$)
- Various attempts to create predictive models of survival (from time of chemotherapy) in DLBCL patients have been published (Comprehensive review: Segal, Biostatistics 2006)

Application in high-dimensional settings (2):

Recent investigation (Molinaro et al., Bioinformatics 2005)

- Simulation scenario: Classification problem with $N = 300$; 150 cases ($Y = 1$), 150 controls ($Y = 0$); $p = 750$ covariates ("genes"); only 8 exhibit an effect $\neq 0$
- Prediction models (LDA, DDA, NN, CART) are developed after "feature" selection (10 "genes" with largest absolute values of t -statistics).
- Training set: $n = 40$, $n = 80$, $n = 120$; remaining observations are used for estimating the true prediction error (misclassification rate).

Application in high-dimensional settings (3):

Molinaro et al. study: Authors' main conclusions:

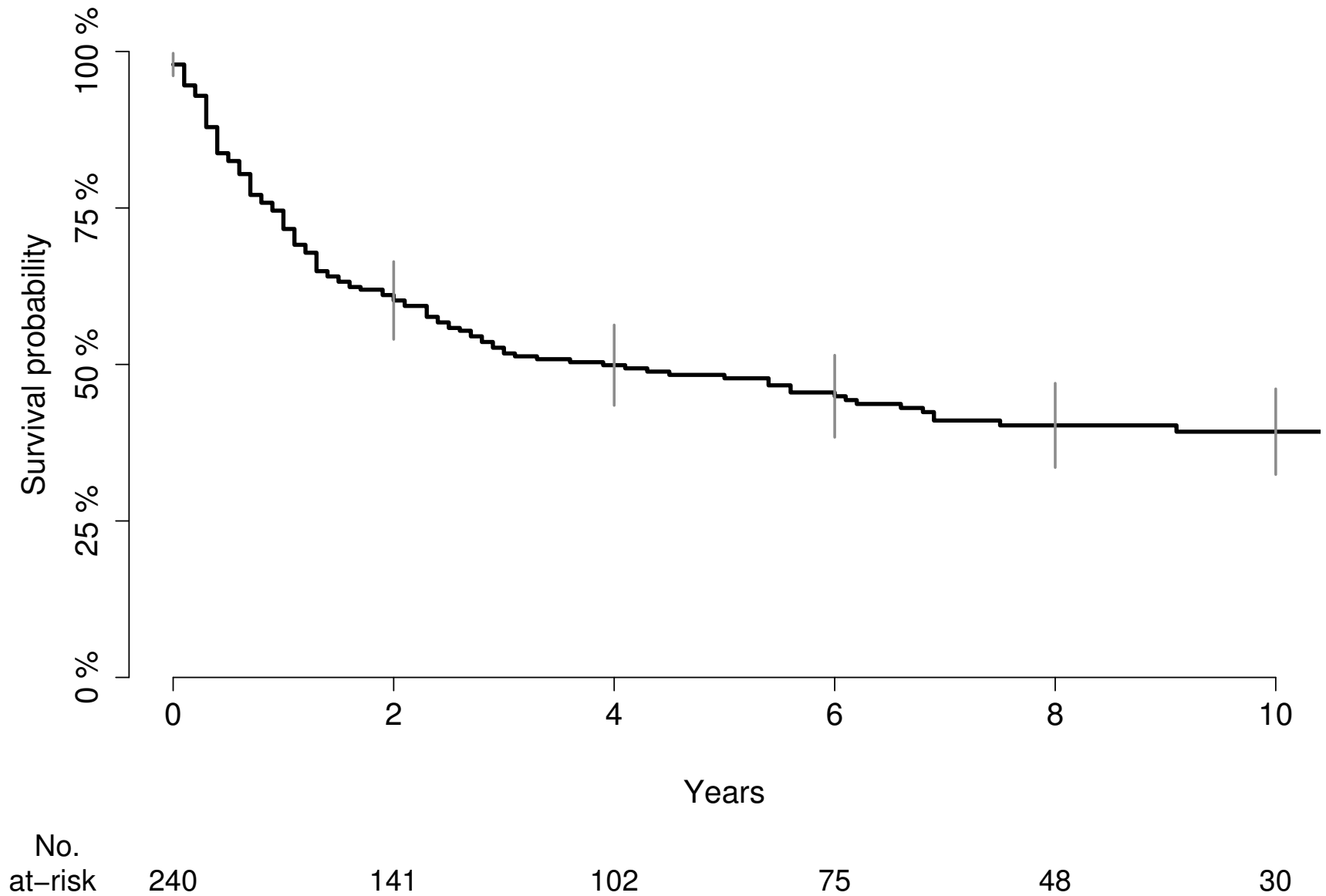
- Leave-one-out cross validation generally performs very well with regard to MSE and bias, the only exception is when a flexible prediction model is used (large variability!)
- The .632+ estimator performs best with moderate to weak effects; some difficulties with "ties" in the small sample situation ($n = 40$)
- Discrepancies between estimators fade when "feature" selection is discarded and when effects decrease
- Differences among estimators diminish as sample size grows ($n = 120$)

Rosenwald DLBCL study (2)

