Using Brier's scoring rule for risk prediction models in medical statistics

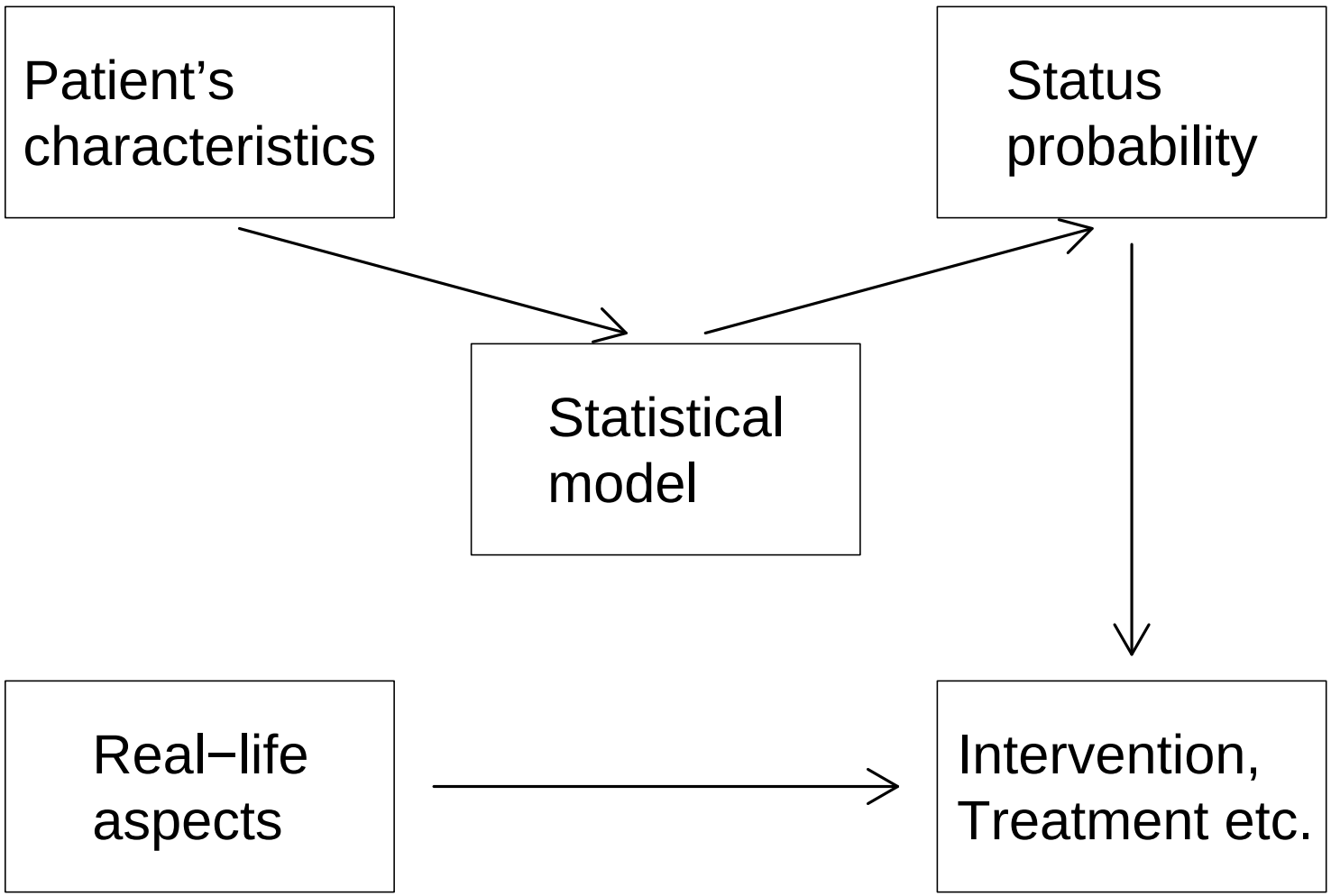
Thomas Gerds, University of Copenhagen
Martin Schumacher, University of Freiburg

Cleveland. September 29th, 2008

Outline

- Motivation
- The epo study
- Some applications
- Pro's and con's
- Survival analysis
- Summary

Using a model to make a decision



Risk prediction in medicine

When a **new** patient seeks advice, then the statistical model should extract and communicate the information inherent in data collected on similar patients.

