Confidence scores for prediction models

Joint work with Mark van de Wiel

Thomas A Gerds

Department of Biostatistics, University of Copenhagen
tag@biostat.ku.dk

6 October 2010

- Motivation
- Predictions in medicine
- Prediction performance
- Confidence scores
- Summary

If you urgently need information ...

If you urgently need information . . .

For example, to answer the following multiple choice question:

Q: What is bagging?

1. A machine learning ensemble meta-algorithm
2. Searching in a *bag*
3. A special case of model averaging
4. The last name of Leo Breiman's first Ph.D student
5. A short name for bootstrap aggregating

then . . .

. . . there are several strategies

Bagging

The results of the k-*nearest neighbor method* can be improved by combining the results of many neighbors, think of *asking the audience* from the well-known tv-show.

More generally, a *weak learner* can be improved by *bagging*[*] .

Random forest[*] combines many decision trees (based on bootstrap) and thereby improves the predictions of a single tree.

[*] Leo Breiman (1996). "Bagging predictors". Machine Learning 24 (2): 123–140
[*] Leo Breiman (2001). "Random Forests". Machine Learning 45 (1), 5-32

Prediction problem

Response:

$$Y_i = \begin{cases} 1 & \text{positive / disease} \\ 0 & \text{negative / non-disease} \end{cases}$$

Predictors:

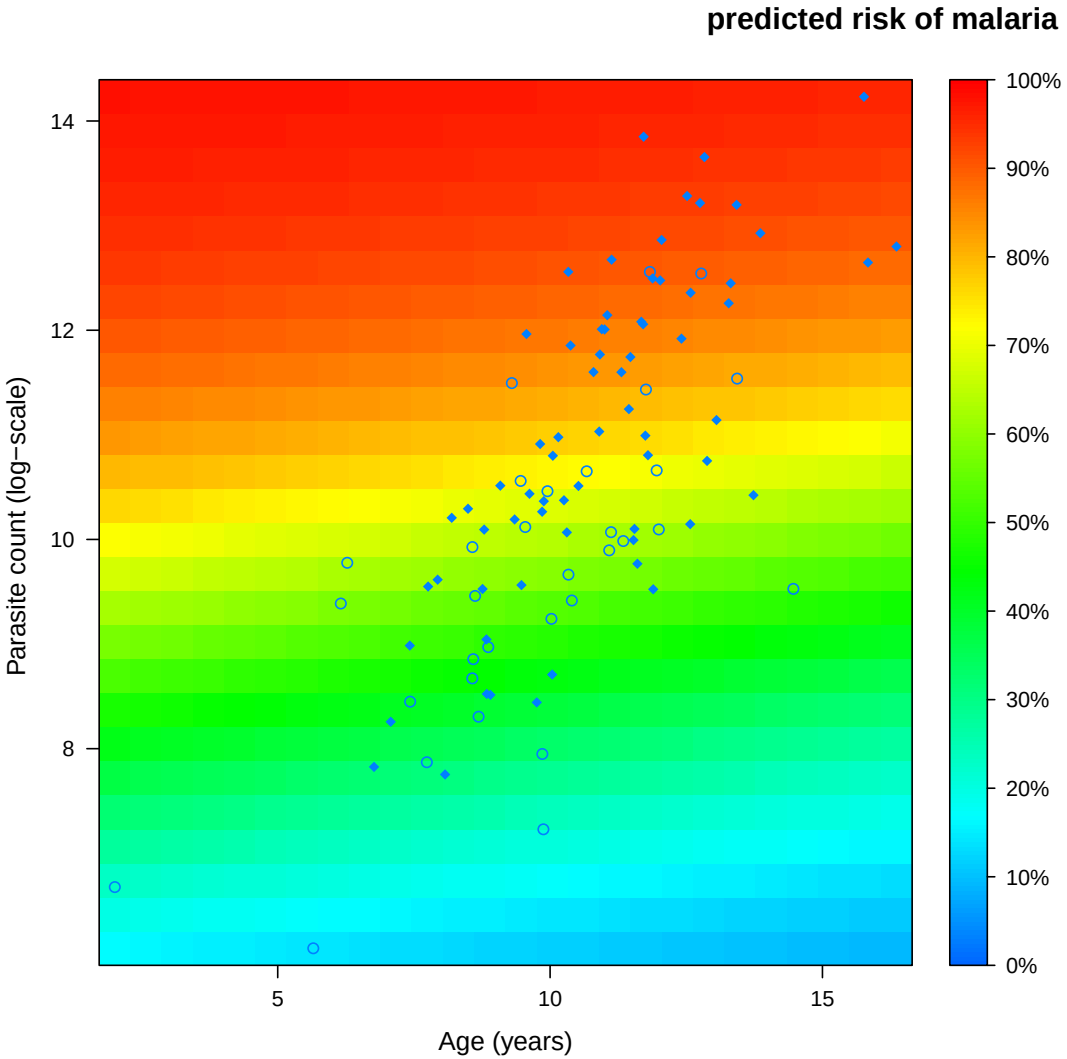
$$X_i = (X_i^1, X_i^2, \ldots, X_i^L)$$

Parameter:

