EpiStat
Epidemiology & Statistics

Development and internal validation of a prediction model – *DRAFT*

Erik Lampa

July 3, 2023

Epistat AB
Uppsala Business Park
Virdings allé 32
754 40 UPPSALA
http://www.epistat.se

Contents

Revision History

Revision	**Date**	**Author(s)**	**Description**
1.0	2023-04-12	EL	Created
1.1	2023-04-26	EL	Fix negative follow-up, move 2 patients with ROS1 and 6 patients with BRAF to Other
1.2	2023-06-22	EL	Update calibration algorithm, update the MR and MR + MET-PET analysis, update the analysis of volume versus radionecrosis and CNS progress analysis from logistic regression to Cox regression
1.3	2023-06-26	EL	Add evaluation of GPA, SIR and BDSM scores

This document contains the development and internal validation of a prediction model for all-cause mortality as well as contrasting MR and MET-PET and the association of the total volume with radionecrosis and CNS progress. All analyses were made using R version 4.2.2 with the glmnet, polspline, boot, survival, Epi and rms add-on packages.

1 Data

The dataset consists of 431 observations and 91 variables. Descriptive statistics for the variables selected for inclusion in the prediction model is shown in table

Table 2. Descriptive statistics

	Overall
	(N=431)
Age	
Mean (SD)	67.0 (9.41)
Median [Min, Max]	68.4 [30.9, 92.7]
Lungadenocarcinoma	
No	109 (25.3%)
Yes	322 (74.7%)
SCC	
No	389 (90.3%)
Yes	42 (9.7%)
SCLC	
No	410 (95.1%)
Yes	21 (4.9%)
LCLC	
No	418 (97.0%)
Yes	13 (3.0%)
NOS	
No	407 (94.4%)
Yes	24 (5.6%)
Adenosquamous	
No	426 (98.8%)
Yes	5 (1.2%)
Performance status	
0-1	384 (89.1%)
2+	47 (10.9%)
Mutational status primary tumour	
No mutation	293 (68.0%)
EGFR	39 (9.0%)
ALK	23 (5.3%)
Other	15 (3.5%)
KRAS	45 (10.4%)
Missing	16 (3.7%)
Extracranial disease	
No	78 (18.1%)
Yes	352 (81.7%)
Missing	1 (0.2%)
Disease outside CNS under control	
No	252 (58.5%)
Yes	178 (41.3%)
Missing	1 (0.2%)
Symptomatic CNS disease	
No	65 (15.1%)
Yes	363 (84.2%)
Missing	3 (0.7%)
Size of largest SRS treated meta (mm)	

Mean (SD)	17.7 (7.78)
Median [Min, Max]	18.0 [2.00, 43.0]
Missing	4 (0.9%)
Volume of the first treated SRS (mm^3)	
Mean (SD)	4.36 (4.91)
Median [Min, Max]	2.81 [0.0150, 38.2]
Volume of the largest treated SRS (mm^3)	
Mean (SD)	3.69 (4.25)
Median [Min, Max]	2.37 [0.00660, 32.5]
CNS metastasis at diagnosis of primary tumour	
No	216 (50.1%)
Yes	215 (49.9%)
Comorbidity, cardiovascular disease	
No	311 (72.2%)
Yes	120 (27.8%)
Comorbidity, pulmonary disease	
No	345 (80.0%)
Yes	86 (20.0%)
Leptomeningeal disease	
No	409 (94.9%)
Yes	20 (4.6%)
Missing	2 (0.5%)

Pateints contributed 592.84 person-years of follow-up time with the median follow-up equal to 0.73 years.

2 Development of a prediction model

2.1 Regression model

A regularized Cox model is used to relate the selected baseline variables to survival. Regularization implies shrinking coefficients in the model towards zero to combat likely overfitting when applied to new data. The amount of shrinkage is determined via a budget dictating the size of the coefficients - a smaller budget leads to coefficients closer to zero while a larger budget gives larger coefficients. The size of the budget can be determined via cross-validation

The Cox model is typically written as

$$h_i(t) = h_0(t)e^{x_i^T\beta}$$

where $h_i(t)$ is the hazard for patient i at time t, $h_0(t)$ is the shared baseline hazard and β are the regression coefficients. The quantity $x_i^T\beta$ is called the linear predictor for patient i and e^{β_j} is the hazard ratio for variable j holding all other variables constant.