Various risk prediction models are created according to recent proposals

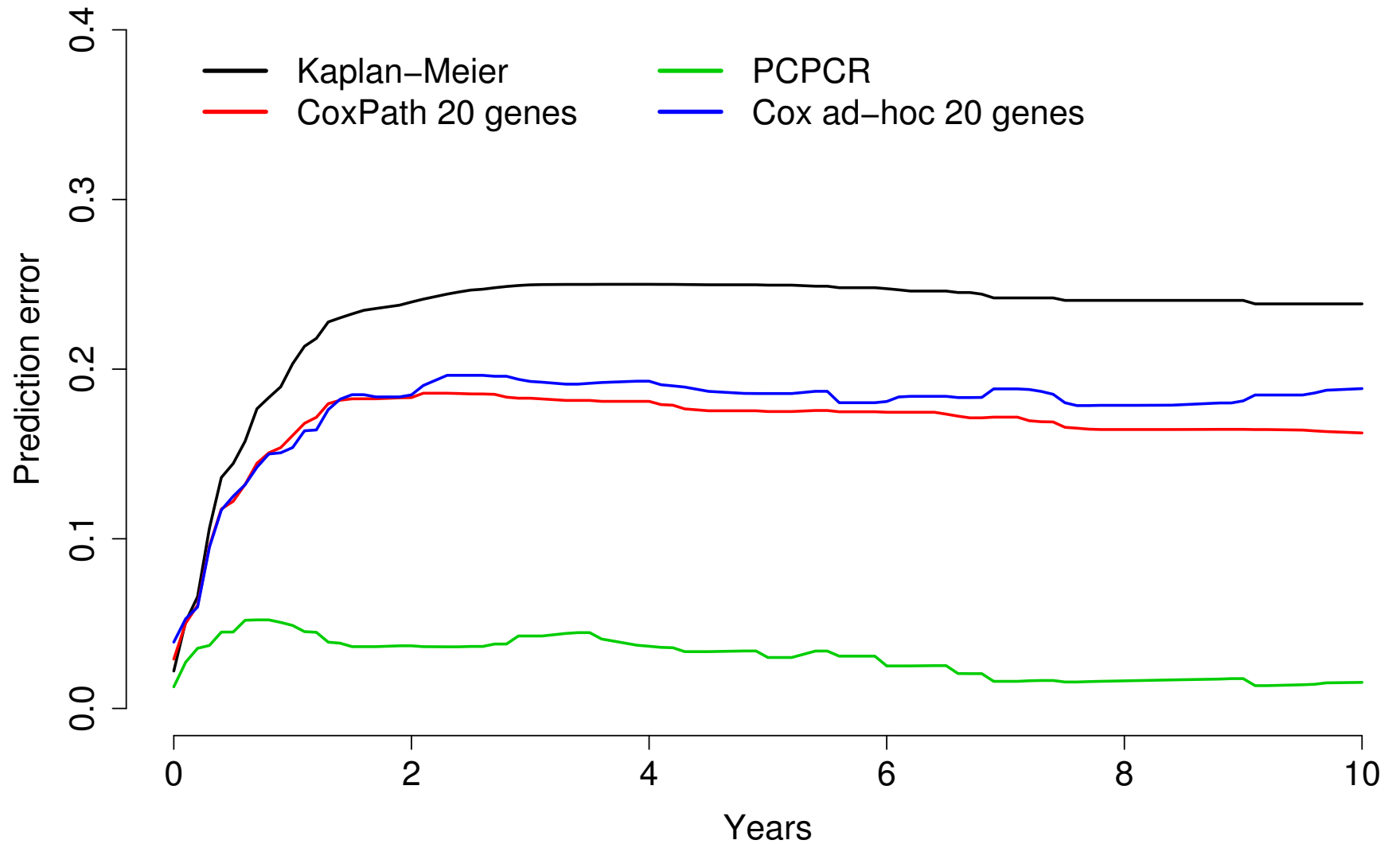
- Partial Cox regression with prior principal component analysis on the gene expression matrix (PCPCR; Li & Gui, Bioinformatics, 2004)
- "Lasso-type" Cox regression with L_1 regularization path algorithm (CoxPath; Park & Hastie, Stanford Techn Rep 2006) using 20 genes
- Cox regression using the ten (twenty) most significant "genes" obtained from univariate analyses (Cox best 10, Cox best 20)

Rosenwald DLBCL study (3)



