# Statistical Analysis

## Contents

## Statistical Analysis

### Descriptive

The intention was to statistically investigate the influence of findings from the hemogram on the progression of tumor diseases within the new therapy. For this purpose, a survival time analysis was carried out using the Cox proportional hazard model.

The parameters leukocytes, lymphocytes, neutrophil granulocytes, monocytes, eosinophil granulocytes, C-reactive protein (CRP), albumin, protein, lactate dehydrogenase (LDH) and magnesium were selected in advance and are listed in the following table.

|                          | mean   | sd     |
| ------------------------ | ------ | ------ |
| leukocytes               | 7.75   | 3.37   |
| lymphocytes              | 1.38   | 1.01   |
| neutrophil granulocytes  | 5.38   | 3.14   |
| monocytes                | 0.70   | 0.28   |
| eosinophil granulocytes  | 0.21   | 0.16   |
| C-reactive protein       | 3.31   | 4.76   |
| albumin                  | 3.88   | 0.44   |
| protein                  | 6.72   | 0.62   |
| lactate dehydrogenase    | 297.96 | 283.11 |
| magnesium                | 0.78   | 0.11   |

### Imputation

As values from the hemogram were often missing, the data set had to be imputed for this purpose, whereby predictive mean matching (using the package `mice` (van Buuren and Groothuis-Oudshoorn 2011)) was used in advance for all metric parameters. For this type of imputation, the data set was imputed twenty times in a stack.

### Model

To calculate the Cox proportional hazard model, the selection of the formula was estimated using the AIC and BIC, corrected for the imputed data (more observations and proportion of imputed values), using forward
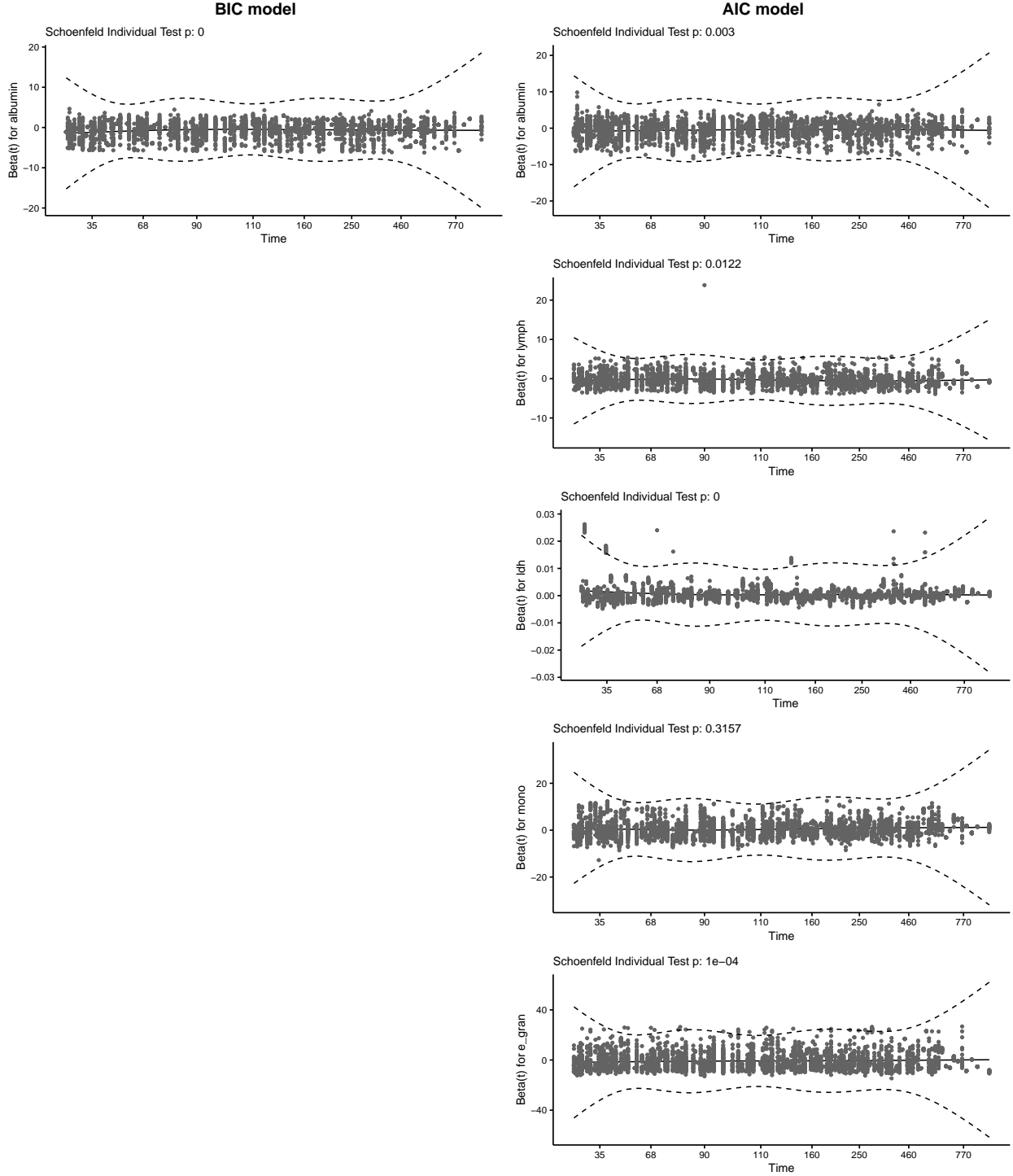
selection, resulting in the following values:

| | BIC model | | AIC model | |
| --- | --- | --- | --- | --- |
| | Covariate | exp(Covariate) | Covariate | exp(Covariate) |
| albumin | -0.69 | 0.5 | -0.50 | 0.61 |
| lymphocytes | | | -0.33 | 0.72 |
| lactate dehydrogenase | | | 0.00 | 1.00 |
| monocytes | | | 0.53 | 1.70 |
| eosinophil granulocytes | | | -0.84 | 0.43 |

In both cases, a higher albumin level is expected to lead to a lower risk of tumor progression. In the case of the AIC model, higher values for lymphocytes and eosinophil granulocytes are also expected to lead to a lower risk of tumor progression. While lactate dehydrogenase has no expected effect on the risk, the expectation for the risk of tumor progression increases with higher levels of monocytes.

## Diagnostic

To investigate whether the models also have proportional hazards (and thus whether the covariates are also independent of time), the Schoenfeld residuals were considered. If the Schoenfeld residuals are considered against time, the individual covariates scatter evenly over time around 0, which means that the proportional hazard assumption appears to be fulfilled.

**BIC model**

Schoenfeld Individual Test p: 0

**AIC model**

Schoenfeld Individual Test p: 0.003

Schoenfeld Individual Test p: 0.0122

Schoenfeld Individual Test p: 0

Schoenfeld Individual Test p: 0.3157

Schoenfeld Individual Test p: 1e−04