The coefficients are found by maximizing the partial likelihood

$$L(\beta) = \prod_{i=1}^{m} \frac{e^{x_{j(i)}^T\beta}}{\sum_{j\in R_i} e^{x_j^T\beta}}$$

where R_i is the set of indices j, i.e. all patients who have not yet experienced the event, at time t_j. The baseline hazard $h_0(t)$ does not come into the equation which is often considered a feature of the Cox model.

The optimization problem could be written in the so called Lagrangian form

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left[\frac{2}{n} \left(\sum_{i=1}^{m} x_{j(i)}^T\beta - \log\left(\sum_{j\in R_j} e^{x_j^T\beta} \right) \right) - \frac{\lambda}{2} \sum_{i=1}^{p} \beta_p^2 \right]$$

The above equation says that we should find the value of β that maximizes the above expression, and that is the estimate of the coefficient, $\hat{\beta}$. The term $\frac{\lambda}{2}\sum_{i=1}^{p}\beta_p^2$ dictates the size of all coefficients and acts like a budget. By increasing the value of λ we get a tighter budget and smaller, in absolute value, regression coefficients.

2.2 Determining the budget

A natural question is how big the budget should be, or how much penalization should be used. One way of finding out is by employing cross-validation

1. Divide the data into K subsets. K is typically 10.
2. for $i = 1, \dots,$ K
   - (a) Set aside subset i. Fit model using the other K - 1 subsets.
   - (b) Evaluate the model on subset i and save the result.
3. Average the K results.

The algorithm uses a sequence of 500 values of λ and 10-fold cross-validation is applied at each value

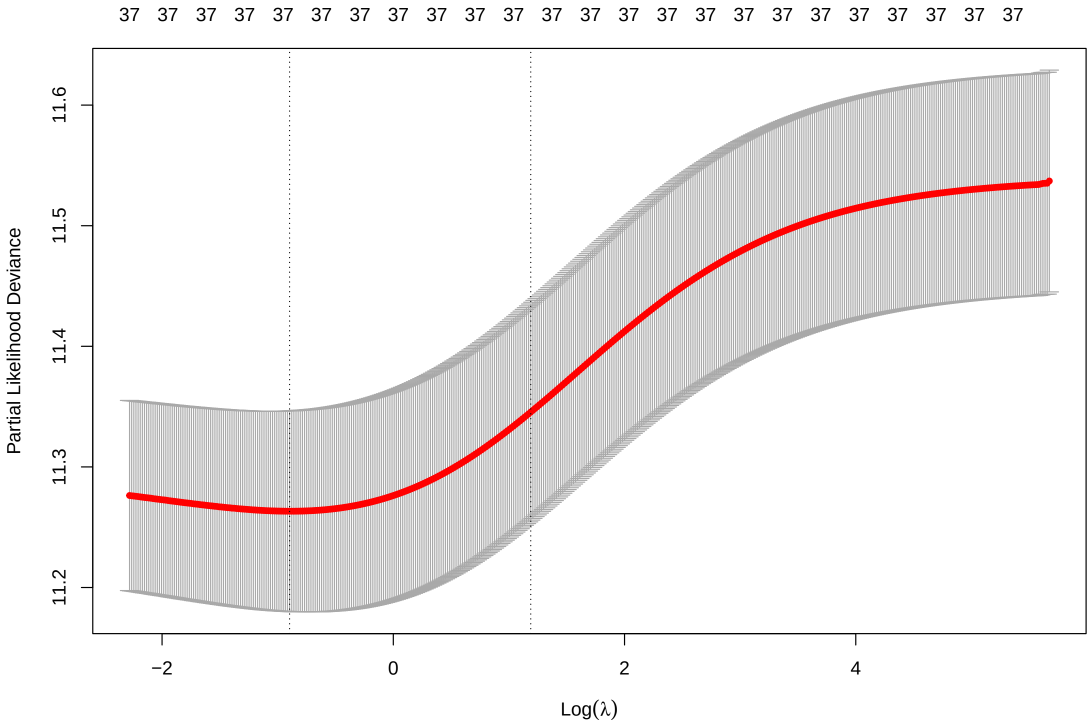

Figure 1. Cross validated partial likelihood deviance (smaller is better) versus log(λ) for a series of 500 λ values. The left dotted vertical line indicates the optimal λ value giving the best fit and the right dotted line is the maximum λ value with a deviance within one standard error of the minimum. The numbers on the top is the number of non-zero coefficients in the model.