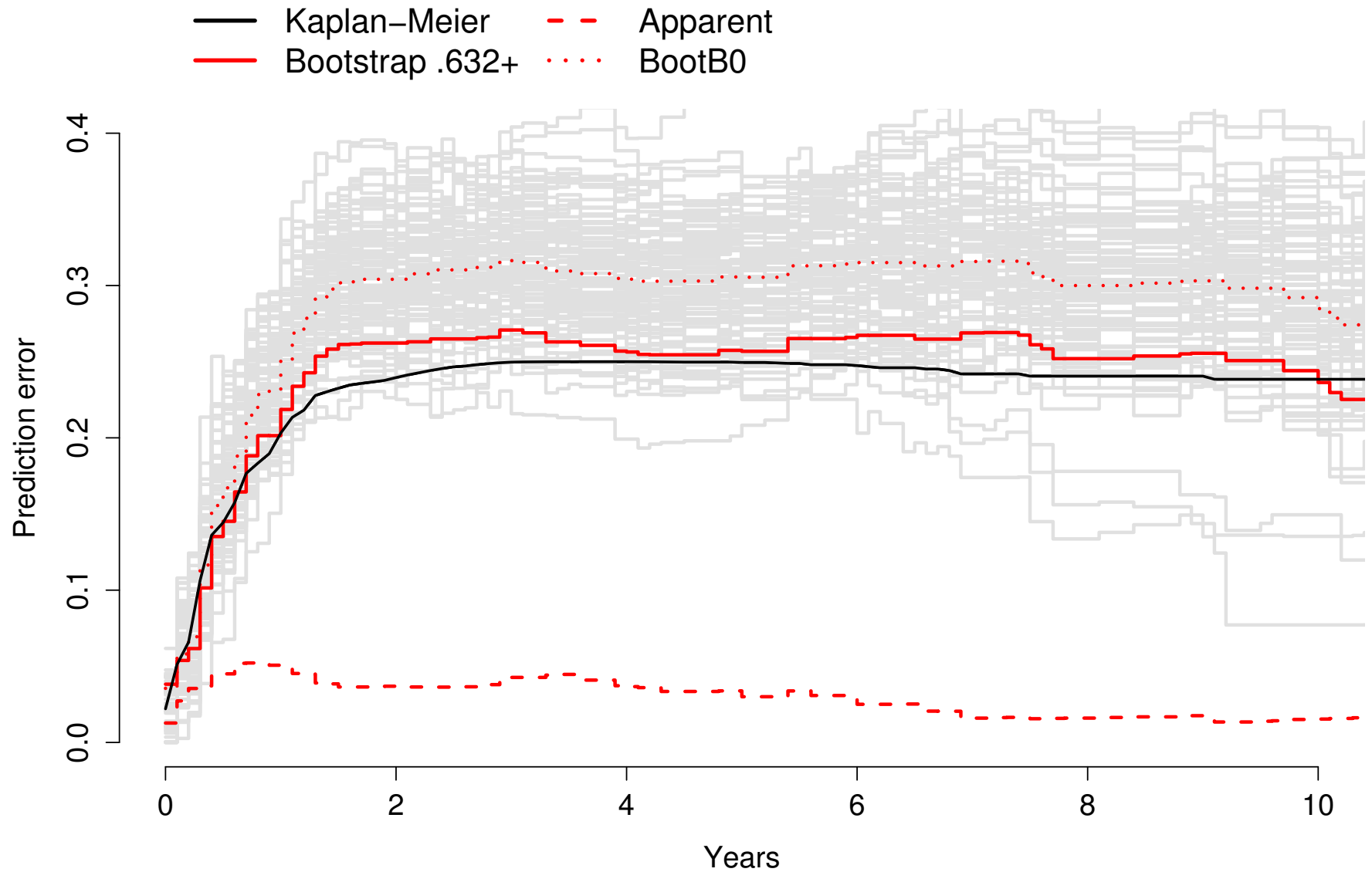
Rosenwald DLBCL study (4)

Apparent error



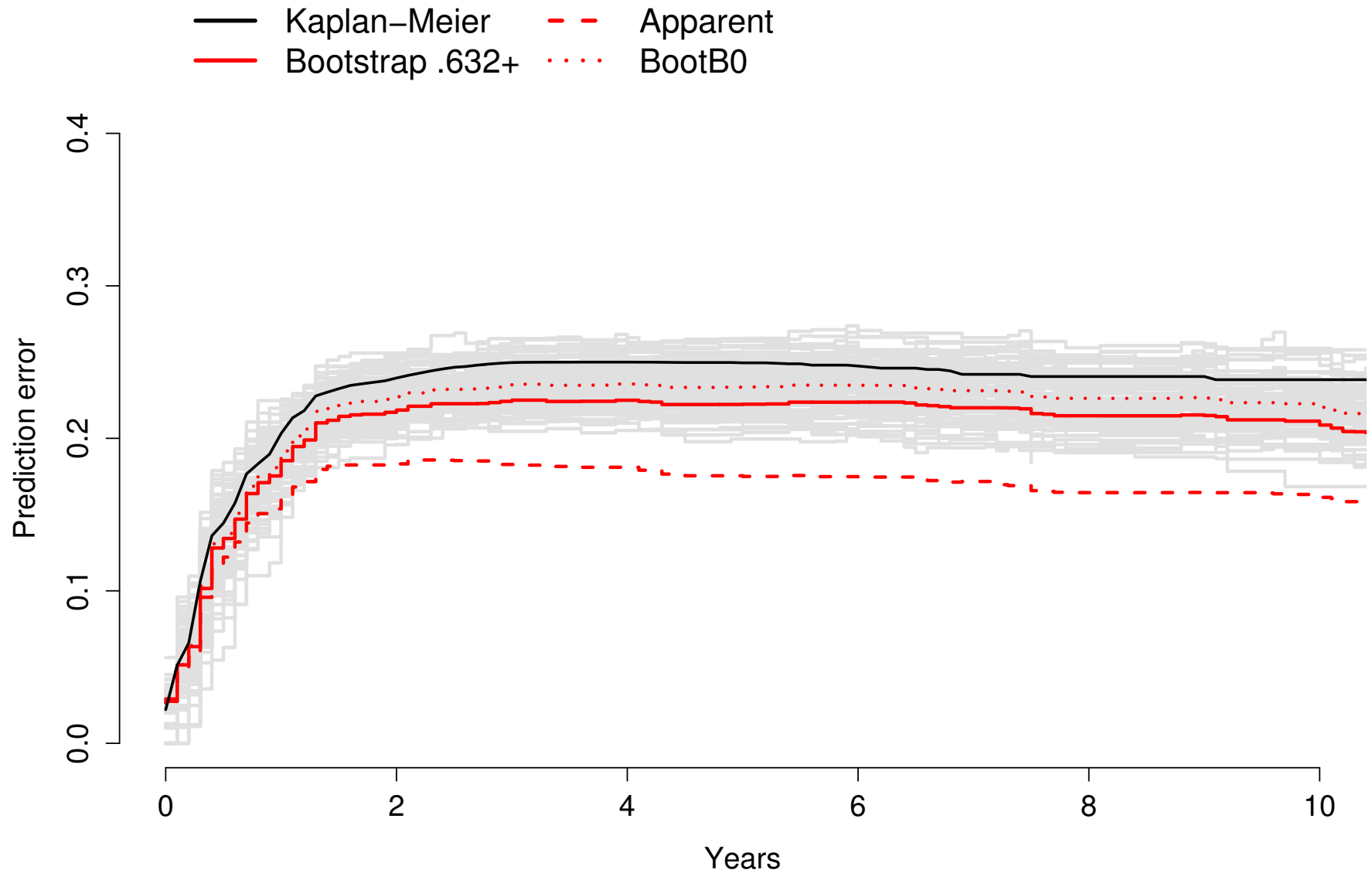
Rosenwald DLBCL study (5)

