Assessment of prediction error of risk prediction models

Thomas Gerds and Martin Schumacher

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

Outline

- Situation
- Measures of prediction error
- Application to prediction of breast cancer survival
- General conclusion
- Considerations for breast cancer risk prediction

Situation (1)

- **Prediction:** $\pi(t|X)$

  predicted probability that an individual will be event-free up to t units of time based on covariate information X available at t = 0

- **Outcome:**

$$1\{T > t\} = \begin{cases} 0, & \text{event occurred before } t \\ 1, & \text{otherwise} \end{cases}$$

  T denotes time to event of interest

- **Goal:** Assessment of predictions $\pi(t|X_i)$ based on a comparison with actually observed outcomes T_i in a sample of n individuals (i = 1,…,n)

Situation (2)

- Prediction: $\pi(t|X)$
  - can be defined for a fixed time t or for a time range
  - should have the properties of a survival probability function
  - is ideally externally derived
  - but otherwise, can be anything: produced by statistical model building, by machine learning techniques or may constitute expert guesses

Measures of prediction error (1)

- General loss function approach

  $E(L(T, X, \pi))$

- Common choices:

$$L(T, X, \pi) = \left[1\{T > t\} - \pi(t|X)\right]^2$$

$$L(T, X, \pi) = -\left[1\{T > t\}\log\pi(t|X) + 1\{T \le t\}\log(1 - \pi(t|X))\right]$$

Measures of prediction error (2)

- Expected quadratic or Brier score

$$E\left(\left[1\{T > t\} - \pi(t|X) \right]^2 \right)$$

"Mean Squared Error of Prediction (MSEP)"

Measures of prediction error (2)

– Expected quadratic or Brier score

$$E\left(\left[1\{T > t\} - \pi(t|X) \right]^2 \right)$$

"Mean Squared Error of Prediction (MSEP)"

– Decomposition

$$\left[1\{T > t\} - \pi(t|X) \right]^2$$

$$= \left[1\{T > t\} - S(t|X) \right]^2 + \left[\pi\{T > t\} - S(t|X) \right]^2$$

S(t|X) denotes the "true" probability that an individual with covariate X will be event-free up to t

Measures of prediction error (3)