The coefficients in the model can be visualized in a similar manner

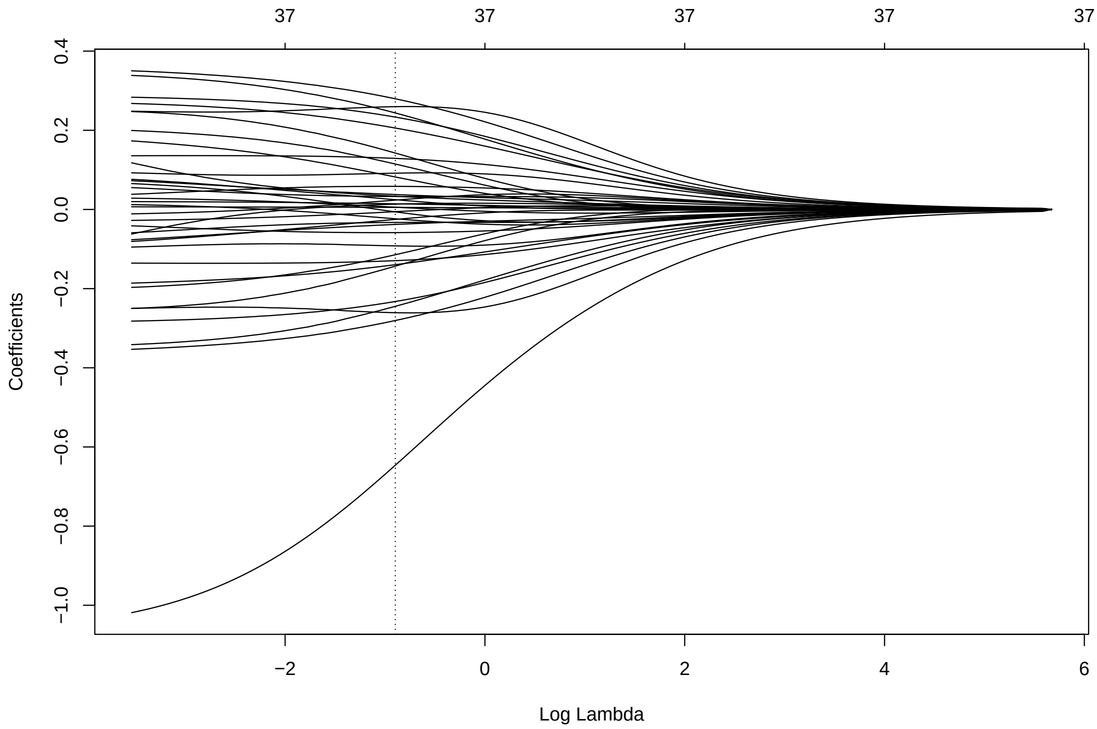

Figure 2. Model coefficients versus $\log(\lambda)$ for a series of 500 λ values. The coefficients get smaller as λ increases to the right and the budget controlling the coefficient sizes get smaller. The dotted line is the optimum value of λ giving the best model fit.

As as evident from figure 2, the optimal amount of shrinkage is not large.

2.3 Validating the prediction model

A prediction model can have great in-sample performance but my fail to be of use if applied in new data. Validation of a prediction model is thus very important. External validation is preferable but is often not possible without external data so a rigorous internal validation is the best option. Cross-validation could be used but set aside data that could be used for model building and the bootstrap is often preferable. The Efron-Gong optimism bootstrap works as follows

1. Fit model in all data and evaluate the performance C_{orig}.
2. Obtain R bootstrap samples, R should be at least 200 - 300.
3. for $i = 1, \ldots, R$
    (a) Fit model from scratch in bootstrap sample i and evaluate the performance in sample i, C_{boot}.
    (b) Evaluate the model in the original data, $C_{b,orig}$.
    (c) Calculate the optimism, $O_i = C_{boot} - C_{b,orig}$.
4. Calculate the average optimism, $\bar{O} = \frac{1}{R}\sum_{i=1}^{R} O_i$.
5. Calculate the optimism corrected performance $C_{orig,corr} = C_{orig} - \bar{O}$.

A common performance measure for Cox models is the C-index which is a measure of discrimination. The C-index, or concordance index, can be thought of the probability that the model will assign a higher risk to a randomly selected case from the patient population than a randomly selected non-case.

The model has an in-sample, or apparent, C-index for one year survival of 0.689. The bootstrap validated C-index, using 250 bootstrap replicates, is 0.670 indicating minor overfitting.

2.4 Calibration

To be useful, a prediction model has to discriminate well and provide risk estimates that are *calibrated*. If a model estimates a one year survival probability of 80% for a patient, we say that the model is well calibrated if 80% survives after one year in a population with the same risk profile as the patient. Calibration can be assessed much like the C-index using the optimism bootstrap and is best visualized using a calibration curve.