PCPCR



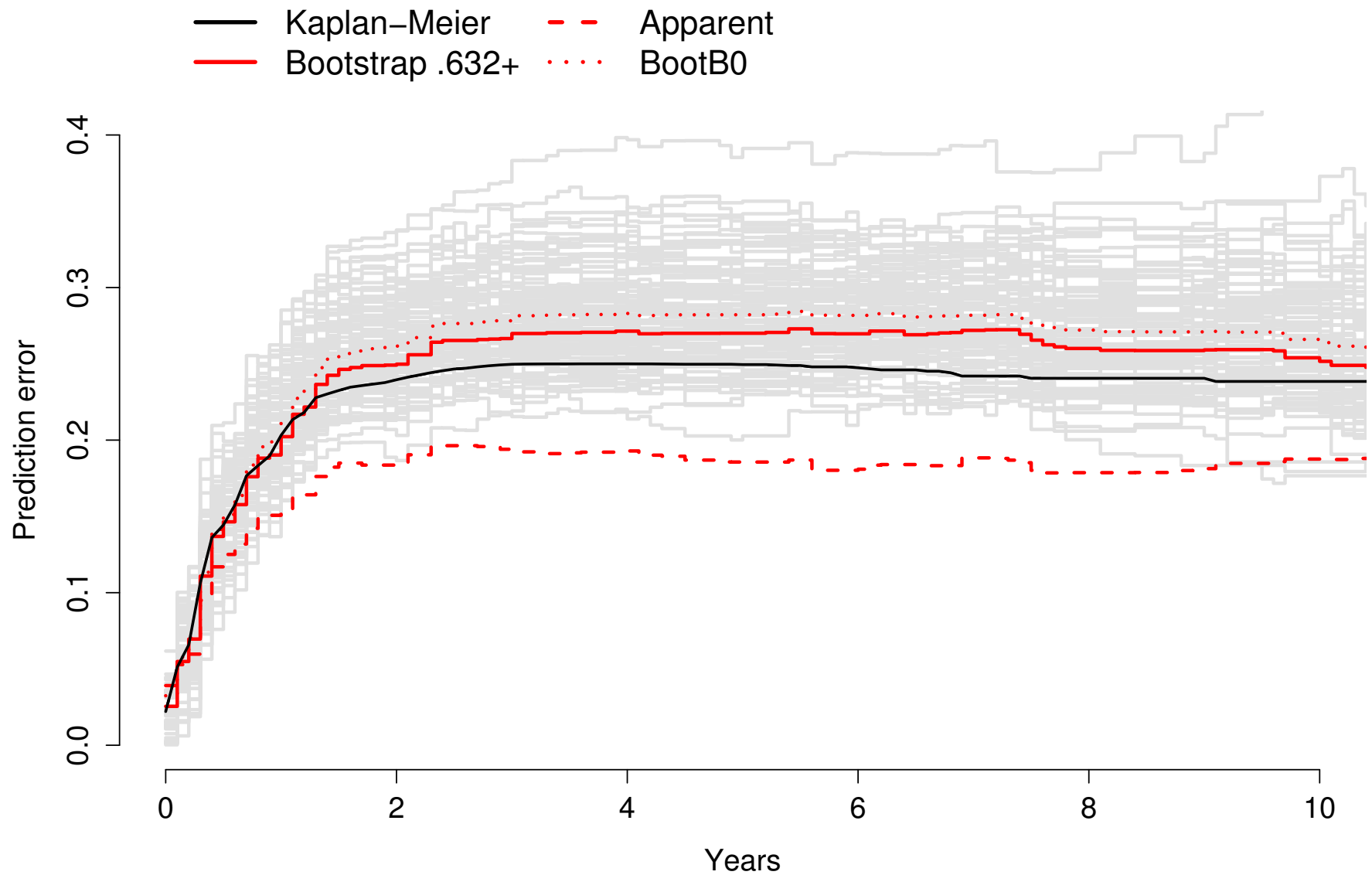
Rosenwald DLBCL study (6)