$$\mathrm{P}(Y_i = 1 | X_i)$$

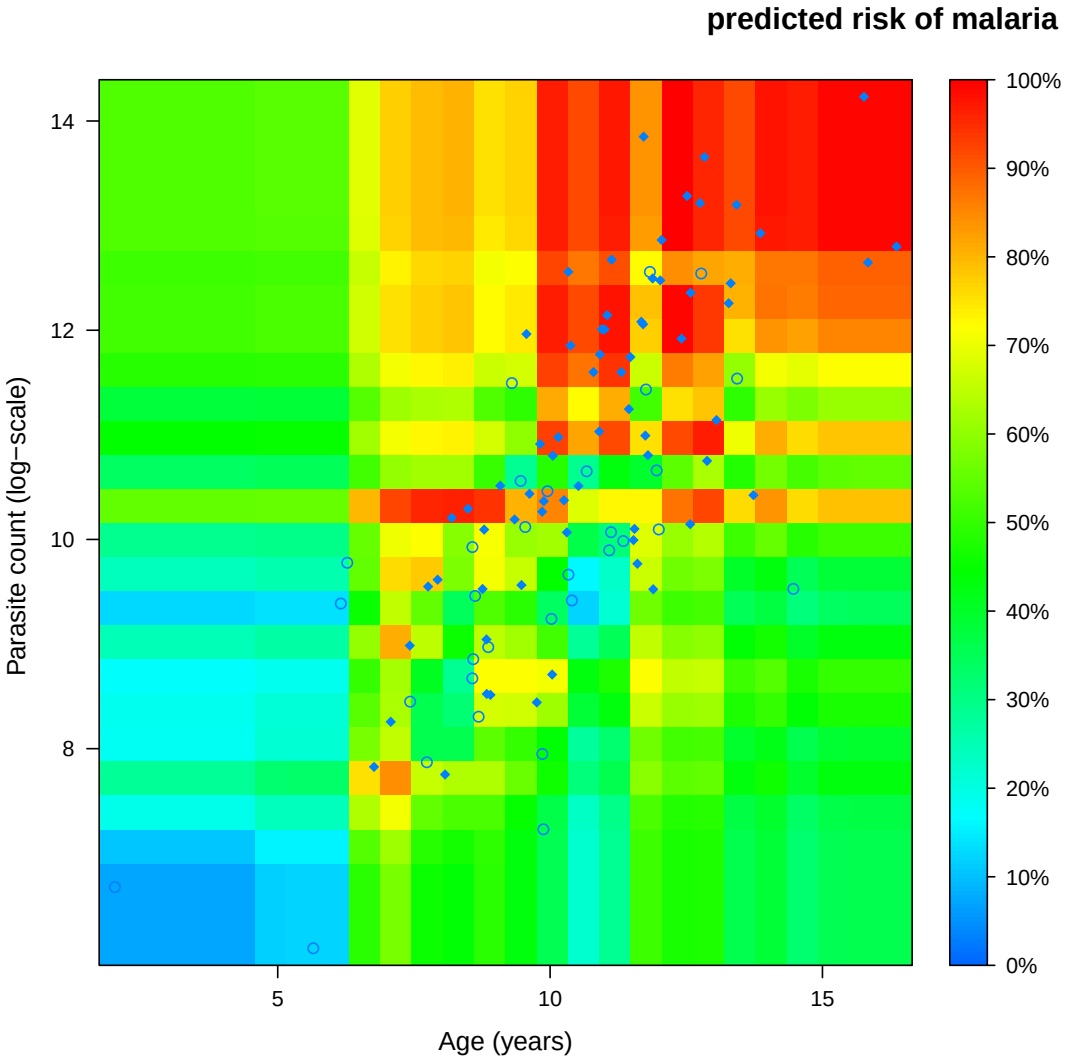
Data set: $D_n = (X_1, Y_1, \ldots, X_n, Y_n)$

Risk plot: logistic regression



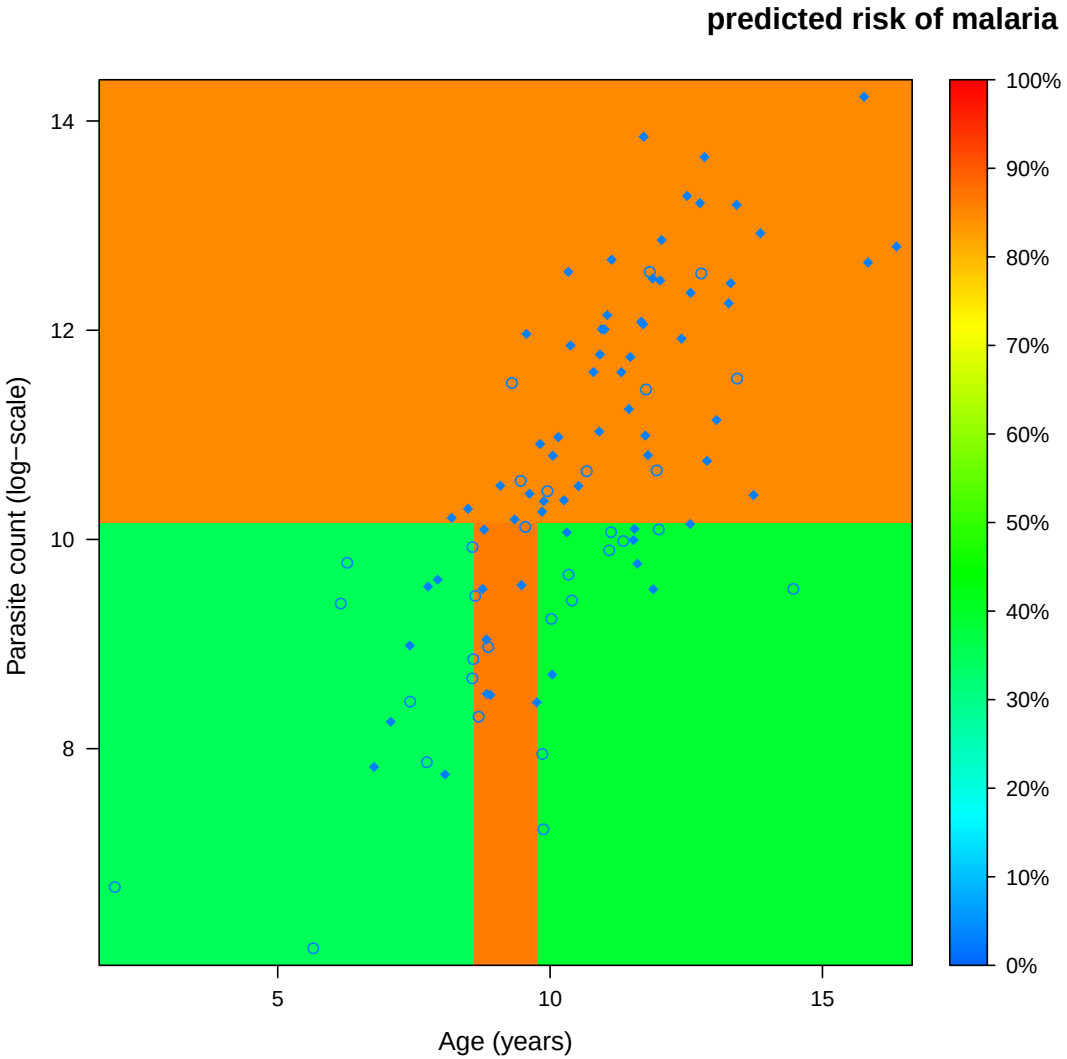

```
predict(glm(fever~age+parasite,data=d,family="binomial"))
```