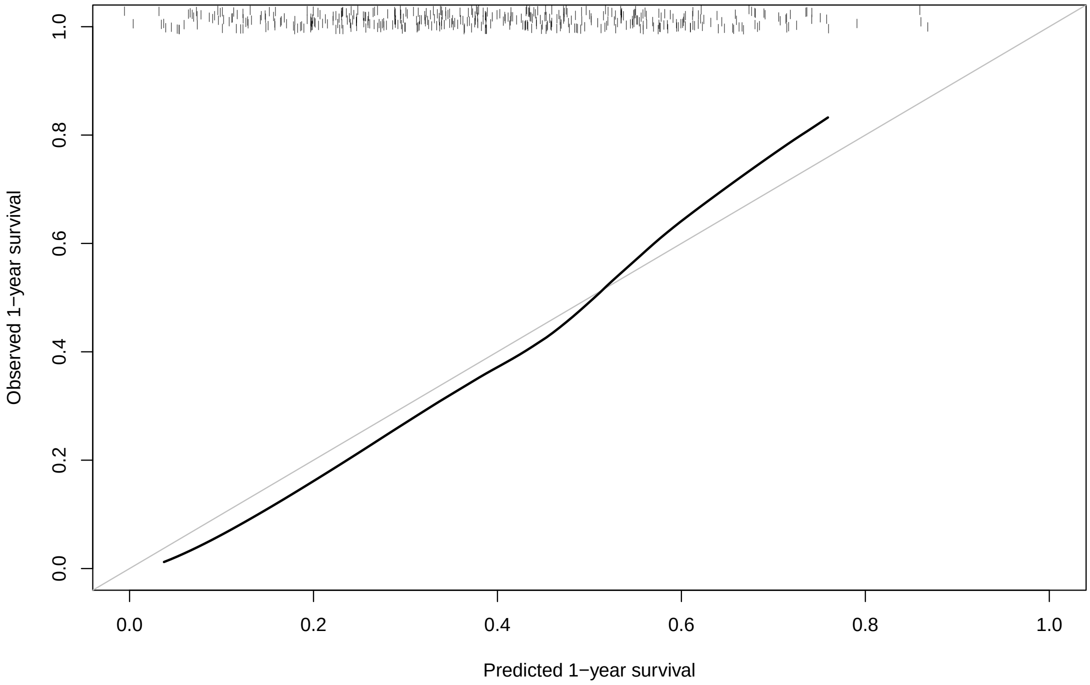

Figure 3. Calibration plot of the predicted (horizontal axis) and the observed (vertical axis) survival probabilities. The grey 45 degree line is the line of perfect calibration where the predicted and the observed survival probabilities match. The black bold line is the optimism corrected estimate of the calibration. Predicted one year survival probabilities are shown as tick marks at the top of the figure.

The calibration is generally very good. Some evidence of overestimation of lower survival probabilities (< 0.5) can also be seen. The mean absolute error (integrated calibration index, ICI) is 0.04 and the 90^{th} quantile is 0.06 meaning that on average, estimated one year survival is off by about 4% from the true value and nine out of ten predictions are within 6%.

2.5 Simplifying the model

A model with 37 coefficients may be to big for clinical use. One strategy of reducing the model is to approximate the model predictions to a high degree using fewer variables

1. Generate model predictions for all patients.
2. Fit a linear model with all variables. This model will fit the data perfectly with $R^2 = 1$.
3. Perform a backwards stepwise deletion of variables until R^2 is below some threshold, e.g 0.95.
4. Refit the linear model with the remaining variables and save the coefficient estimates.

The resulting model will inherit all shrinkage from the full model and with fewer variables but the same predictive ability. Using this strategy, a reduced model with 9 variables and 12 coefficients was found with an $R^2 = 0.98$ against the full model.

2.6 Predicting one year survival

One year survival predictions can be obtained for any patient by using the coefficients and mean values presented in table 3 and the equation

$$\hat{S}(1) = S_0(1)^{e^{(\sum_{i=1}^{p} \beta_i (x_i - \bar{x}_i))}}$$

for variables $i = 1, \ldots, p$. $S_0(1)$ denotes the baseline survival at one year and $\bar{x}_i$ is the mean value for variable i.

Variable	Coefficent	Mean value
Disease outside CNS under control		
Yes	-0.472	0.399
Mutational status primary tumour		
EGFR	-0.351	0.093
ALK	-0.830	0.054
Other	-0.130	0.037
KRAS	-0.208	0.108
Performance status		
2+	0.581	0.115
Age	0.015	66.94
log(Volume of the first treated SRS)	0.073	0.782
Extracranial disease		
Yes	0.486	0.819
CNS metastasis at primary tumour		
Yes	-0.235	0.511
Lungadenocarcinoma		
Yes	-0.217	0.741
Symptomatic CNS disease		
Yes	0.261	0.848

Table 3. Regression coefficients and mean values of the variables in the reduced prediction model.