Coxpath: 20 genes

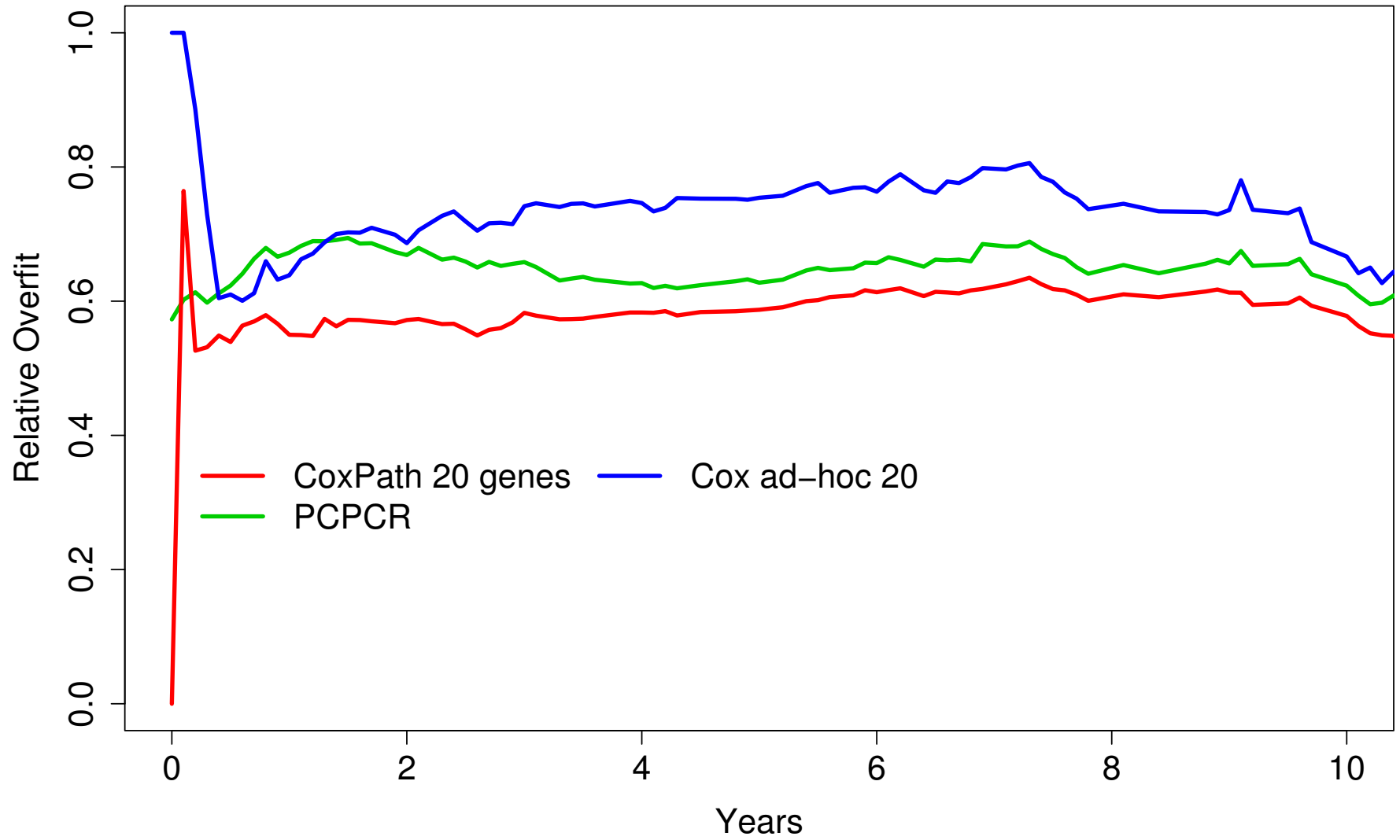


Rosenwald DLBCL study (7)