The coefficients in both models are well distributed, lying at 0 on the one hand and scattering fairly evenly (except for the edges) on the other. However in both cases, except for the monocytes, all covariates are significant (in some cases even strongly significant), so that the proportional hazard assumption may not apply.
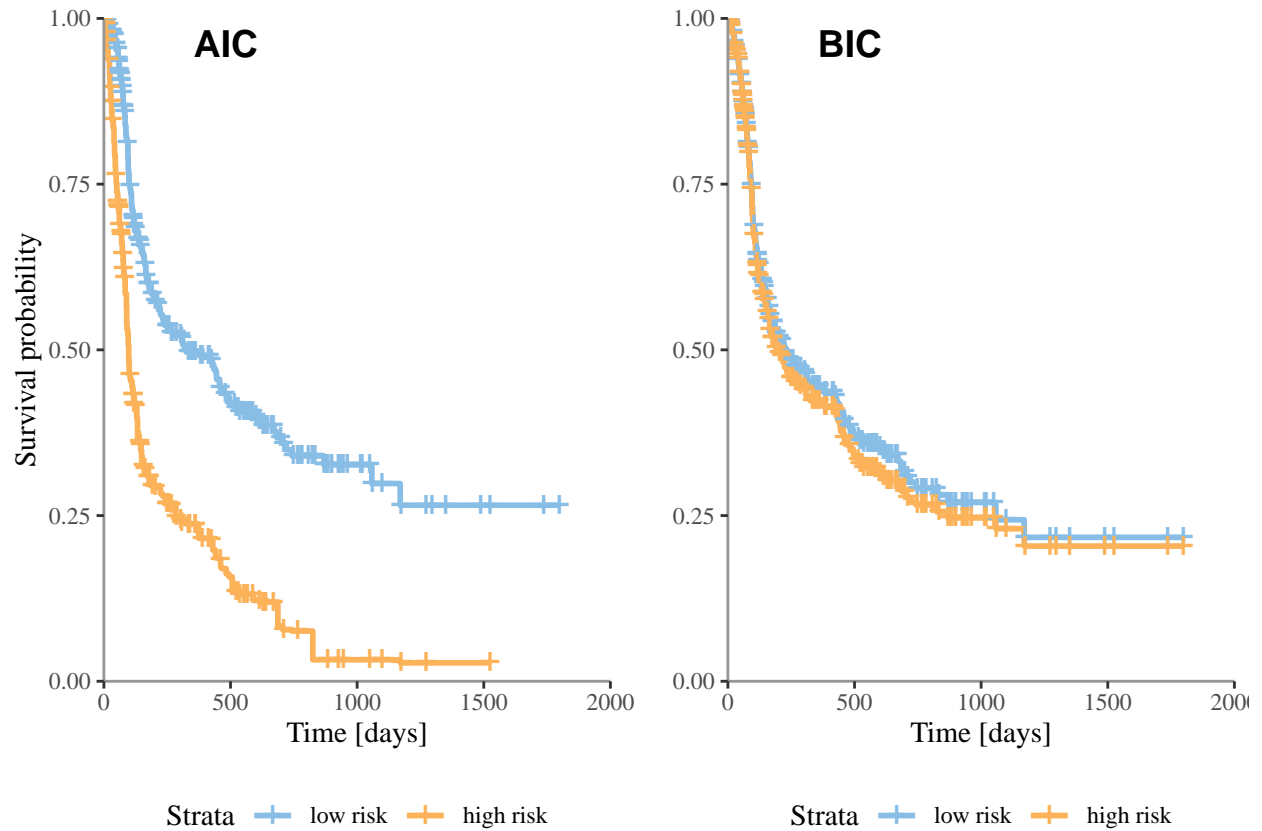
## Cutoff

In order to divide patients into two groups, low risk and high risk, this division was made on the basis of the logrank statistics.

The linear predictor for the AIC and BIC model was calculated for all observations and ordered (from minimum to maximum). Then, the first ordered observation and the remaining observations were divided into two groups and the logrank statistic was estimated. In the next step the two groups were then created based on the first two ordered observations, with the rest of the procedure as above. This procedure was repeated until the logrank statistics were estimated once for all ordered groups.

After all logrank statistics were calculated, the classification that had the maximum logrank statistic was used. Since the best statistics were selected, the log rank test cannot be performed.

For the AIC model, the maximum logrank statistic was at a cutoff value of the linear predictor of -0.12 and for the BIC model at -0.03.

The survival curves took the following form:



## Validation

Cross-validation was implemented to validate the data and thus to assess its quality. The validation was to take place on the basis of the model on the one hand and the cut-off values on the other.