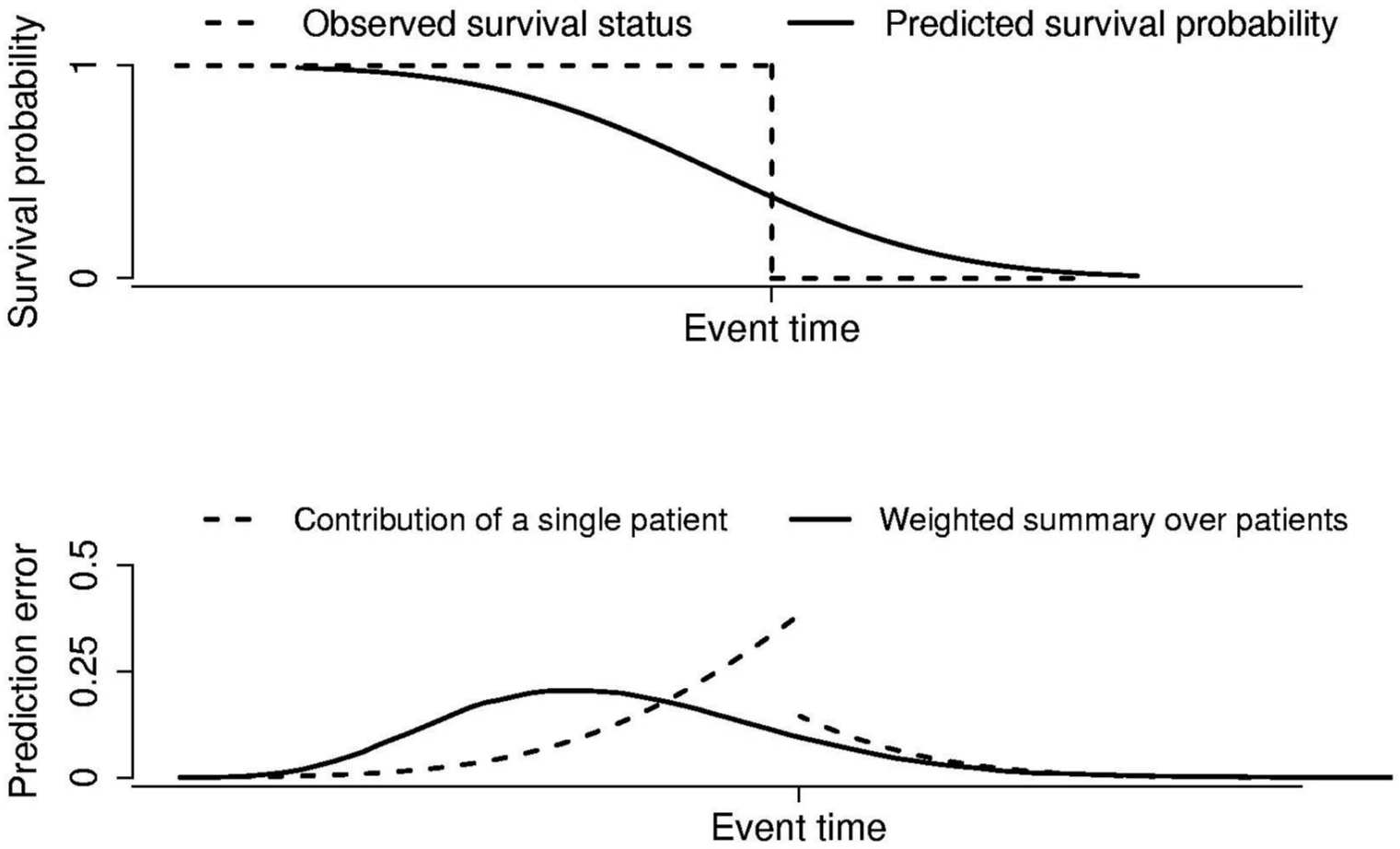
- Empirical quadratic or Brier Score

$$\frac{1}{n}\sum_{i=1}^{n}\left[\,1\{T_i > t\} - \pi(t|X_i)\,\right]^2$$

"Residual Sum of Squares (RSS)"

- MSEP and RSS are time-dependent in survival problems

- Graphical tool: plotting RSS over time


Observed survival status
Predicted survival probability
Survival probability
1
0
Event time
Contribution of a single patient
Weighted summary over patients
Prediction error
0
0.25
0.5
Event time

Measures of prediction error (3)

- Empirical quadratic or Brier Score

$$\frac{1}{n}\sum_{i=1}^{n}\left[\,1\{T_i > t\} - \pi(t|X_i)\,\right]^2$$

"Residual Sum of Squares (RSS)"