Risk plot: random forest



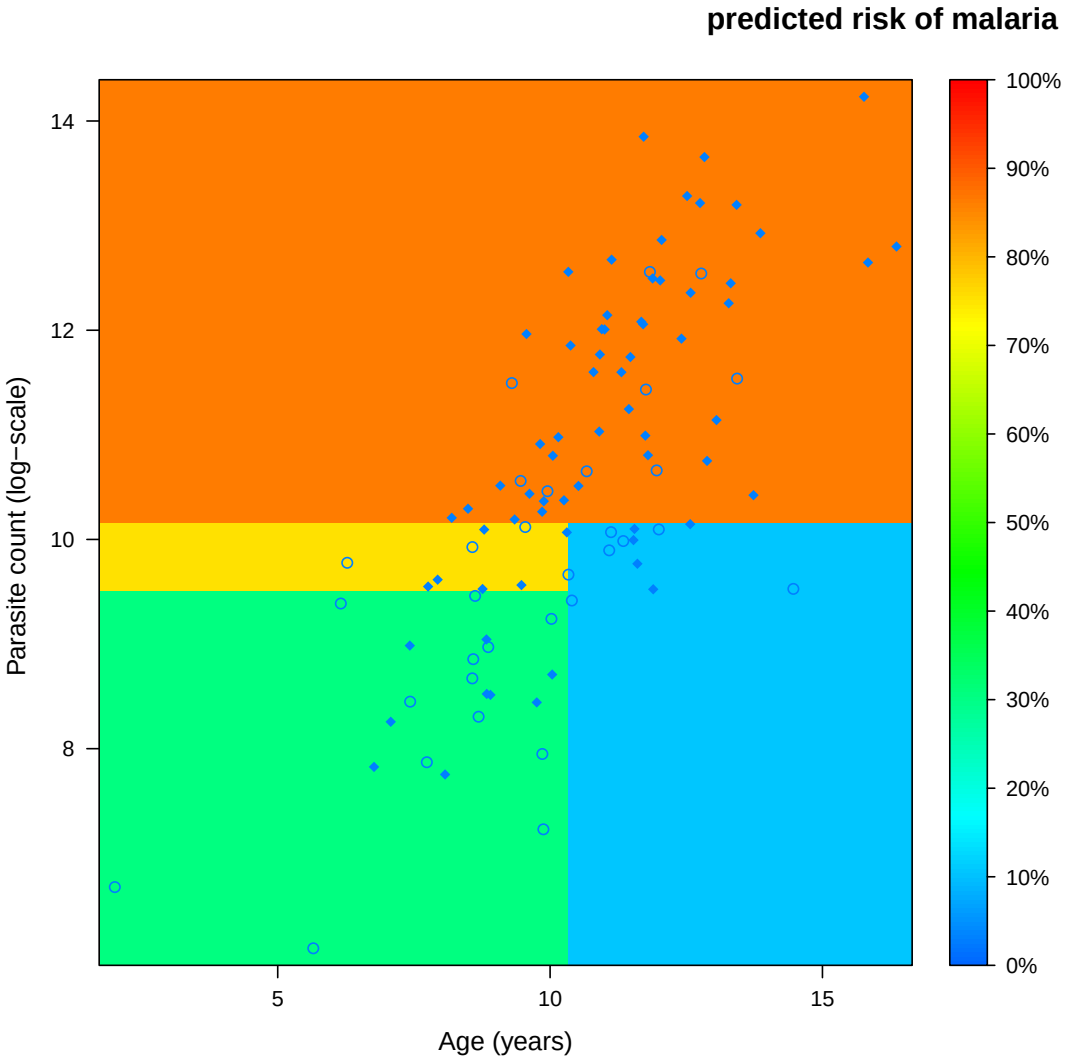

```
predict(randomForest(fever~age+parasite,data=d,ntree=1000))
```

Risk plot: single decision tree



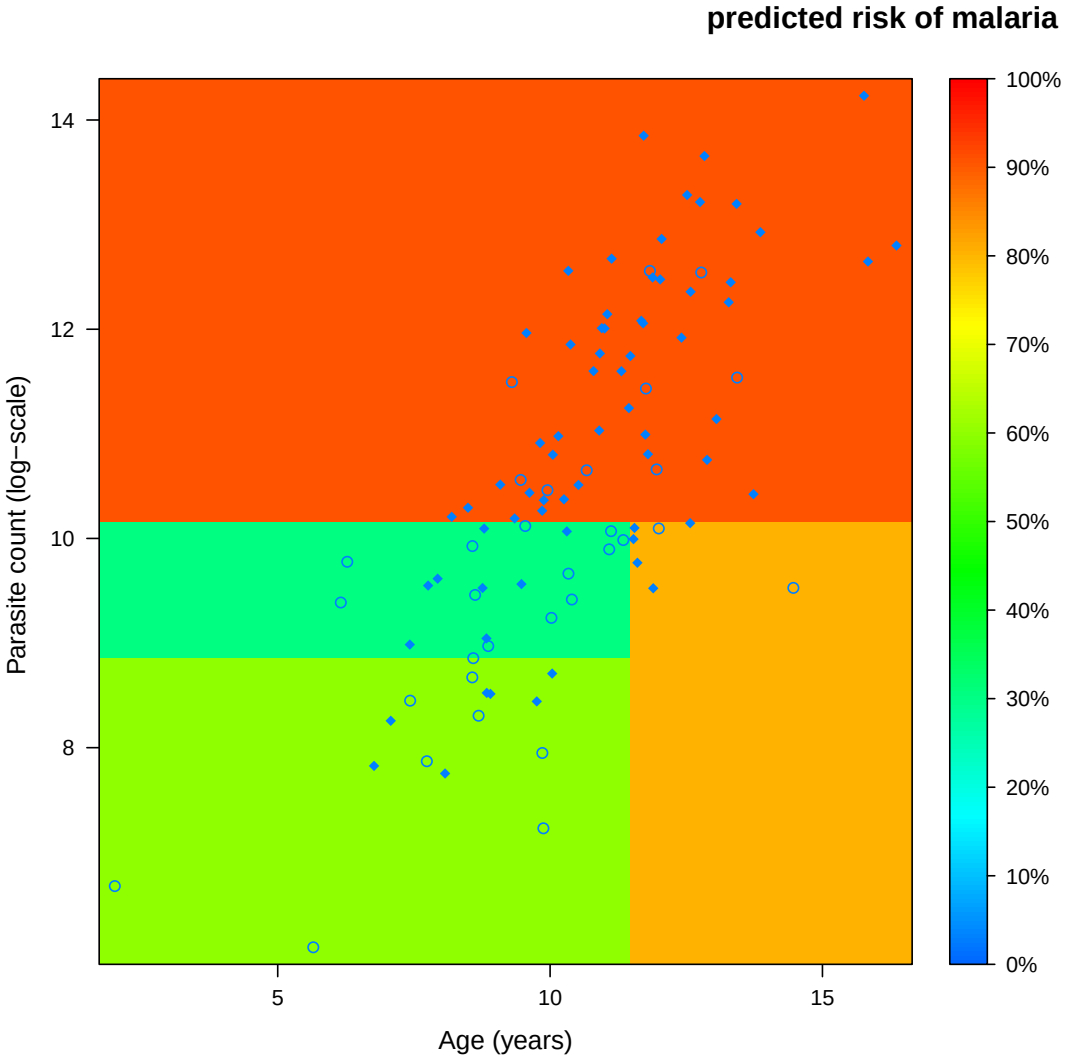

```
predict(rpart(fever~age+parasite,data=d))
```

Tree based on a bootstrap sample




```
predict(rpart(fever~age+parasite,data=d[sample(1:N,replace=T),]))
```

Tree based on a different bootstrap sample




```
predict(rpart(fever~age+parasite,data=d[sample(1:N,replace=T),]))
```

Nice methods, but what is the question?

Who is asking the question?

A patient needs to know:

- Am I diseased? (current status)
- Will I develop the disease? (future status)
- Should I stop smoking?
- Do I really need chemotherapy?