### Model

The model was validated in the same way as described above. For this purpose, the data set was imputed five times and a five-fold cross-validation was performed on the basis of each data set generated. The Concordance Index (check whether the estimated risk from the model matches the event time for two patients) and

the integrated Brier Score (mean square difference between prediction and observation for all patients and timepoints) were used as performance measures.
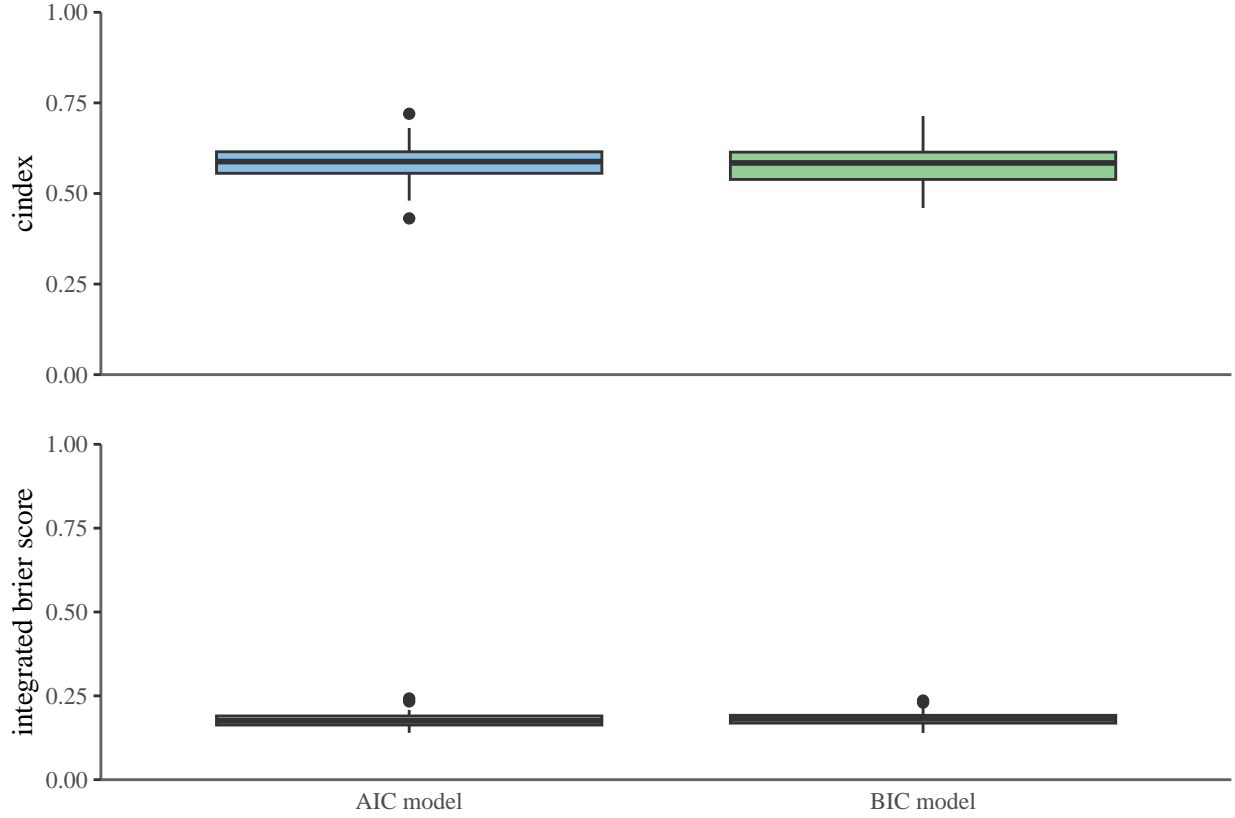


The performance measures are relatively similar for both models. The Cindex is small in both the AIC (0.58 (0.05)) and BIC (0.58 (0.05)) model. The Brier score, on the other hand, appears to be more optimistic for both models (AIC: 0.18 (0.02) and BIC: 0.18 (0.02)). Since the performance measures for the different models are close to each other and even similar between the two selected models, the models appear to be valid.