On a high level a model performs well if it provides information that leads to successful medical decisions.

Since in real life other aspects enter into the decision making, it is often sufficient to assess and validate if the model can predict accurately the **probability for the status of a patient**.

Example: epo study[1]

Anaemia is a deficiency of red blood cells and/or hemoglobin and an additional risk factor for cancer patients.

This randomized placebo controlled trial includes 149 head and neck cancer patients. Treatment with 300 U/kg epoetin beta (epo) should enhance the hemoglobin level and thereby improve survival chances.

Henke et al. 2006 identified the c20 expression (erythropoietin receptor status) as a new biomarker for the prognosis of locoregional progression-free survival.

[1] Henke et al. Do erythropoietin receptors on cancer cells explain unexpected clinical findings? J Clin Oncol, 24(29):4708-4713, 2006.

Predictors

1. Age
   min: 41 y, median: 59 y, max: 80 y
2. Gender
   male: 85%, female: 15%
3. Baseline hemoglobin level
   mean: 12.03 g/dl, std: 1.45
4. Treatment arm
   epo: 50%, placebo 50%
5. Resection
   complete: 48%, incomplete: 19%, no resection: 34%
6. Erythropoietin receptor status
   pos: 32%, neg: 68%

Logistic regression model

Epo treatment was **successful** (n=68) when the hemoglobin level increased sufficiently during 7 weeks of radiotherapy and **not successful** (n=87) otherwise.

Variable	Coef	$CI_{95\%}$	p-value
(Intercept)	-17.301	(-25.981 ; -10.101)	< 0.0001
age	-0.032	(-0.094 ; 0.025)	0.281
Gender:female	1.552	(-0.093 ; 3.259)	0.066
HbBase	1.181	(0.689 ; 1.777)	< 0.0001
Treatment:epo	4.505[2]	(3.174 ; 6.202)	< 0.0001
Resection:Incompl	0.557	(-1.023 ; 2.201)	0.492
Resection:Compl	1.419	(0.121 ; 2.854)	0.039
epoReceptor:pos	1.759	(0.541 ; 3.152)	0.008

[2] That means everyone should be treated?

Assessing the predictive power of the logistic model

Patient no.	Treatment successful (%)	Predicted probability (%)	Residual	Brier's scoring rule
	Y_i	P_i	$Y_i - P_i$	$(Y_i - P_i)^2$
1	0	2.31	-2.31	0.05
2	0	1.91	-1.91	0.04
3	100	98.11	1.89	0.04
4	100	79.58	20.42	4.17
.	.	.	.	.
.	.	.	.	.
147	0	84.09	-84.09	70.71
148	100	96.64	3.36	0.11
149	0	11.93	-11.93	1.42
			apparent Brier score:	8.69

Definition

The **Brier score** for a model that predicts P_i for patient i out of N patients is

$$BS_N = \frac{1}{N}\sum_{i=1}^{N}(Y_i - P_i)^2$$

For a **given** model it **estimates** the expected squared difference between patient status and predicted probability.

Interpretation

The lower the Brier score of a model the better the predictive performance.

Benchmarks

- Coin toss: Brier score = 33 %
- Perfect prediction: Brier score= 0
- Performance of a model that ignores all covariates (null model)

Comparison to a model that ignores all covariates

Patient no.	Treatment successful (%)	Predicted probability (%)	Residual	Brier's scoring rule
	Y_i	P_i	$Y_i - P_i$	$(Y_i - P_i)^2$
1	0	44.3	-44.3	19.62
2	0	44.3	-44.3	19.62
3	100	44.3	55.7	31.02
4	100	44.3	55.7	31.02
.	.	.	.	.
.	.	.	.	.
147	0	44.3	-44.3	19.62
148	100	44.3	55.7	31.02
149	0	44.3	-44.3	19.62
			apparent Brier score:	24.67

The performance of a model

The **generalization error** of a risk prediction model is the accuracy that can be expected for a new patient.

A commonly used invalid estimate is called the *re-substitution estimate* (also known as the *apparent error* or the *training error*).

Valid estimates can be based on **external data** or obtained by repeated partition of the data into training and validation sets.