The community wants a risk prediction model

A basic researcher wants a biologically plausible model

A statistician wants a widely applicable strategy

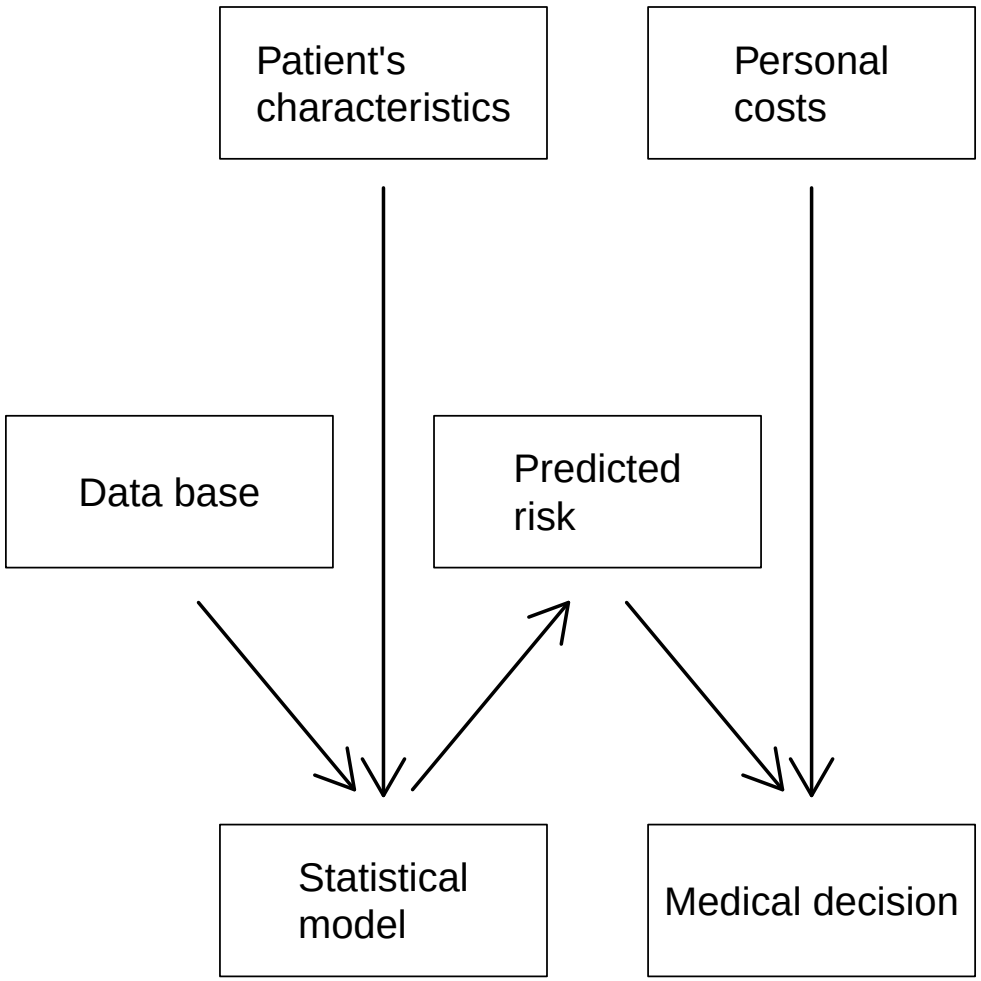
Common aims

To develop **statistical strategies** that select useful diagnostic and predictive models based on data.

To build a **statistical model** that predicts the risk of future subjects beneficial or adverse status (diseased, dead, pregnant, employed) based on a *bag* of data from former subjects.

To improve existing prediction models by including **new predictor variables** (genes, blood measurements)

Using a model to make a decision



Prediction model

A prediction model m is a mapping from subject individual predictor values to the risk of an event:

$$(X_i^1, X_i^2, \ldots, X_i^L) \rightarrow \boxed{\begin{array}{l}\text{Cox regression}\\ \text{Support Vector Machines}\\ \text{Bump hunting}\\ \text{Lars and his three cousins}\\ \text{Cart and RandomForests}\\ \text{Logistic regression}\end{array}} \rightarrow m(X_i)^*$$

$^* m(X_i) \approx \mathrm{P}(Y_i = 1|X_i)$

Prediction modelling strategy

A prediction modelling strategy S_n is a mapping from training data

$$D_n = \{(Y_1, X_1), \ldots, (Y_n, X_n)\}$$

to the set of prediction models:

$D_n \rightarrow$ Cox regression
Support Vector Machines
Bump hunting
Lars and his three cousins
Cart and RandomForests
Logistic regression $\rightarrow \mathcal{S}(D_n) = M_n$

In summary

A probabilistic **risk prediction** based on strategy S for the unknown status Y_i of a new patient X_i is the result of applying both mappings:

$$\mathcal{S} : D_n \mapsto M_n : X_i \mapsto [0, 1]$$

$$\mathcal{S}(D_n)(X_i) = M_n(X_i) \in [0, 1]$$

Prediction performance

Using Brier's score, define

a) the prediction performance of a **deterministic** model m

$$\tilde{\mathrm{BS}}(m) = \mathrm{E}_{Y_i, X_i}\left[\{Y_i - m(X_i)\}^2\right],$$

b) the **conditional prediction performance** of a selected model

$$\mathrm{BS}(M_n) = \mathrm{E}_{Y_i, X_i}\left[\{Y_i - \mathcal{S}(D_n)(X_i)\}^2 \mid D_n\right],$$

c) the **expected prediction performance** of a strategy at sample size n

$$\mathrm{EBS}(\mathcal{S}, n) = \mathrm{E}_{D_n}\left(\mathrm{E}_{Y_i, X_i}\left[\{Y_i - \mathcal{S}(D_n)(X_i)\}^2 \mid D_n\right]\right).$$

Example: GBSG-2 study

The GBSG-2 study is a prospective controlled clinical trial on the treatment of primary node positive breast cancer which included **686** patients.

The prognostic factors:

age, tumor size and grade, number of positive lymph nodes, estrogen and progesterone receptors.

are available to predict the **recurrence free survival status***

$$Y_i(t) = \mathcal{I}\{T_i > t\}.$$

*Note: We deal with censored data using inverse of the probability of censoring weighed (IPCW) statistics.