**Cutoff**

The procedure described above was used to calculate the optimum cut-off. The data set was imputed five times and the best cutoff was calculated for each imputed data set based on the log rank statistics of the linear predictor. Five different cut-off values were obtained for the AIC and BIC models as follows:
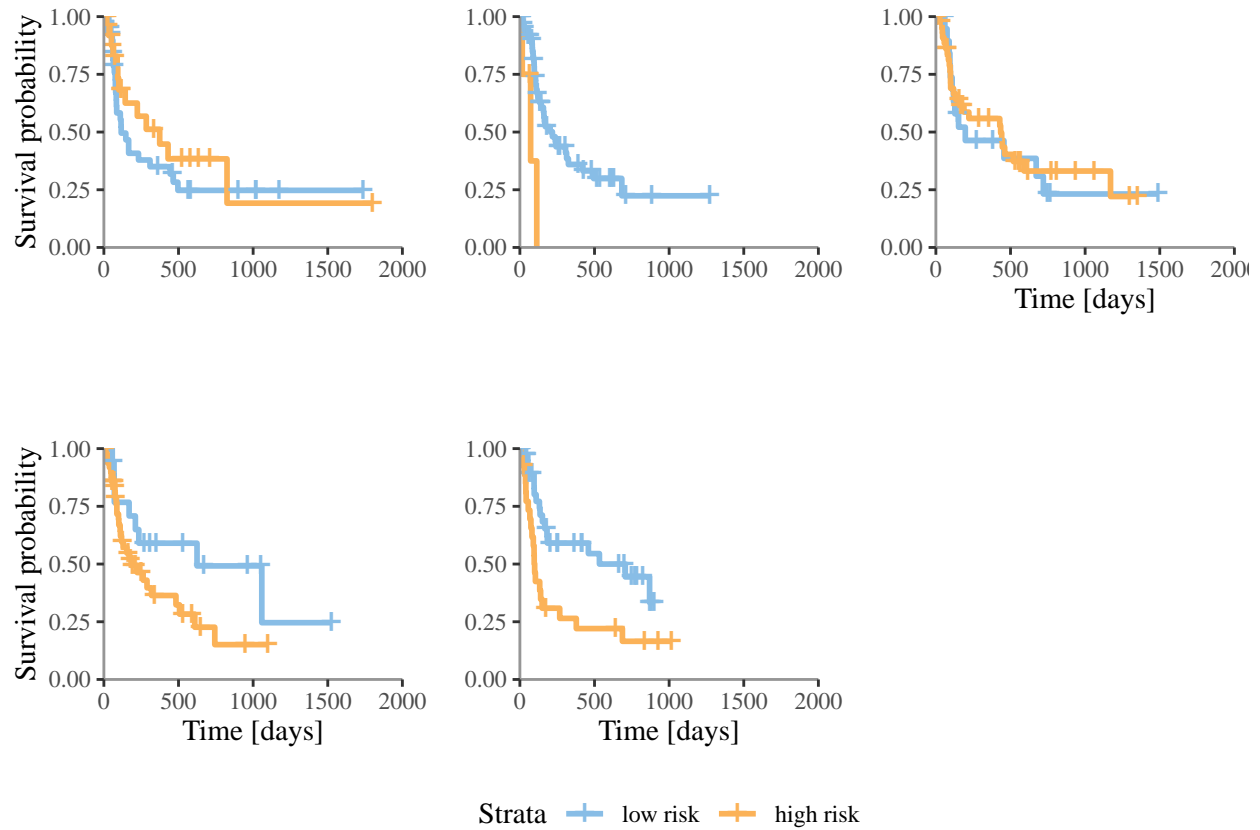
[h]