For example, suppose a patient presents with the following variables

- Disease outside CNS under control: No
- Mutational status of primary tumour: EGFR
- Performace status: 2+
- Age: 65

- log(Volume of the first treated SRS): -1
- Extracranial disease: No
- CNS metastasis at diagnosis of primary tumour: Yes
- Lungadenocarcinoma: No
- Symptomatic CNS disease: Yes

Using the coefficients and mean values in table 3, along with the baseline survival estimate of 0.385, a prediction of the one year survival probability can be estimated as (equation shortened for readability)

$$\hat{S}(1) = 0.385^{e^{-0.472(0-0.399)-0.351(1-0.093)-0.830(0-0.054)+\ldots+0.261(1-0.848)}} = 0.391$$

A way to visualize the model, and to provide a simple prediction tool, is to create a nomogram.

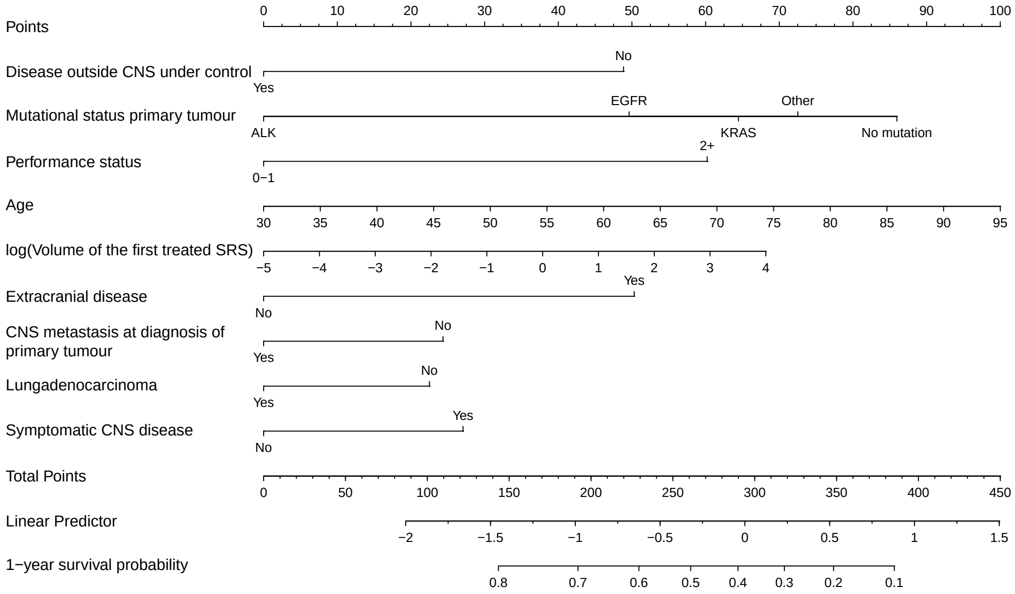

Figure 4. Nomogram depicting the reduced model. Each variable contribute points, measured by the top ruler, to the prediction. The predicted probability can be estimated by following a vertical line from the total points down to the last row which shows the one year survival probability estimates. The example patient gets a total score of 290 which is slighly less than 0.4.

3 Contrasting MR and MET-PET

The output below contains the 2x2 table with MET-PET diagnosis is correct (rows) tabulated against MR diagnosis is correct (columns) along with various statistics