Bootstrap crossvalidation

Models with different complexity and different potential overfitting can be compared with the boostrap-crossvalidated Brier score

$$BS^* = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{M_b}\sum_{i \notin Boot[b]}(Y_i - P_i^*)^2$$

where P_i^* is the probability predicted for patient i when the model is trained using the bootstrap sample $Boot[b]$.

Different models

- LRM(0) ignores all predictive factors
- LRM(1) is the model discussed so far
- LRM(2) excludes the erythropoietin receptor status

- LRM(3) is obtained from automated backward elimination
- LRM(4) is an ad-hoc strategy which dichotomizes age to find the minimal p-value for the effect of age.

- CART is a classification tree
- RF is a random forest (many classification trees combined to a majority vote)

Comparing different models with the Brier score

	$LRM^{(0)}$	$LRM^{(1)}$	$LRM^{(2)}$	$LRM^{(3)}$	$LRM^{(4)}$	CART	RF
Apparent performance	24.67	8.69	9.58	8.63	8.28	10.04	3.00
Bootstrap crossvalidation	25.19	11.58	11.85	11.9	13.7	12.52	11.02
Difference	0.52	2.89	2.27	3.27	5.42	2.48	8.02

Accuracy for two sample patients

PatNr	Age	Gender	HbBase	Resection	Treat	epoRec
134	51	male	12.6	Compl	Epo	positive
151	50	male	14.3	Incompl	Placebo	negative

PatNr 134: treatment success
PatNr 151: no treatment success

	PatNr	$LRM^{(0)}$	$LRM^{(1)}$	$LRM^{(2)}$	RF
Prediction	134	44.3	97.39	93.13	97.6
	151	44.3	18.73	46.83	10.8
Brier Score	134	31.03	0.07	0.47	0.06
	151	19.62	3.51	21.93	1.17
Pair concordant		no	yes	yes	yes

Advantages of the Brier score

General. It can be used to assess predictions by *any* model for binary and continuous and right censored response variables.

Mathematical. It estimates a well-defined parameter in the population

Philosophical. It is a strictly proper scoring rule, thus the true model would win any comparison

Practical. It has an interpretation for a single patient

Reliable. Easily implemented with validation procedures

Disadvantages of the Brier score

Fine tuning. It is a summary and a model with overall good performance may predict poorly for a single patient.

Cases and controls. It does not distinguish the performance of the model for cases and controls (Moskowitz & Pepe, Stat Med, 2004). It can not directly be used in case control studies.

Rare diseases. It is difficult to see differences in populations with small prevalence

Intuition. It does not penalise very small forecasted probabilities when they should be giving larger probabilities to the same extent that we penalise such forecasts with our intuition. Intuition apparently uses fractional or logarithmic rather than differences in probability. (Stephen Jewson, 2008, arXiv:physics/0401046)

Censored survival times

The locoregional progression free survival time T_i (in the epo study) can be represented by the **time-dependent** status:

$$Y_i(t) = \begin{cases} 0 & \text{Patient alive} \\ 1 & \text{Patient dead} \end{cases}$$

Suppose a survival model predicts the survival probability $S_i(t)$ for patient i based on **baseline** characteristics.

The weighted time-dependent Brier score yields a prediction error curve for the model:

$$\text{pec}(t; \mathit{model}) := \frac{1}{N} \sum_i \widehat{W}_i(t)\{Y_i(t) - S_i(t)\}^2$$

$\widehat{W}_i(t)$ are weights that account for right censoring.