Table 1: Optimal cutoff Validation

| AIC | BIC |
| --- | --- |
| 0.23 | 0.14 |
| 0.19 | 0.20 |
| 0.09 | -0.12 |
| 0.24 | -0.13 |
| 0.42 | -0.05 |

which thus provide survival time curves for the AIC model
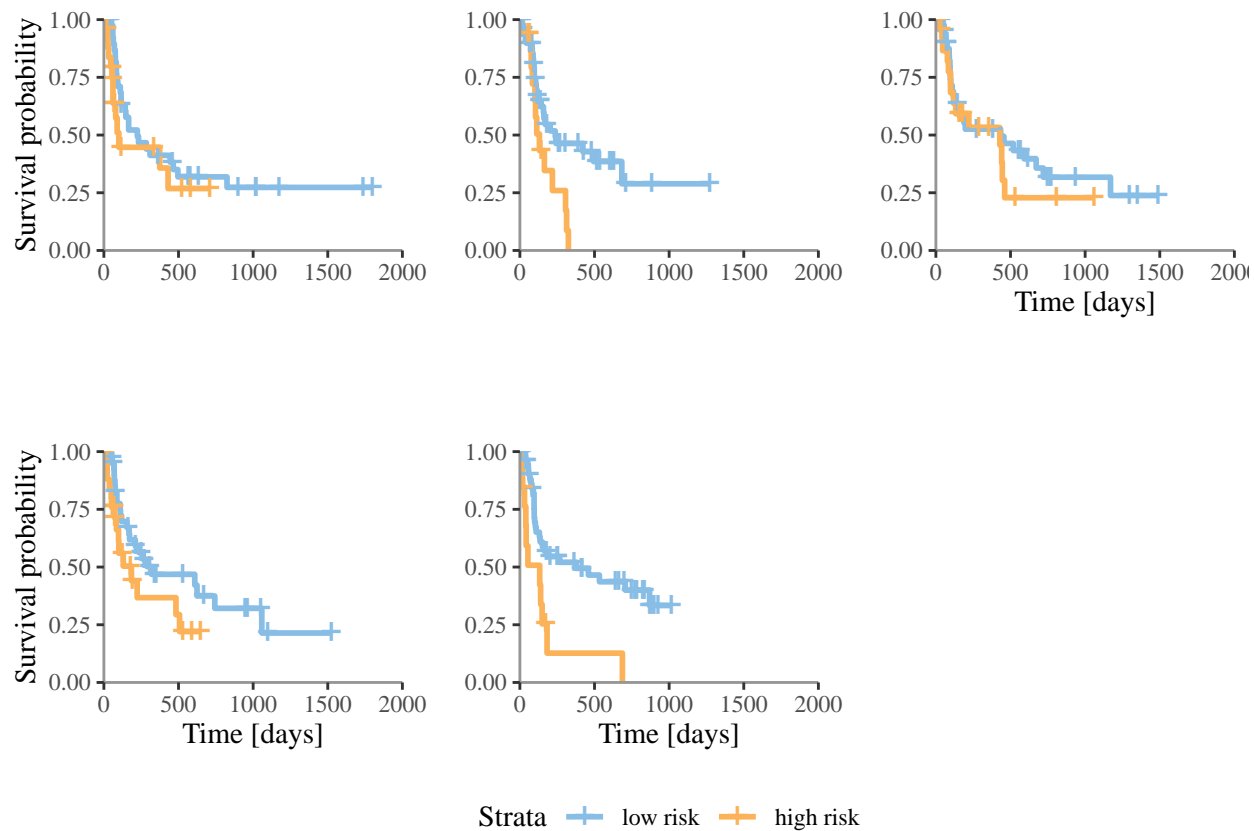
and the BIC model

. It can be seen that the validity of the AIC model is generally better. Even though the differences from the smallest to the largest value in the AIC (0.33) and BIC (0.33) models are exactly the same, the survival time curves of the AIC models do not differ as much as in the BIC model.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.