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No   3   2
##        Yes  2  20
##
##                Accuracy : 0.8519
##                  95% CI : (0.6627, 0.9581)
##     No Information Rate : 0.8148
##     P-Value [Acc > NIR] : 0.4223
##
##                   Kappa : 0.5091
##
##  Mcnemar’s Test P-Value : 1.0000
##
##             Sensitivity : 0.9091
##             Specificity : 0.6000
##          Pos Pred Value : 0.9091
##          Neg Pred Value : 0.6000
##               Precision : 0.9091
##                  Recall : 0.9091
##                      F1 : 0.9091
##              Prevalence : 0.8148
##          Detection Rate : 0.7407
##    Detection Prevalence : 0.8148
##       Balanced Accuracy : 0.7545
##
##        ’Positive’ Class : Yes
##
```

The sensitivity and specificity of MET-PET are 0.909 and 0.600 if MR is considered to be the gold standard.

The probability of radionecrosis and CNS progress can be modeled using logistic regression where the independent variable indicates whether the patient was examined by MR, MR + MET-PET or none of them. CNS progress is defined as either CNS progress inside SRS treated meta or CNS progress outside SRS treated meta.

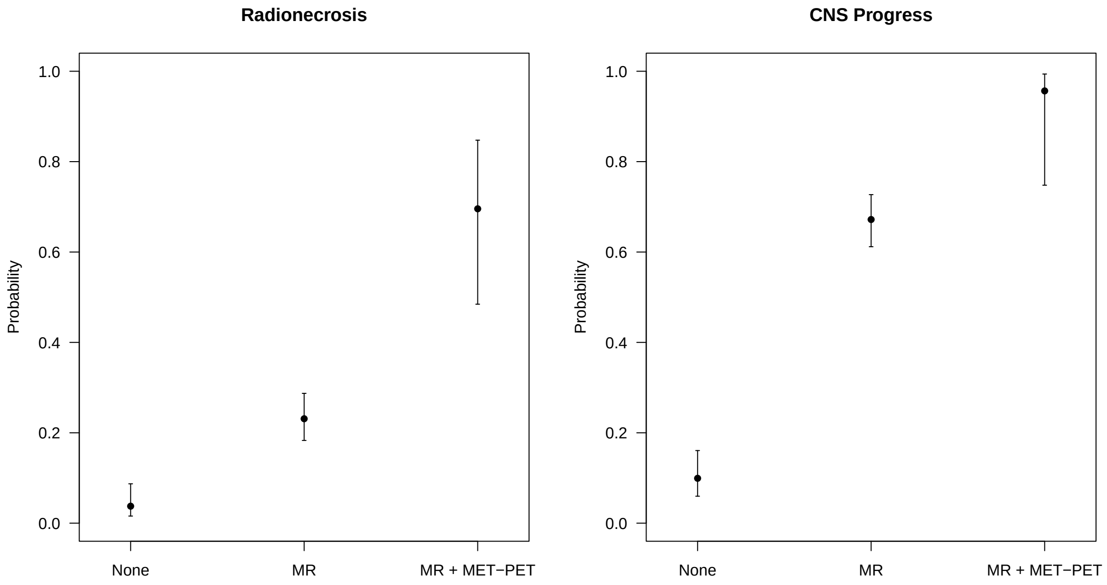

Figure 5. Probability of detecting radionecrosis (left panel) or CNS progress (right panel) depending on whether an MR or an MR + MET-PET examination was done.

4 Associations between total volume and radionecrosis and CNS progress.

Associations between total volume and the time in months to radionecrosis and CNS progress are analyzed using Cox regression models. Total volume is modeled using penalized splines which allows for non-linear associations.

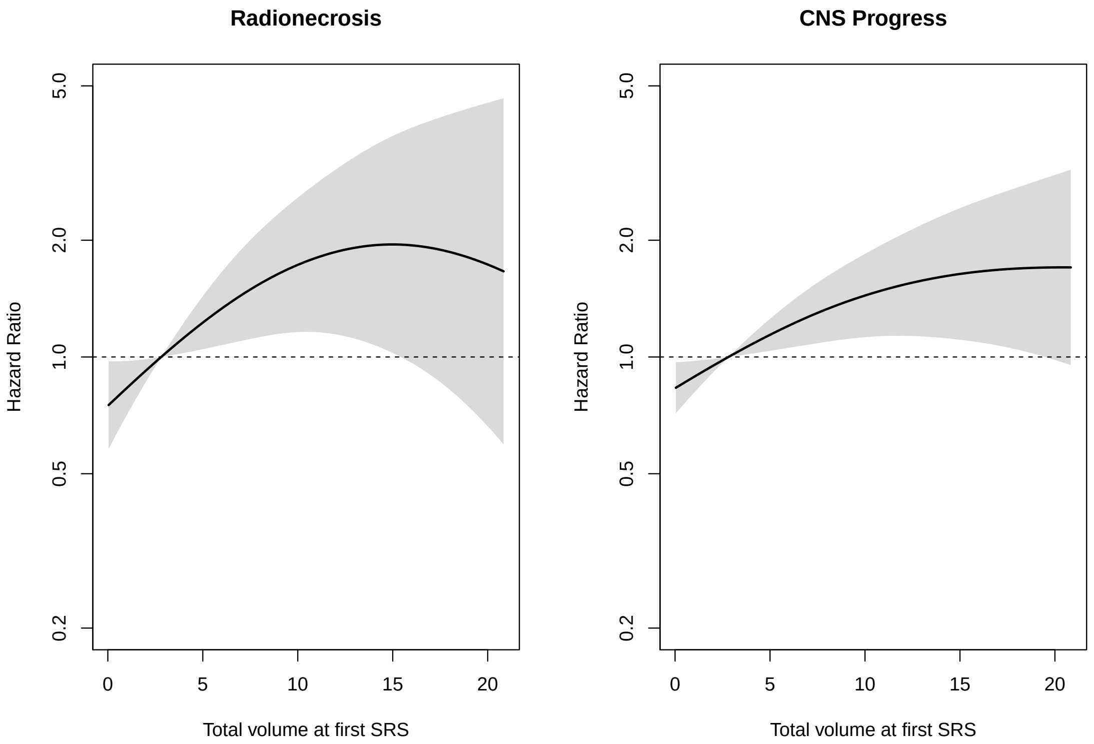

Figure 6. Hazard ratios for radionecrosis (left panel) and CNS progress (right panel) as functions of total volume of the first SRS. Shaded regions are 95% pointwise confidence intervals.