- Incorporation of censored observations

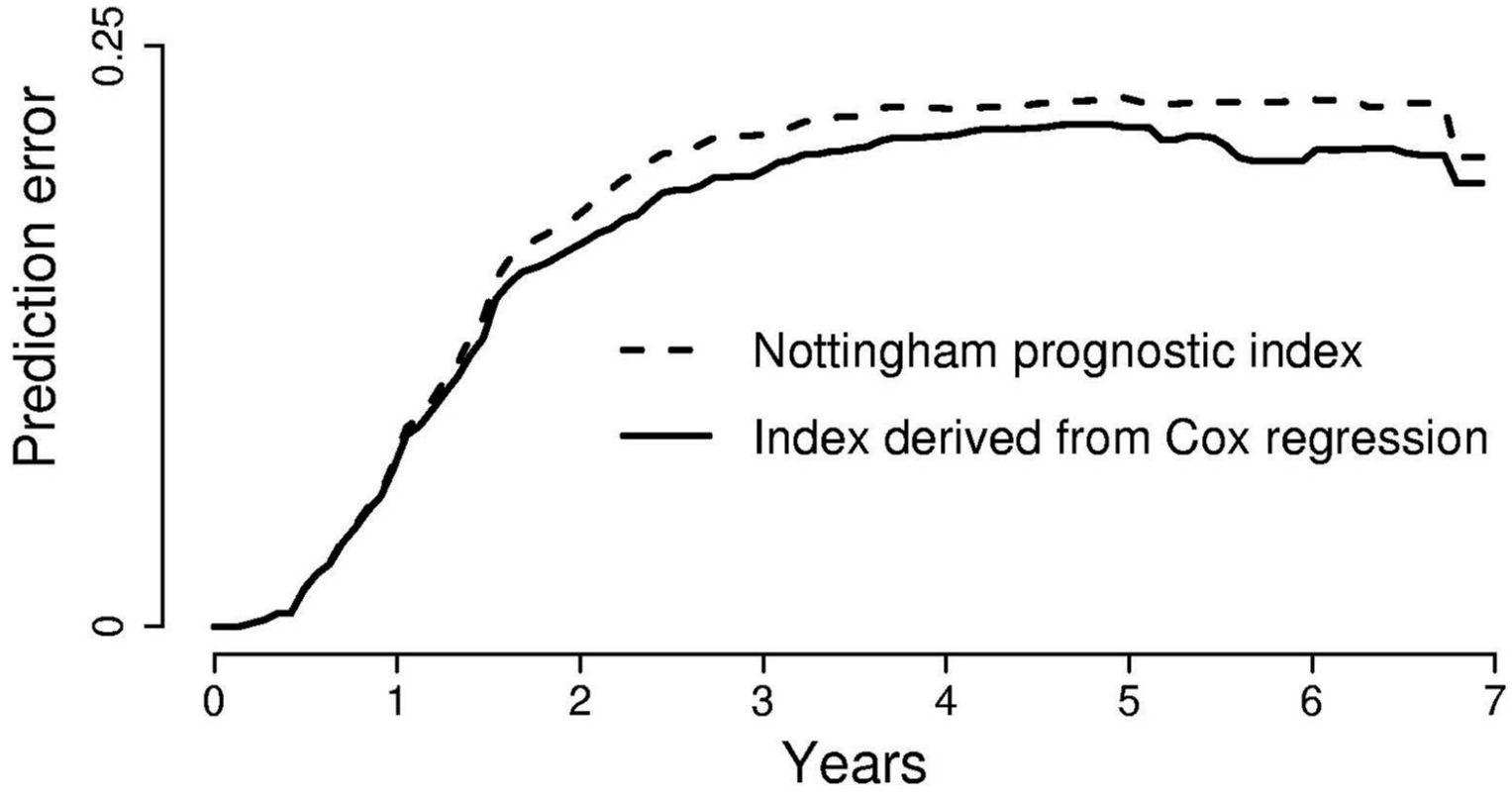
$$\frac{1}{n}\sum_{i=1}^{n} w_i(t)\left[\,1\{T_i > t\} - \pi(t|X_i)\,\right]^2$$

"Weighted Residual Sum of Squares (WRSS)"