Cox: adhoc best 20 genes

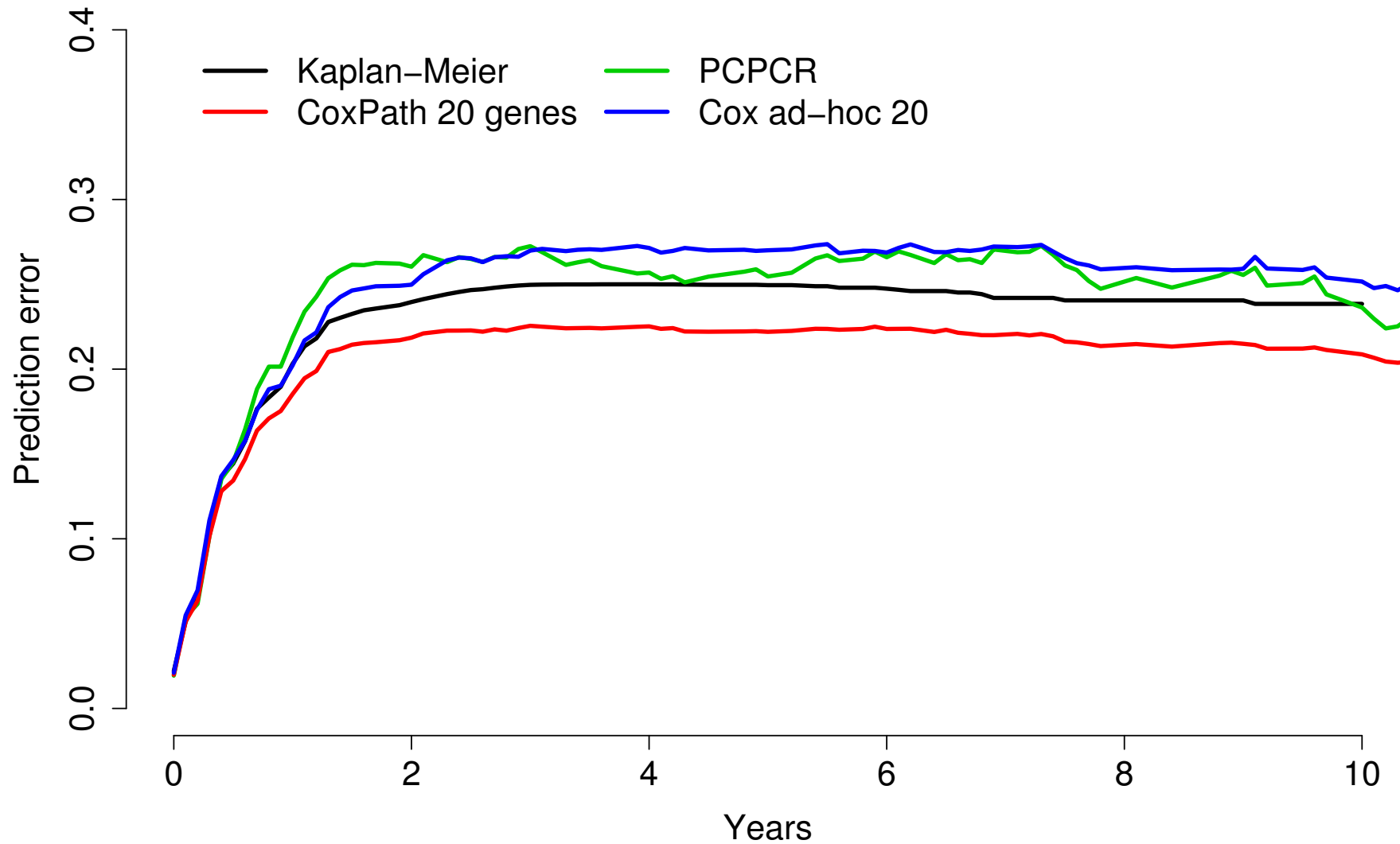


Rosenwald DLBCL study (8)



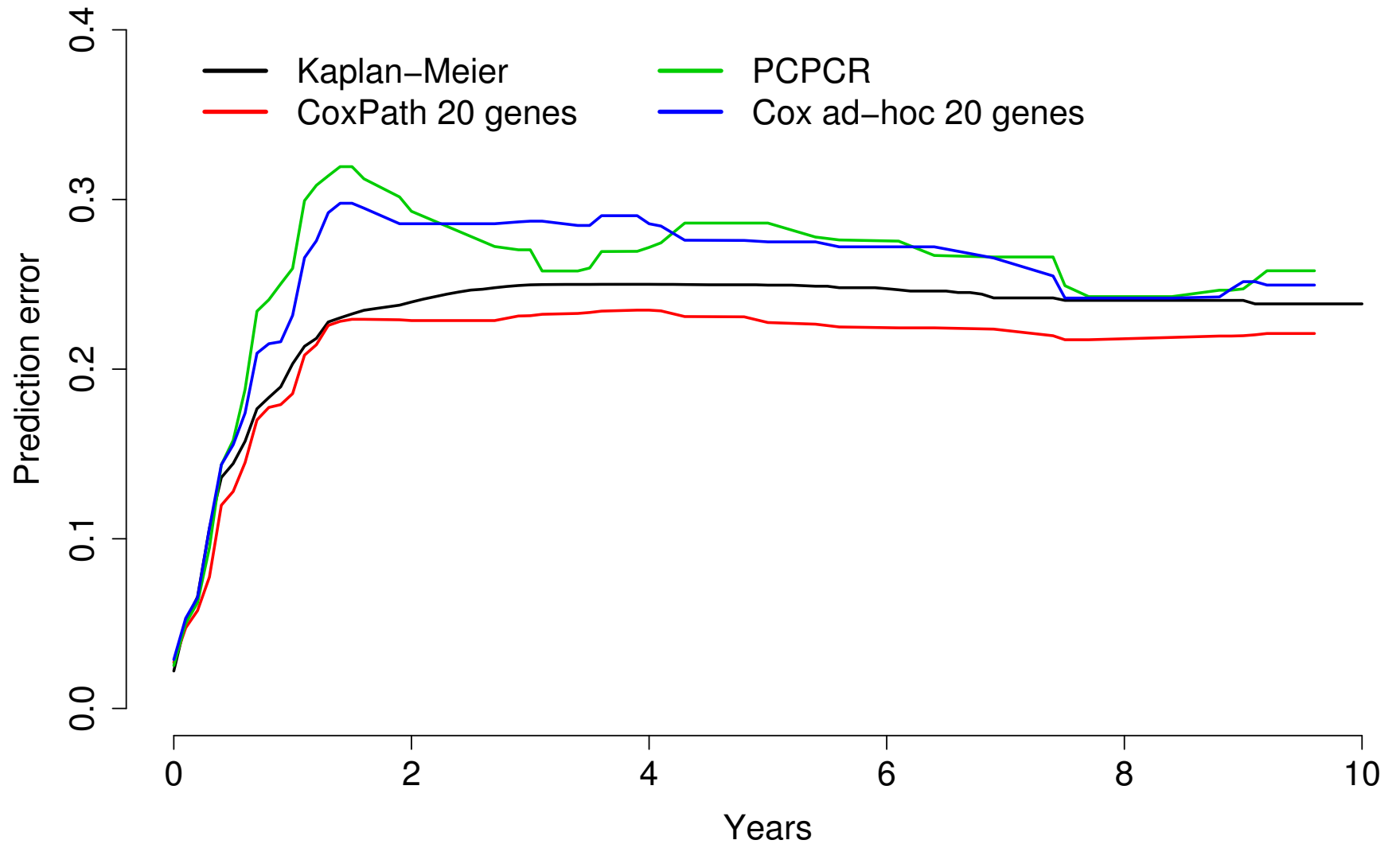
Rosenwald DLBCL study (9)