Assessing the importance of the baseline hemoglobin level

Cox: Cox regression model using

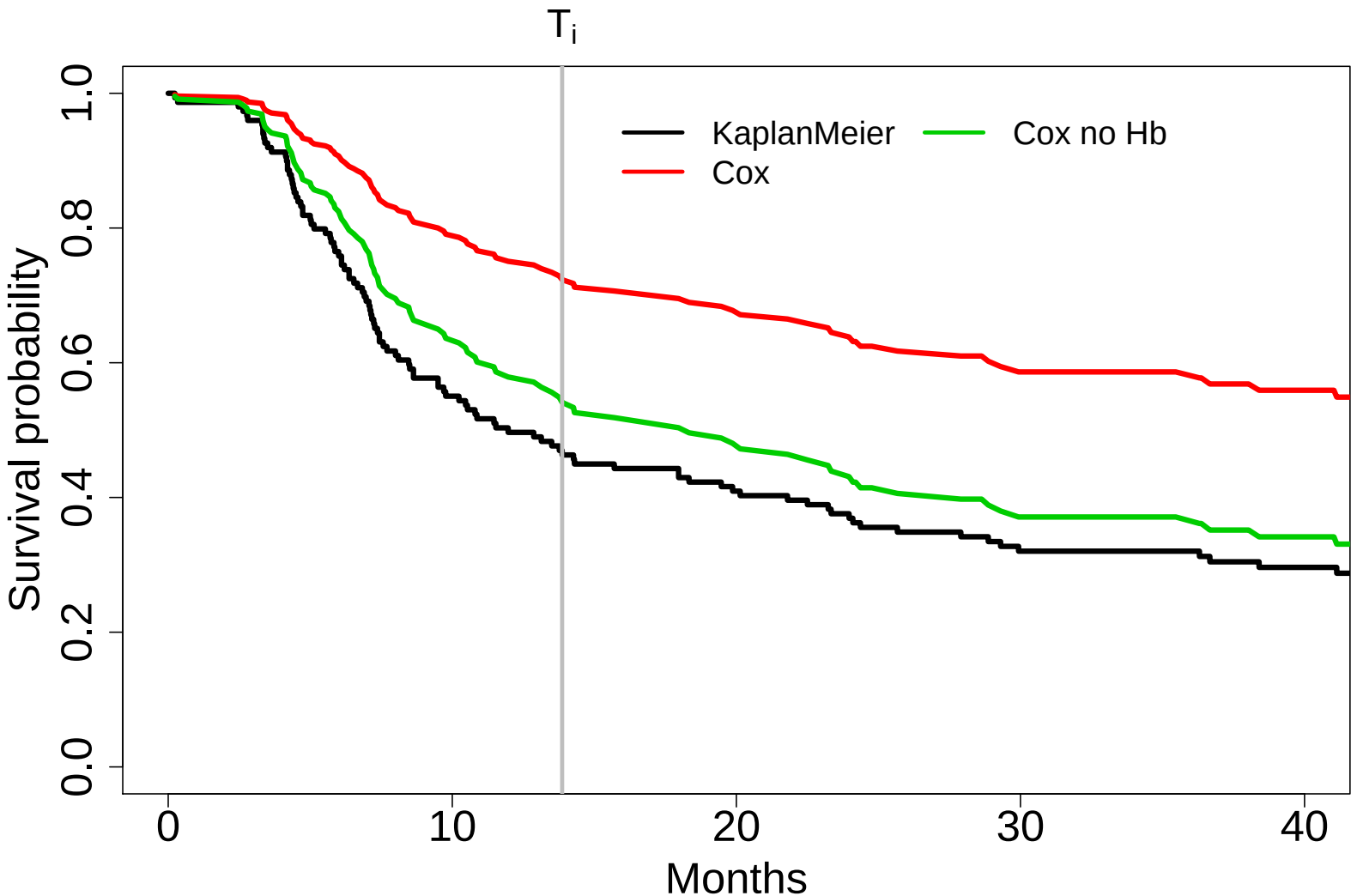
```
age + HbBase + Resection + Treat
```