Application to prediction of breast cancer survival

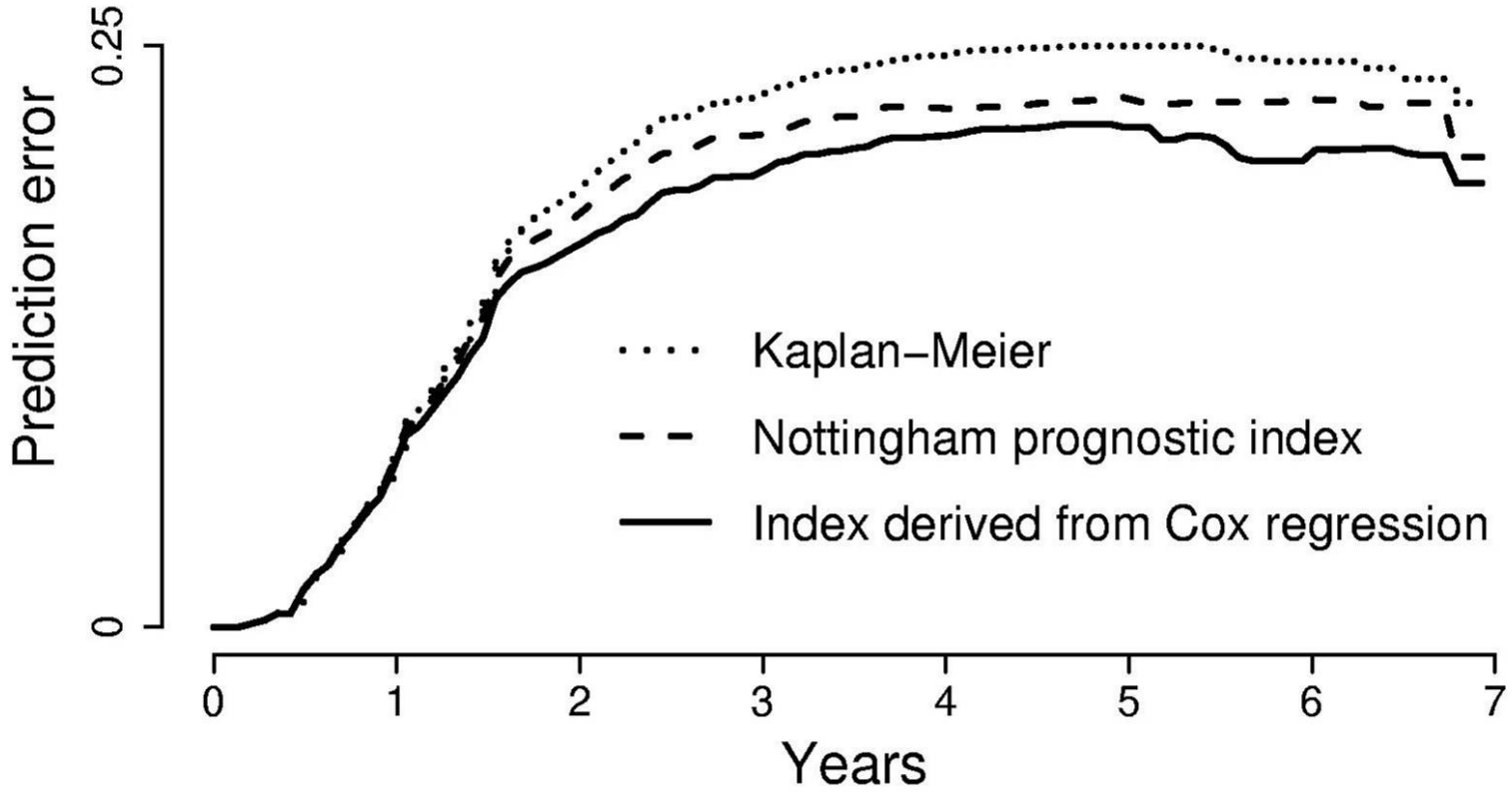
GBSG-2-study (German Breast Cancer Study Group)

- 686 patients with complete information on prognostic factors
- Two thirds are randomized, otherwise standardized treatment
- Median follow-up 5 years, 299 events for event-free survival
- Prognostic factors considered: age, tumor size, tumor grade, number of positive lymph nodes, progesterone receptor, estrogen receptor
- Predictions for individual patients are derived in terms of conditional event-free probabilities given the covariate combination by means of the Nottingham Prognostic Index and a Cox regression model with all six prognostic factors


Prediction error
0.25
0
Nottingham prognostic index
Index derived from Cox regression
0
1
2
3
4
5
6
7
Years

Which benchmark value?

- "Naive" prediction $\pi(t|X) = 0.5$ for all t and X gives a Brier score value of 0.25
- Common prediction $\pi(t)$ for all individuals ignoring the available covariate information ("pooled Kaplan Meier")