4.1 Prediction of readionecrosis and CNS progress

Calibration curves for the prediction of radionecrosis and CNS progress can be constructed in a similar fashion as before using the Efron-Gong bootstrap (figure 7). While the calibration curves were close to optimal, the ranges of predicted risks were narrow.

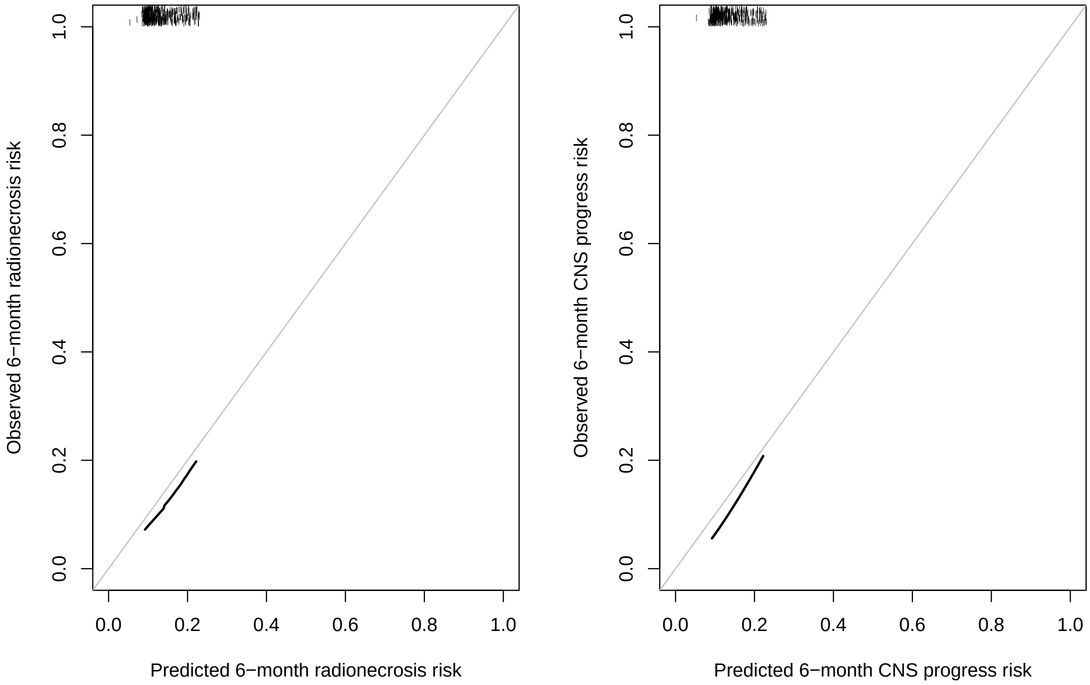

Figure 7. Calibration plots of the predicted (horizontal axis) and the observed (vertical axis) risks of radionecrosis (left) and CNS progress (right). The grey 45 degree lines are the lines of perfect calibration where the predicted and the observed risks match. The black bold lines are the optimism corrected estimates of the calibration. Predicted six month risks are shown as tick marks at the top of the figures.

5 Contrasting survival predictions with GPA, SIR and BSBM

Scores resulting from GPA, SIR and BSBM were fitted to the data using Cox models yielding re-calibrated survival predictions. All predictions were internally validated using the bootstrap and calibration curves were created and contrasted with the KI prediction model. Table 4 gives the regression coefficients and botstrap validated C-statistic for the models. Let x denote the score from a specific model. The one year survival probability from a specific model can be calculated as

$$\hat{S}(1) = 0.385^{e^{\beta\cdot(x-\bar{x})}}$$

where β and $\bar{x}$ can be read from columns two and three in table 4.

For example, a GPA score of 2 translates to a one year survival probability of $0.385^{e^{-0.52*(2-1.66)}} = 45\%$

Table 4. Regression coefficients, mean values and optimism corrected C-statistics for the different prediction models refitted to the data.

Prediction model	Coefficient	Mean value	C-statistic
GPA	-0.52	1.66	0.632
SIR	-0.26	5.27	0.625
BSBM	-0.46	1.51	0.624