Rival strategies

in *R* notation:

$$\begin{aligned}\text{Cox} = &\, cph(Surv(time, status) \sim age + tsize + grade.bin \\ &+ pnodes + progrec + estrec, data = GBSG2, surv = TRUE)\end{aligned}$$

$$\begin{aligned}\text{MFP} = &\, mfp(Surv(time, status) \sim fp(I(age/50), df = 4, select = 0.05) \\ &+ grade.bin + fp(I(exp(-.12 * pnodes)), df = 4, select = .05) \\ &+ fp(I(progrec), df = 4, select = .05), data = GBSG2, family = cox)\end{aligned}$$

$$\begin{aligned}\text{RSF} = &\, rsf\,(Survrsf\,(time, status) \sim age + tsize + grade.bin \\ &+ pnodes + progrec + estrec, data = GBSG2, forest = TRUE)\end{aligned}$$

$$\begin{aligned}\text{CoxSpline} = &\, cph(Surv(time, status) \sim rcs(age) + rcs(tsize) + grade.bin + pnodes \\ &+ rcs(progrec) + rcs(estrec), data = GBSG2, surv = TRUE)\end{aligned}$$

Estimation of expected performance

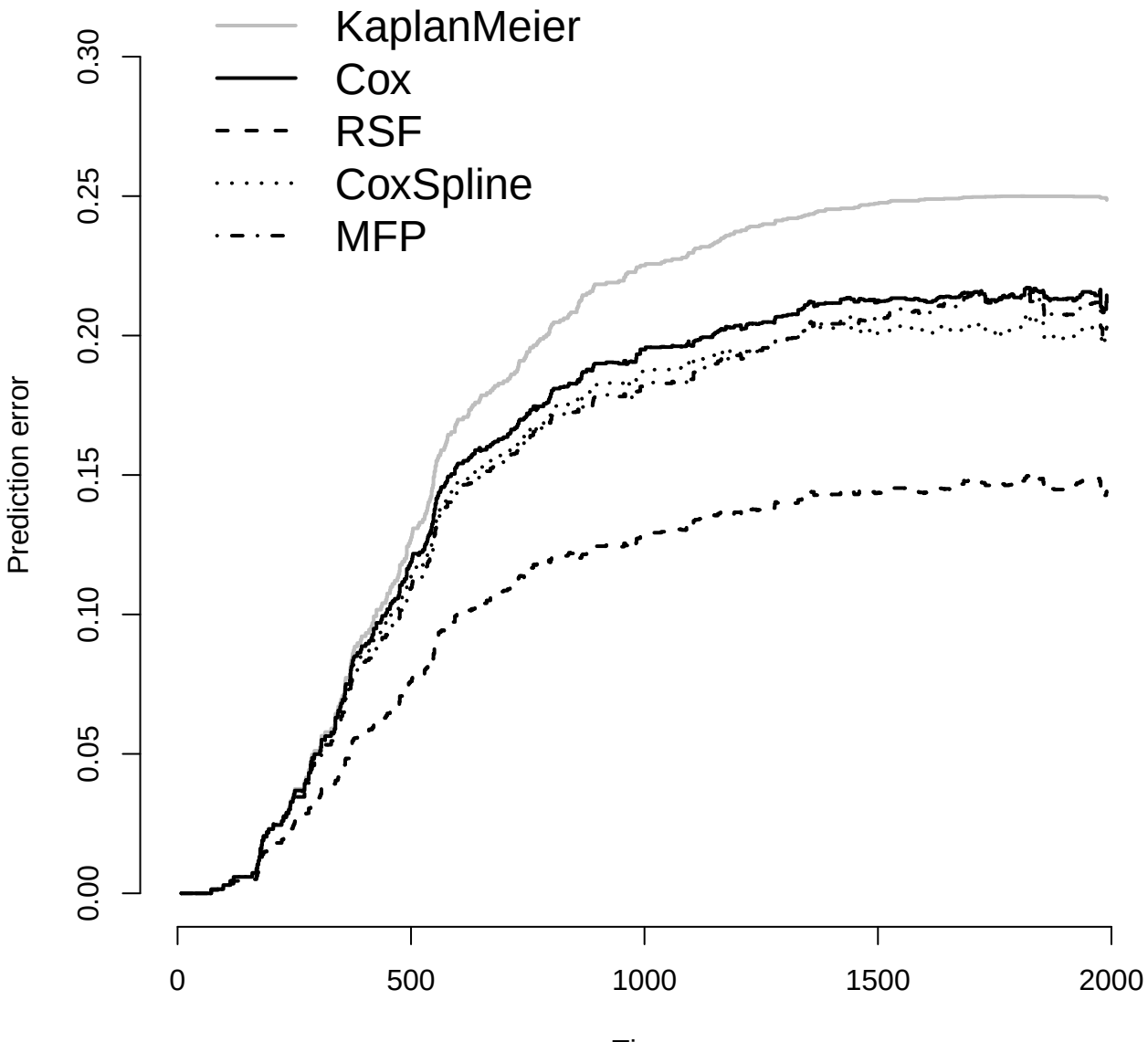
Apparent or re-substitution estimate:

$$\texttt{AppErr}(t) = \frac{1}{n} \sum_{i \in D_n} W^*(t, X_i) \left\{ Y_i(t) - \mathcal{S}(D_n)(t, i) \right\}^2$$

Overrestimates the conditional performance of the model

*Inverse of the probability of censoring weights

Apparent performance



Estimation of expected performance

Generate B bootstrap training sets $D_1^*, \ldots, D_B^*$

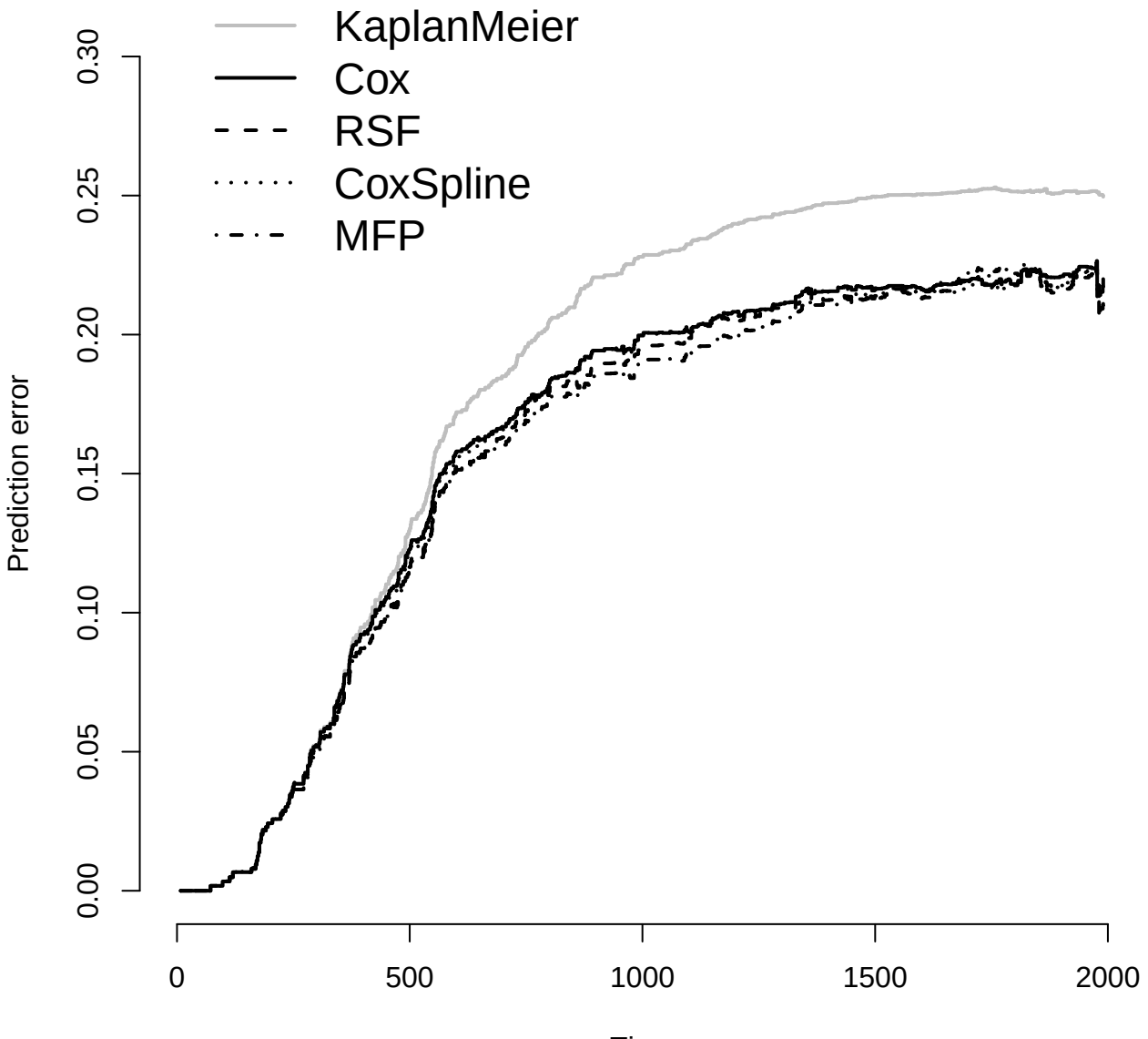
- n times with replacement (ordinary bootstrap)
- $m < n$ times without replacement (subsampling bootstrap).