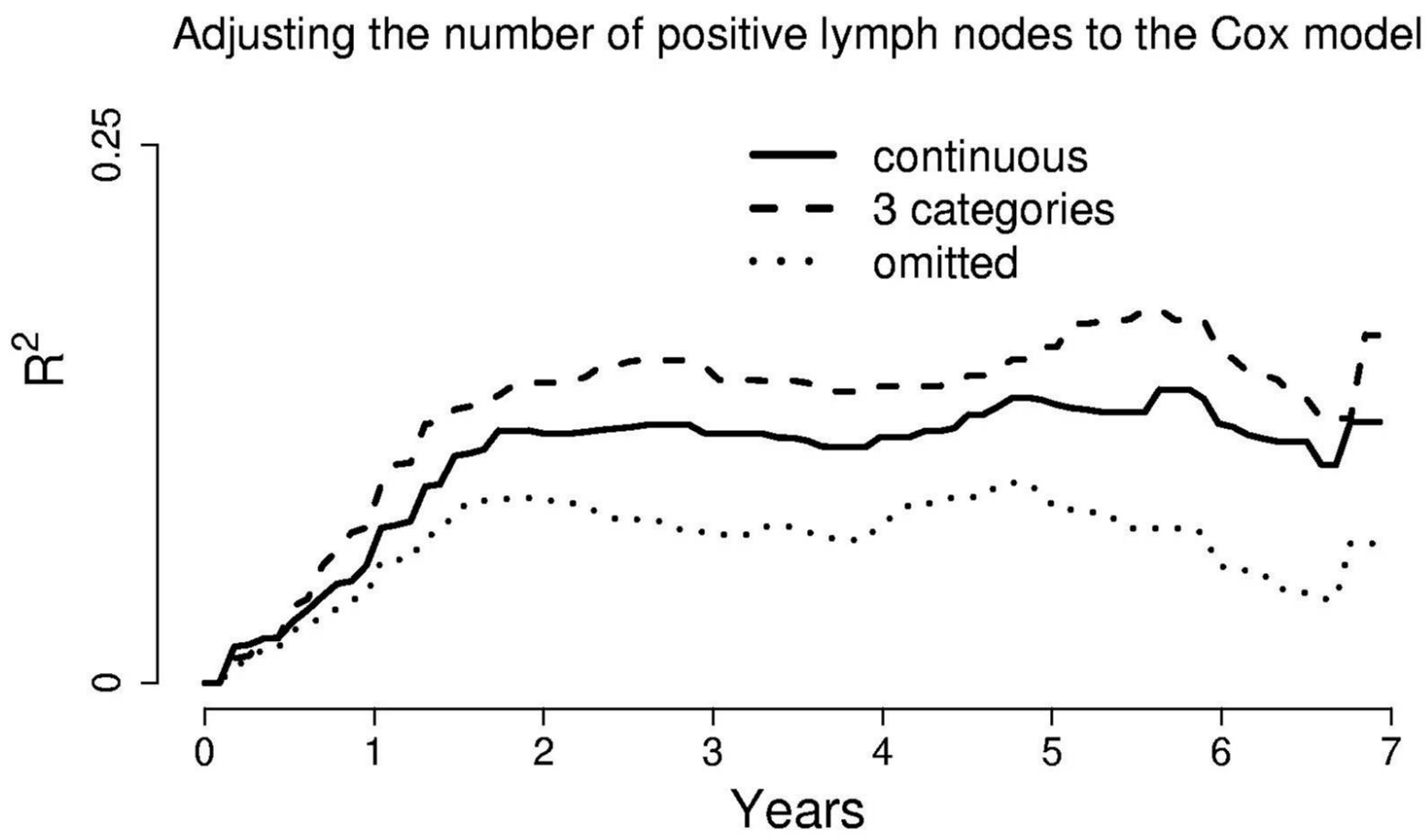

Prediction error
0.25
0
Kaplan–Meier
Nottingham prognostic index
Index derived from Cox regression
0
1
2
3
4
5
6
7
Years