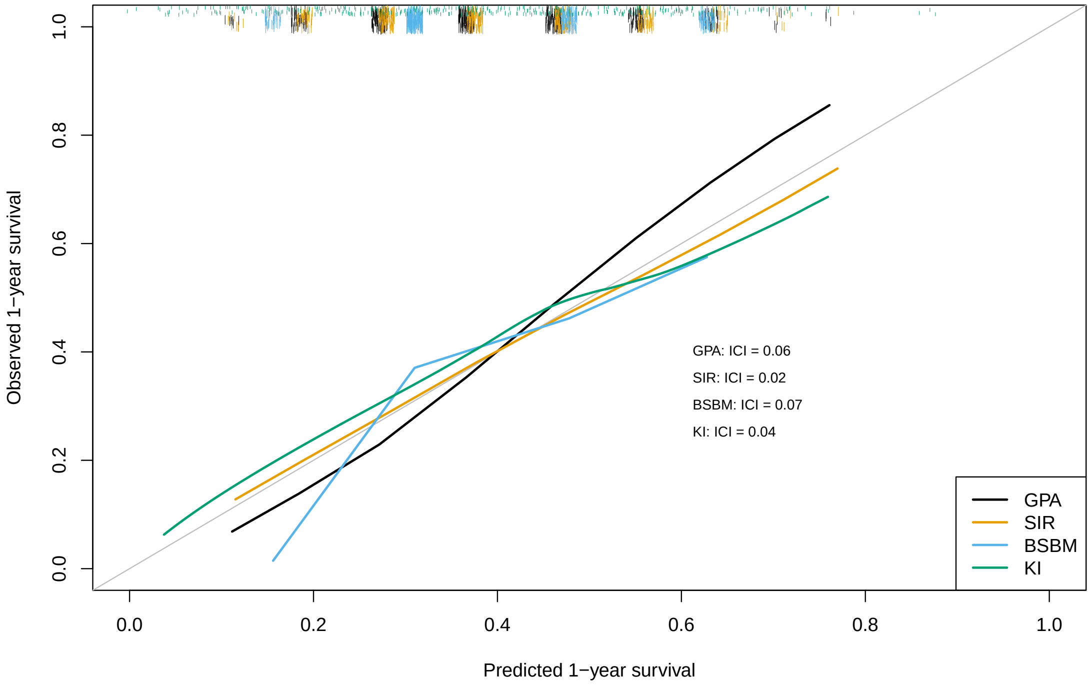

Figure 8. Calibration plot of the predicted (horizontal axis) and the observed (vertical axis) one year survival for the four prediction models. The grey 45 degree line is the line of perfect calibration where the predicted and the observed survival probabilities match. The bold lines are the optimism corrected estimates of the calibration from the different models. Predicted one year survival from the different models are shown as tick marks at the top of the figure. ICI values are the mean absolute errors between the predicted survival probabilities and the observed survival fraction.

It can be seen in figure 8 that the GPA score is underfitting the data, that is low survival probabilities are estimated too high and high survival probabilities are estimated too low, while the KI model shows some overfitting. The calibration is nearly optimal for the SIR score while the BSBM score is miscalibrated for survival probabilities < 0.30. The three models GPA, SIR and BSBM only predicts a small number of survival probabilities while the KI model provides a continuous risk prediction and has superior discrimination.

6 Summary

- One-year survival could be well predicted from a model containing 9 variables. The optimism corrected C-index was 0.67 and the calibration was very good. The average absolute error in one-year survival probabilities was low, 0.04, and the calibration, figure 3, was very good.
- The sensitivity and specificity comparing MET-PET to MR was 0.91 and 0.60 respectively.
- The total volume of the first SRS provided accurate risk estimates for both radionecrosis and CNS progress at 6 months.
- The KI prediction model performed well when contrasted with GPA, SIR and BSBM. SIR predictions showed almost optimal calibration but discrimination, as measured by the C-index, was highest for the KI prediction model.