Bootstrap cross-validation estimate:

$$\texttt{BootCV}(t) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{n_b} \sum_{i \notin D_b^*}^{n} \tilde{W}(t,i) \left\{ Y_i(t) - \mathcal{S}(D_b^*)(t, X_i) \right\}^2$$

Underestimates the expected performance of the strategy at sample size n because the bootstrap samples contain less information than the full sample.

Bootstrap cross-validation performance (B=200,m=500)



The .632+ bootstrap estimate*

With

$$\hat{\omega}_{632+}(t) = .632 / \left(1 - .368 \frac{BootCV(t) - AppErr(t)}{NoInf(t) - AppErr(t)}\right)$$

where

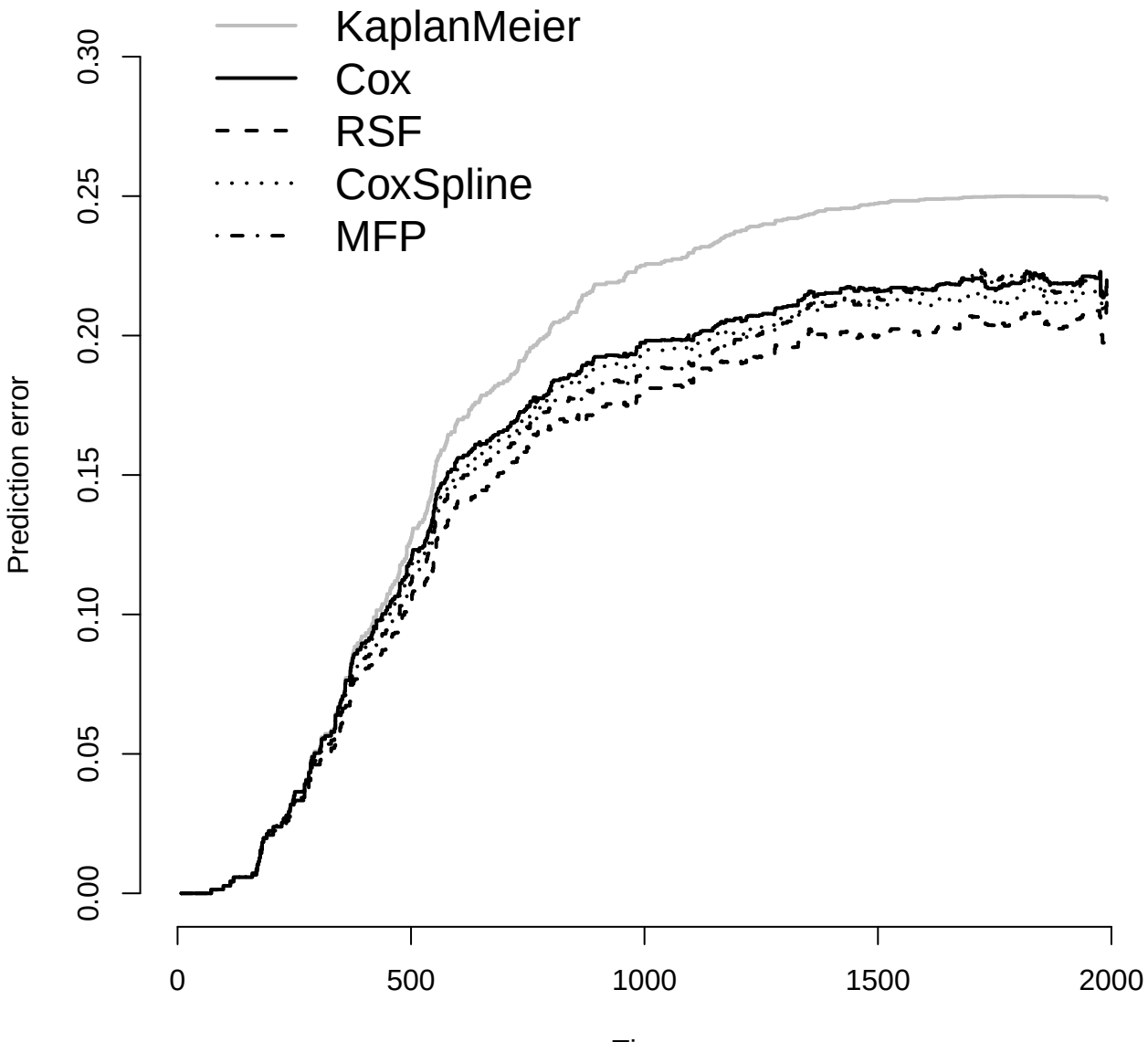
$$NoInf(t) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \bar{W}(t, i, j) \{Y_i(t) - \mathcal{S}(D_n)(t, X_j)\}^2$$

define

$$\texttt{Bootstrap.632+} = (1 - \hat{\omega}_{632+}) \quad \texttt{AppErr} + \hat{\omega}_{.632+} \quad \texttt{BootCV}$$

*Efron, B. and R. Tibshirani (1997)

.632+ bootstrap estimate



Split sample test*

P-values for one-sided differences in expected prediction performance at sample size m=500 (test set size =186)

Hypothesis	t= 500	t= 1000	t= 1500	t= 2000
$KaplanMeier \leq Cox$	0.0002	< 0.0001	0.0002	0.0031
$KaplanMeier \leq MFP$	0.0004	< 0.0001	0.0004	0.0050
$KaplanMeier \leq CoxSpline$	0.0027	0.0003	0.0005	0.0028
$Cox \leq MFP$	0.0131	0.0089	0.0598	0.1689
$Cox \leq CoxSpline$	0.0889	0.1445	0.1007	0.2408
$MFP \leq CoxSpline$	0.8351	0.9051	0.5361	0.5043

To derive an interpretation for n=686 we need to assume that all strategies improve consistently when the sample size increases.