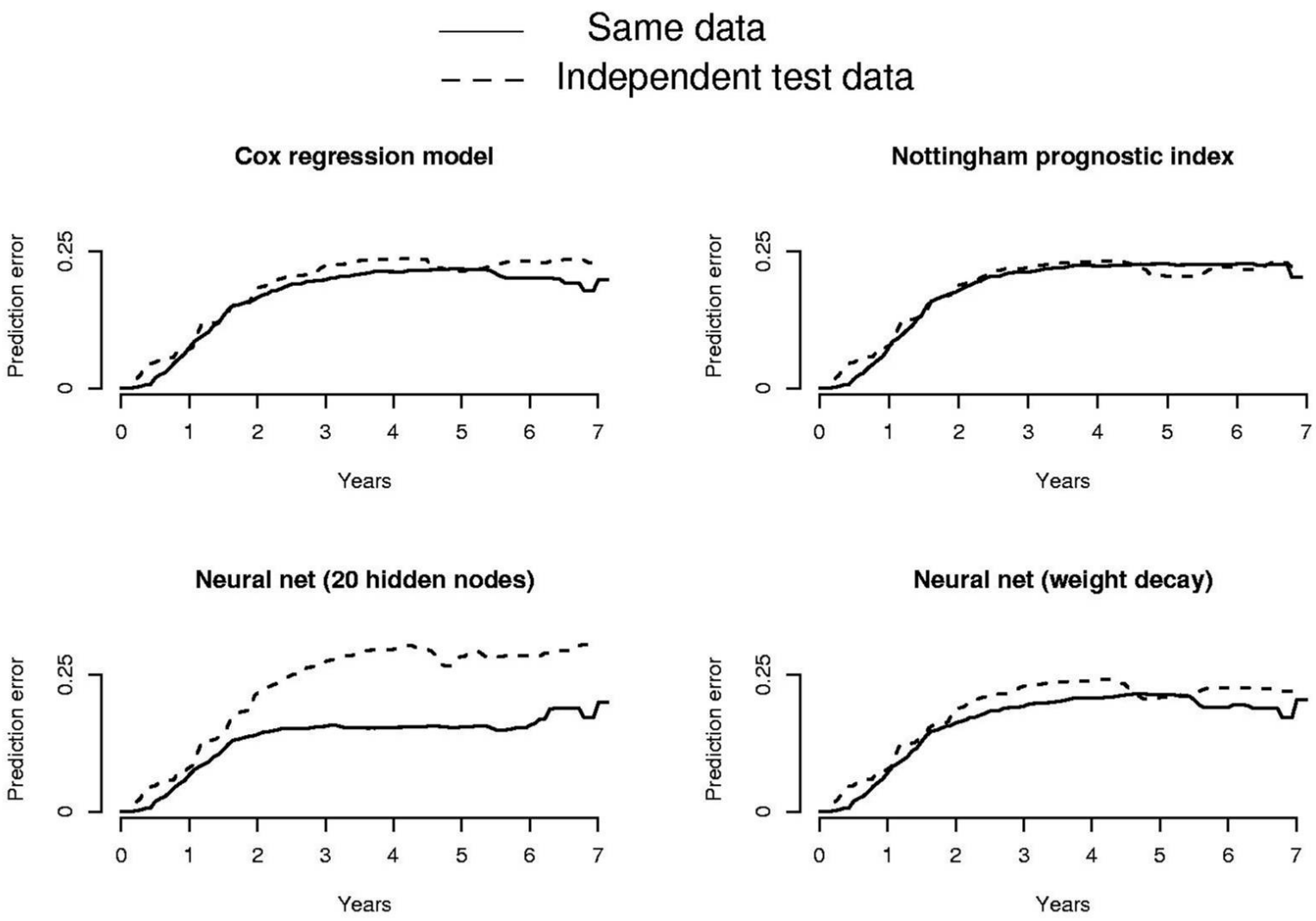
Which benchmark value?

- "Naive" prediction $\pi(t|X) = 0.5$ for all t and X gives a Brier score of 0.25
- Common prediction $\pi(t)$ for all individuals ignoring the available covariate information ("pooled Kaplan Meier")'

- Calculation of R^2-measures for checking various aspects of prediction models

$$R^2(t) = 1 - \frac{WRSS(t,\pi(t|X))}{WRSS(t,\pi(t))}$$


Adjusting the number of positive lymph nodes to the Cox model
continuous
3 categories
omitted
R^2
0.25
0
0
1
2
3
4
5
6
7
Years


Same data
Independent test data
Cox regression model
Nottingham prognostic index
Neural net (20 hidden nodes)
Neural net (weight decay)
Prediction error
0
0.25
0 1 2 3 4 5 6 7
Years

General conclusion

The quadratic or Brier score

- is the mean squared error of prediction (MSEP) when predictions are made in terms of event(-free) probabilities
- allows the assessment of any kind of predictions based on individual covariate values
- can be estimated even in the presence of right censoring by a weighted residual sum of squares in a nonparametric way
- is a valuable tool to detect overfitting
- allows the calculation of R^2-measures
- can be adapted to the situation of competing risks and dynamic updating of predictions