Cox no hb: Cox regression model using

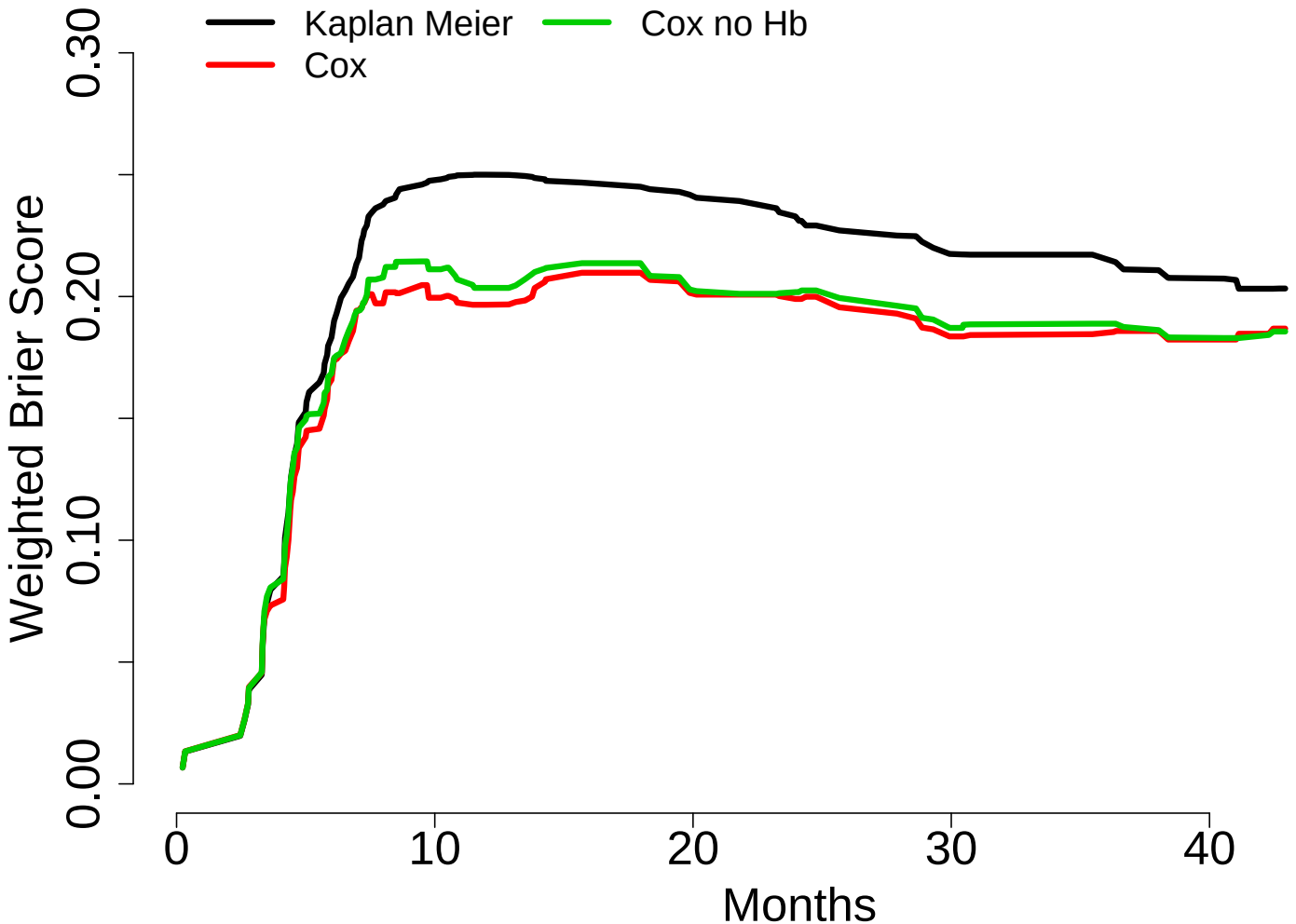
```
age + Resection + Treat
```

Benchmark model: Kaplan-Meier using no covariates

Sample Patient 151



Estimation method: .632+ bootstrap crossvalidation



Summary

For binary and continuous and right censored outcome the **Brier score** can be used ...

- to find predictive or diagnostic markers
- to assess the predictive performance of a traditional statistical model
- to assess an algorithmic (black box) model
- to detect overfitting
- to compare simple to complex models
- for focussed and automated model selection

Rpackages: `Design,pec`

Brier, G. W. (1950).
Verification of forecasts expressed in terms of probability.
Monthly Weather Review 78, 1–3.

Redelmeier, D., D. Bloch, and D. Hickam (1991).
Assessing predictive accuracy: how to compare Brier scores.
Journal of Clinical Epidemiology 44, 1141–6.

Gneiting, T. and A. E. Raftery (2007).
Strictly proper scoring rules, prediction, and estimation.
Journal of the American Statistical Association 102(477), 359–378.

Gerds, T. A., T. Cai, and M. Schumacher (2008).
The performance of risk prediction models.
Biometrical Journal 50(4), 457–479.

Gerds, T. A. and M. Schumacher (2006).
Consistent estimation of the expected Brier score in general survival models.
Biometrical Journal 48, 1029–1040.

Gerds, T. A. and M. Schumacher (2007).
On Efron type measures of prediction error for survival analysis.
63, 1283–1287.