*van de Wiel, Berkhof,van Wieringen (2009)

Decomposition of the expected prediction performance

Introducing the expected prediction of strategy $\mathcal{S}$ at sample size n:

$$\mathrm{E}_{D_n}\{S(D_n)(x)\} = \mathrm{E}_{D_n}\{M_n(x)\} = m_n(x)$$

yields *

$$\mathrm{EBS}(\mathcal{S}, n) = \underbrace{\mathrm{E}_{X_i,Y_i}\left[\{Y_i - m_n(X_i)\}^2\right]}_{\text{Model accuracy}} + \underbrace{\mathrm{E}_{D_n}\left[\mathrm{E}_{X_i}\{\mathcal{S}(D_n)(X_i) - m_n(X_i)\}^2\right]}_{\text{Model uncertainty}}$$

*$\mathrm{E}_{X_i,Y_i}\mathrm{E}_{D_n}\{\mathcal{S}(D_n)(X_i) - m_n(X_i)\} = 0$

Model uncertainty

Traditional prediction models as derived from logistic or Cox regression

- first select variables and functional form based on the data
- and then estimate parameters, like regression coefficients and baseline risk, to predict risk based on the same data.

This may yield substantial model uncertainty even in large sample sizes

...

ELSEVIER

Journal of Clinical Epidemiology 57 (2004) 1138–1146

Journal of Clinical Epidemiology

Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality

Peter C. Austin[a,b,c,*], Jack V. Tu[a,b,c,d,e]

Abstract

Objectives: Automated variable selection methods are frequently used to determine the independent predictors of an outcome. T objective of this study was to determine the reproducibility of logistic regression models developed using automated variable selecti methods.

Study Design and Setting: An initial set of 29 candidate variables were considered for predicting mortality after acute myocard infarction (AMI). We drew 1,000 bootstrap samples from a dataset consisting of 4,911 patients admitted to hospital with an AMI. Usi each bootstrap sample, logistic regression models predicting 30-day mortality were obtained using backward elimination, forward selecti and stepwise selection. The agreement between the different model selection methods and the agreement across the 1,000 bootstrap samp were compared.

Results: Using 1,000 bootstrap samples, backward elimination identified 940 unique models for predicting mortality. Similar resu were obtained for forward and stepwise selection. Three variables were identified as independent predictors of mortality among all bootstr samples. Over half the candidate prognostic variables were identified as independent predictors in less than half of the bootstrap sampl

Conclusion: Automated variable selection methods result in models that are unstable and not reproducible. The variables selected independent predictors are sensitive to random fluctuations in the data.

Keywords: Regression models; Multivariate analysis; Variable selection; Logistic regression; Acute myocardial infarction; Epidemiology

Parameter of interest

Similarly, for most machine learning methods the insides of the models selected based on different bootstrap sets may be pretty unstable.

However, here we are not interested in the insides model, we are interested in the predictions.

Thus, it makes sense to compare modelling strategies in how confident they are about the predictions, at fixed X_i and also across the population.

Confidence scores at individual x

Subject specific value

$$\mathcal{C}_n(\mathcal{S}, x) = 1 - \sqrt{\mathrm{E}_{D_n}\left\{\mathcal{S}(D_n)(x) - m_n(x)\right\}^2}.$$

For most strategies there is no explicit formula (even not asymptotically) for estimating $\mathcal{C}_n$.

Bootstrap estimate

Generate B bootstrap training sets $D_1^*, \ldots, D_B^*$

- n times with replacement (ordinary bootstrap)
- $m < n$ times without replacement (subsampling bootstrap).

The variation of $M_b^* = \mathcal{S}(D_b^*)$, $b = 1, \ldots, B$ around the bagged predictions

$$m_B^*(x) = \frac{1}{B} \sum_{b=1}^{B} M_b^*(x)$$

yield a bootstrap estimate of the confidence score at x:

$$\hat{\mathcal{C}}_{n,B}(\mathcal{S}, x) = 1 - \sqrt{\frac{1}{B} \sum_{b=1}^{B} \{M_b^*(x) - m_B^*(x)\}^2}.$$

Estimate of population level confidence

A population level confidence score can be estimated by two alternative approaches, either by predicting everyone in the original sample using all bootstrap models:

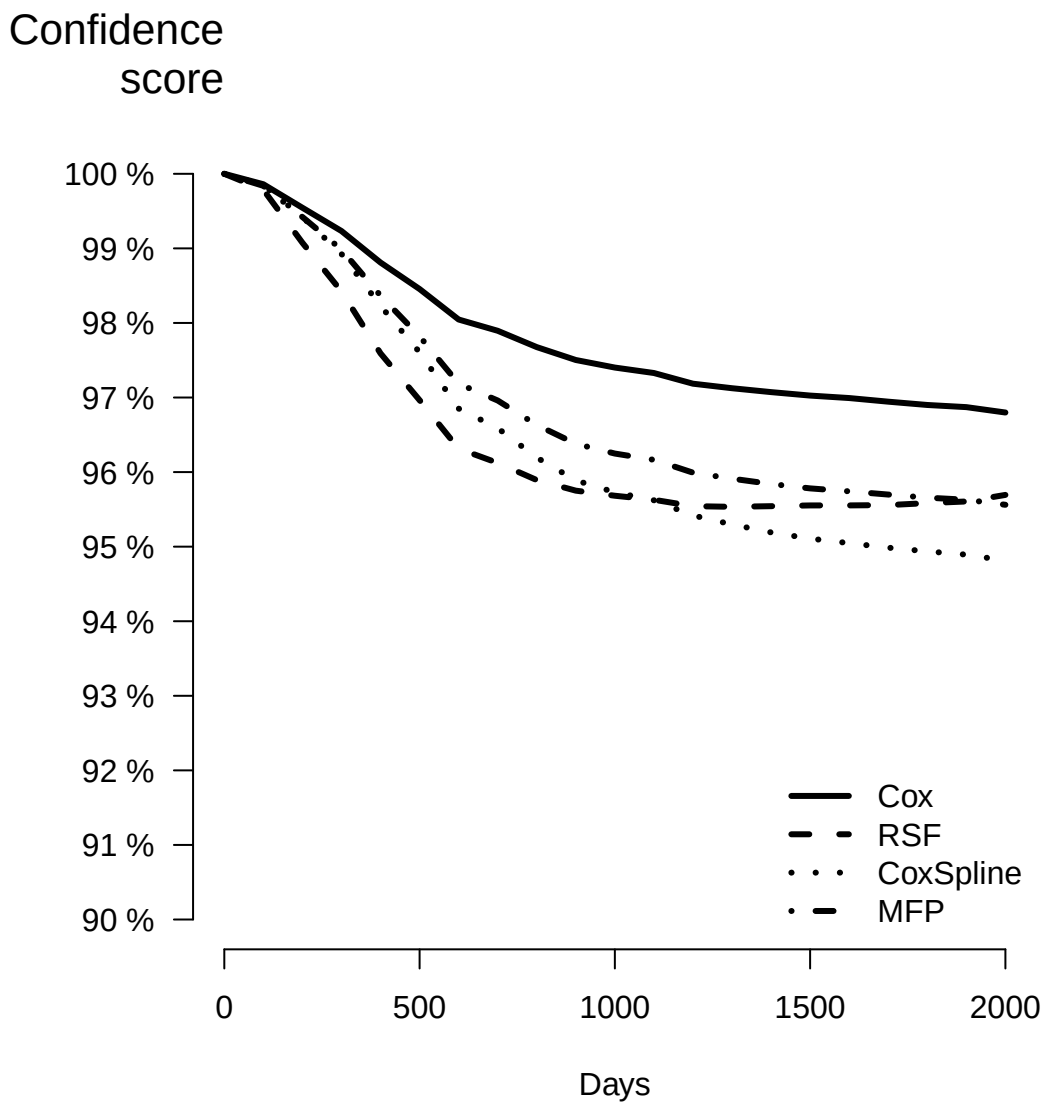
$$\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{C}}_n(\mathcal{S}, X_i),$$