Considerations for breast cancer risk prediction

- **Prediction: $\pi(t|X)$**

  predicted probability that a woman will develop breast cancer up to time t based on covariate information including age available at t = 0 (entry into program or study; time when prediction is performed)

- **Outcome:**

  $$1\{T \le t\} = \begin{cases} 0, & \text{development of breast cancer before t} \\ 1, & \text{otherwise} \end{cases}$$

  T denotes time from entry into program to development of breast cancer

- **Intention:** Assessment of predictions for t = 5y based on aggregated data published by Costantino et al. JNCI 1999; constant prediction ignoring all covariate information is used as benchmark value

Table 3. Comparison of the expected cases of total breast cancer (invasive and all *in situ*) predicted from model 1 to the observed cases among white women in the placebo arm of the Breast Cancer Prevention Trial

Age group, y	Predicted 5-year risk, %	No. of women	Observed (*O*) breast cancers	Expected (*E*) breast cancers	*E/O*	95% confidence intervals
≤49	<2.32	111	1	1.93	1.93	0.35–76.25
	2.32–2.65	499	11	9.60	0.87	0.49–1.75
	2.66–3.28	521	25	12.89	0.52	0.35–0.80
	3.29–4.73	614	17	19.32	1.14	0.71–1.95
	>4.73	587	29	31.42	1.08	0.75–1.62
	Total	2332	83	75.16	0.91	0.73–1.14
50–59	<2.32	304	8	5.35	0.67	0.34–1.55
	2.32–2.65	468	14	9.80	0.70	0.42–1.28
	2.66–3.28	362	6	8.47	1.41	0.65–3.85
	3.29–4.73	326	13	10.43	0.80	0.47–1.15
	>4.73	347	13	17.69	1.36	0.80–2.56
	Total	1807	54	51.75	0.96	0.73–1.28
≥60	<2.32	784	21	9.75	0.46	0.30–0.75
	2.32–2.65	232	8	4.77	0.60	0.30–1.38
	2.66–3.28	308	12	7.61	0.63	0.36–1.23
	3.29–4.73	244	9	8.30	0.92	0.49–2.02
	>4.73	262	17	14.01	0.82	0.51–1.41
	Total	1830	67	44.44	0.66	0.52–0.86
All ages	<2.32	1199	30	17.03	0.57	0.40–0.84
	2.32–2.65	1199	33	24.17	0.73	0.52–1.06
	2.66–3.28	1191	43	28.97	0.67	0.50–0.93
	3.29–4.73	1184	39	38.05	0.98	0.71–1.37
	>4.73	1196	59	63.13	1.07	0.83–1.41
Grand total		5969	204	171.34	0.84	0.73–0.97

Costantino et al., Journal of the National Cancer Institute, Vol. 91, No. 18, September 15, 1999