Bootstrap .632+



Rosenwald DLBCL study (10)

Learnset (n=160); testset (n=80)



Rosenwald DLBCL study (11)

Summary of findings

- Prediction models developed show prognostic potential in terms of apparent error; PCPCR seems to be best
- Overfitting is substantial leading to far too optimistic estimates of prediction error
- The $.632+$ estimates of true prediction error indicate that, with the exception of CoxPath, the developed prediction models are no better than the Kaplan-Meier benchmark value ignoring all covariate information
- The $.632+$ estimates are in nice agreement to those obtained with usual data-splitting (cf. Segal, Biostatistics 2006)

Discussion and conclusion (1)

What is new?

- Generalization of $.632+$ estimator for application to time-to-event data
- Predictions are provided in terms of event-free probabilities
- Incorporation of time dimension leads to prediction error curve
- Inverse probability of censoring (IPC) weighting scheme takes censoring properly into account
- Methodology can also be used for other loss functions

Discussion and conclusion (2)

What is gained?

- The $.632+$ estimator enables to track the true prediction error curve without splitting data into training and test set
- For standard regression modelling the $.632+$ estimator essentially reduces to the $.632$ estimator, for flexible modelling approaches to bootstrap cross-validation
- Prediction error curves show the relative merits of a risk prediction model when compared to the Kaplan-Meier benchmark value
- Graphical display of relative overfitting rate aids identification of overfitting

Discussion and conclusion (3)

What is important?

- For flexible modelling strategies overfitting can be markedly reduced by regularization (e.g. weight decay for ANNs); "fine-tuning" can be performed with the $.632+$ estimator
- The $.632+$ estimator will only "work" when all steps of the data-dependent modelling process are repeatedly performed within each bootstrap sample (especially important for $p \gg n$)
- Application in high-dimensional settings needs further investigation
- So no free lunch (as always!) but available information can be used more efficiently

References (1)

- Brier GW: Verification of forecasts expressed in terms of probability. *Monthly Weather Rev*, 1950; 78; 1-3.
- Efron B: Estimating the error rate of a prediction rule: Improvement on cross-validation. *J Am Stat Assoc*, 1983; 78: 316-331.
- Efron B and Tibshirani R: Improvements on cross-validation: The .632+ bootstrap method. *J Am Stat Assoc*, 1997; 92: 548-560.
- Fu WJ, Carroll RJ, Wang S: Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics*, 2005; 21: 1979-1986.
- Gerds TA and Schumacher M: Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical J*, 2006; 48; to appear.
- Graf E, Schmoor C, Sauerbrei W, Schumacher M: Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 1999; 18: 2529-2545.
- Korn EL and Simon R: Explained residual variation, explained risk, and goodness of fit. *Am Stat*, 1991; 45: 201-206.

References (2)

- Li H, Gui J: Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 2004; 20 Suppl 1: i208-i215.
- Martin R and Yu K: Assessing performance of prediction rules in machine learning. *Pharmacogenomics*, 2006; 7: 534-550.
- Molinaro AM, Simon R, Pfeiffer RM: Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 2005; 21: 3301-3307.
- Park MY, Hastie T: L_1 regularization path algorithm for generalized linear models. Dept. of Statistics, Stanford University. Techn. Rep. February 28, 2006.
- Schemper M and Henderson R: Predictive accuracy and explained variation in Cox regression. *Biometrics*; 2000; 56: 249-255.
- Schumacher M, Graf E, Gerds T: How to assess prognostic models for survival data: a case study in oncology. *Method Inform Med*, 2003; 42: 564-571.
- Schumacher M, Holländer N, Schwarzer G, Sauerbrei W: Prognostic Factor Studies. In: John Crowley and Donna Pauler Ankerst (Hrsg): *Handbook of Statistics in Clinical Oncology*. , 2. Auflage. Boca Raton, FL: Chapman & Hall /CRC, 2006; 289-333.