or by only predicting everyone who is not in the current bootstrap training set:

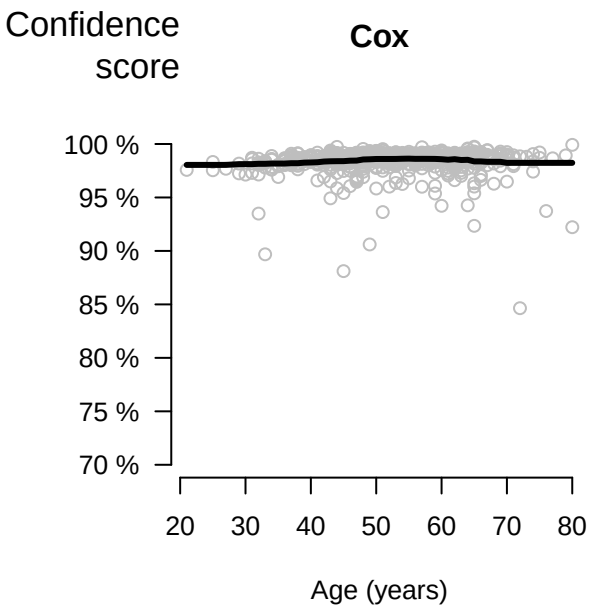
$$\frac{1}{n}\sum_{i=1}^{n}\left[1-\sqrt{\frac{1}{K_i}\sum_{b:i\notin D_b^*}\left\{M_b^*(X_i)-m_B^*(X_i)\right\}^2}\right].$$

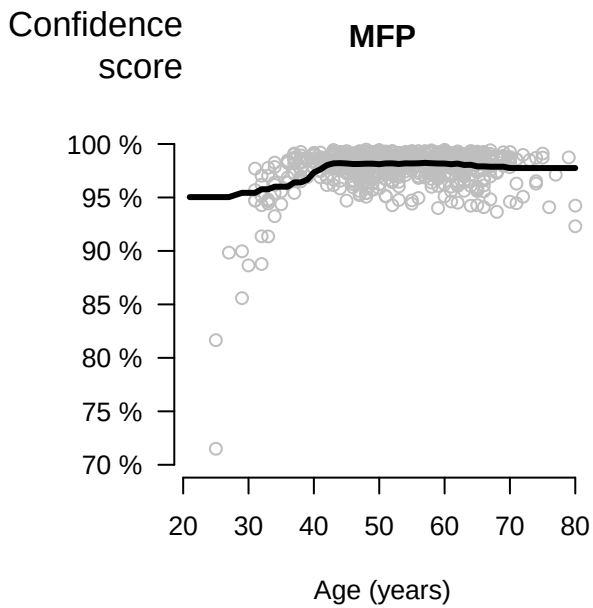
$K_i = \sum_{b=1}^{B}\mathcal{I}\{i \notin D_b^*\}$.

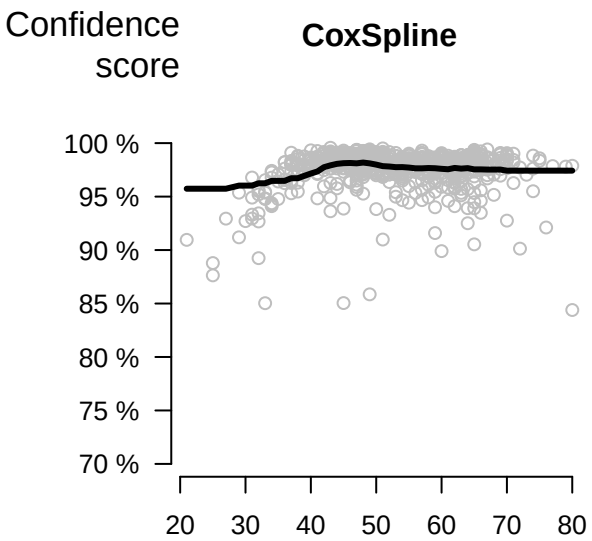
Overall confidence scores

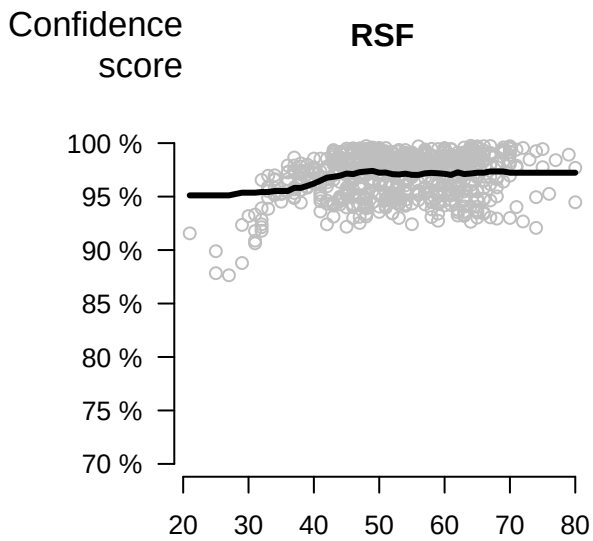


Partial confidence scores along patients' age









Conclusions

- A statistical prediction in medicine is the result of two mappings:
  1. the strategy selects a model
  2. the model predicts the probability of an event
- The prediction performance can be decomposed into model accuracy and model confidence
- The model uncertainty is part of the commonly used estimates of prediction performance.
- The bootstrap 632+ estimate likes random forests.
- The variability of individual predictions due to model uncertainty may be systematically higher for one modelling strategy
- The variability of individual predictions may depend on the patient characteristics.
- Confidence scores may be useful for the patient, and as a *model free* measure of model uncertainty for comparing strategies.