Estimated prediction error based on aggregated data

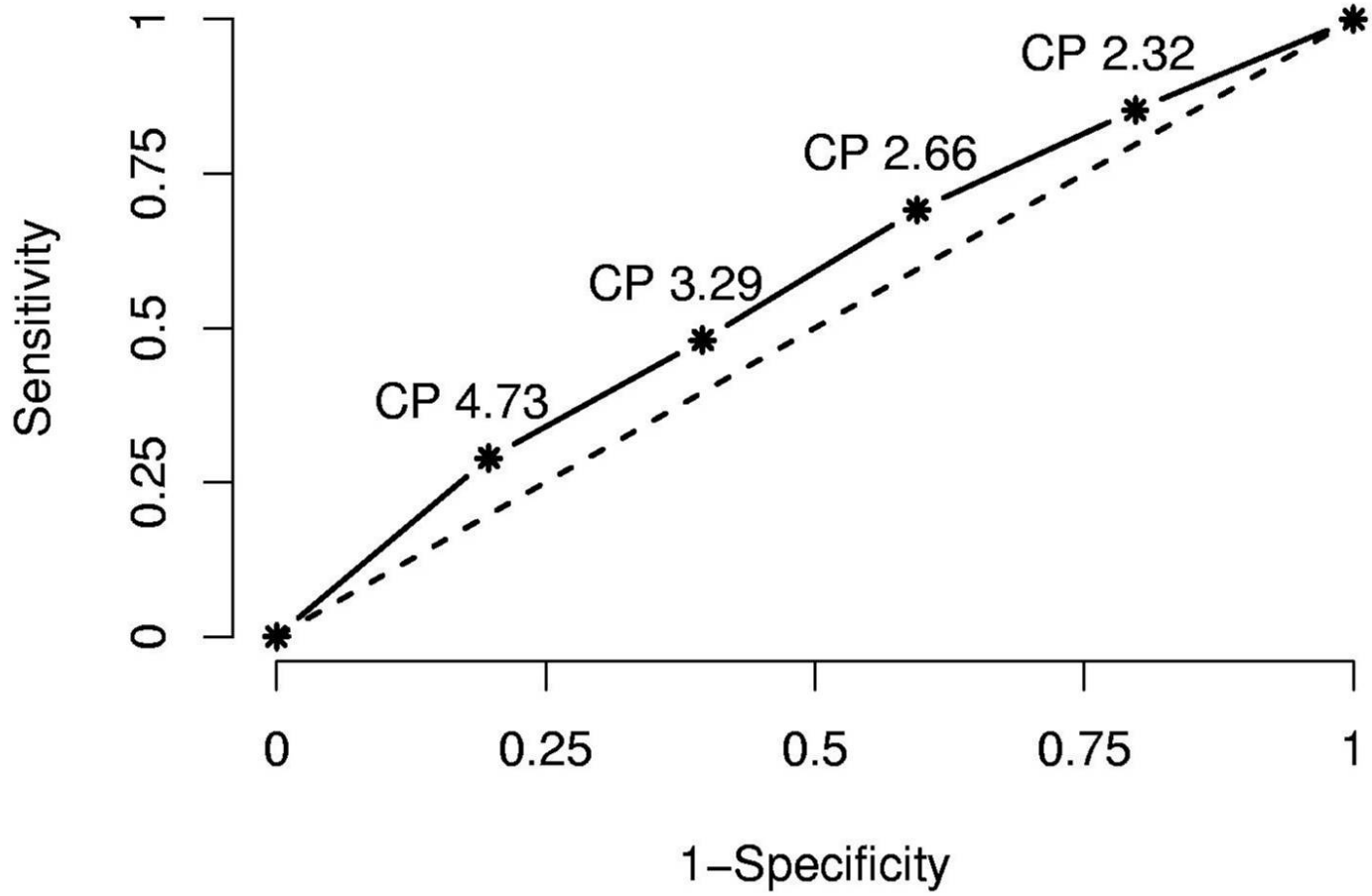
	Brier Score		Logarithmic score	
Age group, y	model 1	const. pred.	model 1	const. pred.
≤ 49	0.03428	0.03432	0.15289	0.15367
50 – 59	0.02900	0.02899	0.13446	0.13434
≥ 60	0.03519	0.03527	0.15625	0.15702
All ages	0.03296	0.03301	0.14834	0.14897

Estimated relative risk (RR) for predicted risk quintiles (model 1, all ages)

Predicted 5-year, % risk	No. of women	Observed breast cancer	RR
< 2.32	1199	30	1
2.32 – 2.65	1199	33	1.1
2.66 – 3.28	1191	43	1.436
3.29 – 4.73	1184	39	1.316
> 4.73	1196	59	1.972

"Diagnostic" properties of predicted risk quintiles (model 1, all ages)

Cutpoint Pred. 5-year risk, %	Sensitivity	Specificity	Pos. pred. value	Neg. pred. value
2.32	0.853	0.203	0.036	0.975
2.66	0.690	0.405	0.039	0.974
3.29	0.480	0.604	0.041	0.971
4.73	0.289	0.803	0.049	0.970


CP 2.32
CP 2.66
CP 3.29
CP 4.73
Sensitivity
0
0.25
0.5
0.75
1
0
0.25
0.5
0.75
1
1−Specificity