References (3)

- Segal MR: Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. *Biostatistics*; 2006; 7:268-285.
- Simon R, Radmacher MD, Dobbin K, McShane LM: Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Nat Cancer Inst*, 2003; 95: 14-18.
- Van Houwelingen HC, Bruinsma T, Hart AAM, van't Veer LJ, Wessels LFA: Cross-validated Cox regression on microarray gene expression data. *Stat Med*, in press, Early View on Wiley InterScience.
- Van der Laan MJ and Robins JM: *Unified Methods for Censored Longitudinal Data and Causality*. Springer 2003.
- Wehberg S, Schumacher M: A comparison of nonparametric error rate estimation methods in classification problems. *Biometrical J*, 2004; 46: 35-47.

Application in high-dimensional settings (4)

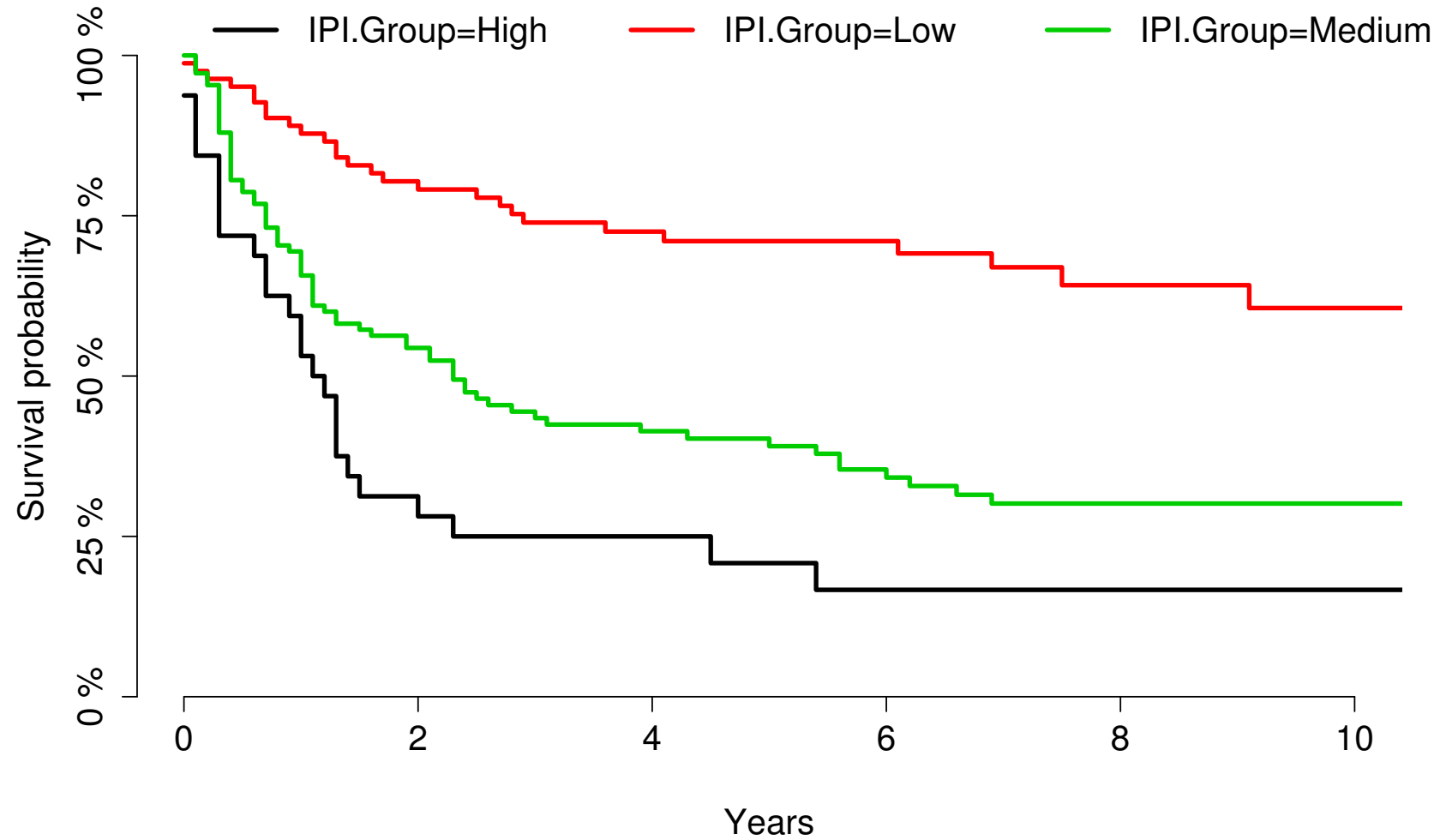
Idea: Combine advantages of bootstrap resampling with leave-one-out cross-validation! (Fu et al., Bioinformatics 2005)

- Draw B bootstrap samples X_b^* each of size n from X (with replacement)
- Perform leave-one-out cross-validation for prediction model r_b^* in order to obtain $\widehat{Err}_{CV}(t, r_b^*)$
- Calculate the "bagging estimator" of prediction error

$$\frac{1}{B} \sum_{b=1}^B \widehat{Err}_{CV}(t, r_b^*)$$

- Limited experiences (small sample sizes (20-50); $p = 1, 5, 10$; classification problem) so far.

Rosenwald DLBCL study (12)



IPI.Group=High	32	10	6	4	3	1
IPI.Group=Low	82	63	50	37	22	14
IPI.Group=Medium	108	56	39	28	19	13

Rosenwald DLBCL study (13)

Kaplan–Meier in IPI strata

