Statistical Analysis

Contents

Statistical Analysis **1**
Descriptive . . . 1
Imputation . . . 1
Model . . . 2
Diagnostic . . . 2
Cutoff . . . 4
Validation . . . 5

Bibliography **7**

Statistical Analysis

Descriptive

The aim was to estimate a prognostic model for the risk of tumor disease progression using the findings from the hemogram. 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. A predictive mean matching (with the package `mice` (van Buuren and Groothuis-Oudshoorn 2011)) was carried out in advance for all metric parameters, whereby multiple imputation was performed.

Model

To calculate the Cox proportional hazard model, the variables were selected based on the AIC and the BIC using forward selection. Since observations occur multiple times, the AIC and BIC were corrected for the multiple observations on the one hand and for the imputed values on the other. According to Wood, White, and Royston (2008) (and thus also applied from Cheng et al. (2021)), let the data set be repeated M times and let f be the mean proportion of missing values in all variables. Then the loglikelihood in the AIC and BIC is adjusted by the factor

$$\frac{M}{1-f} \quad .$$

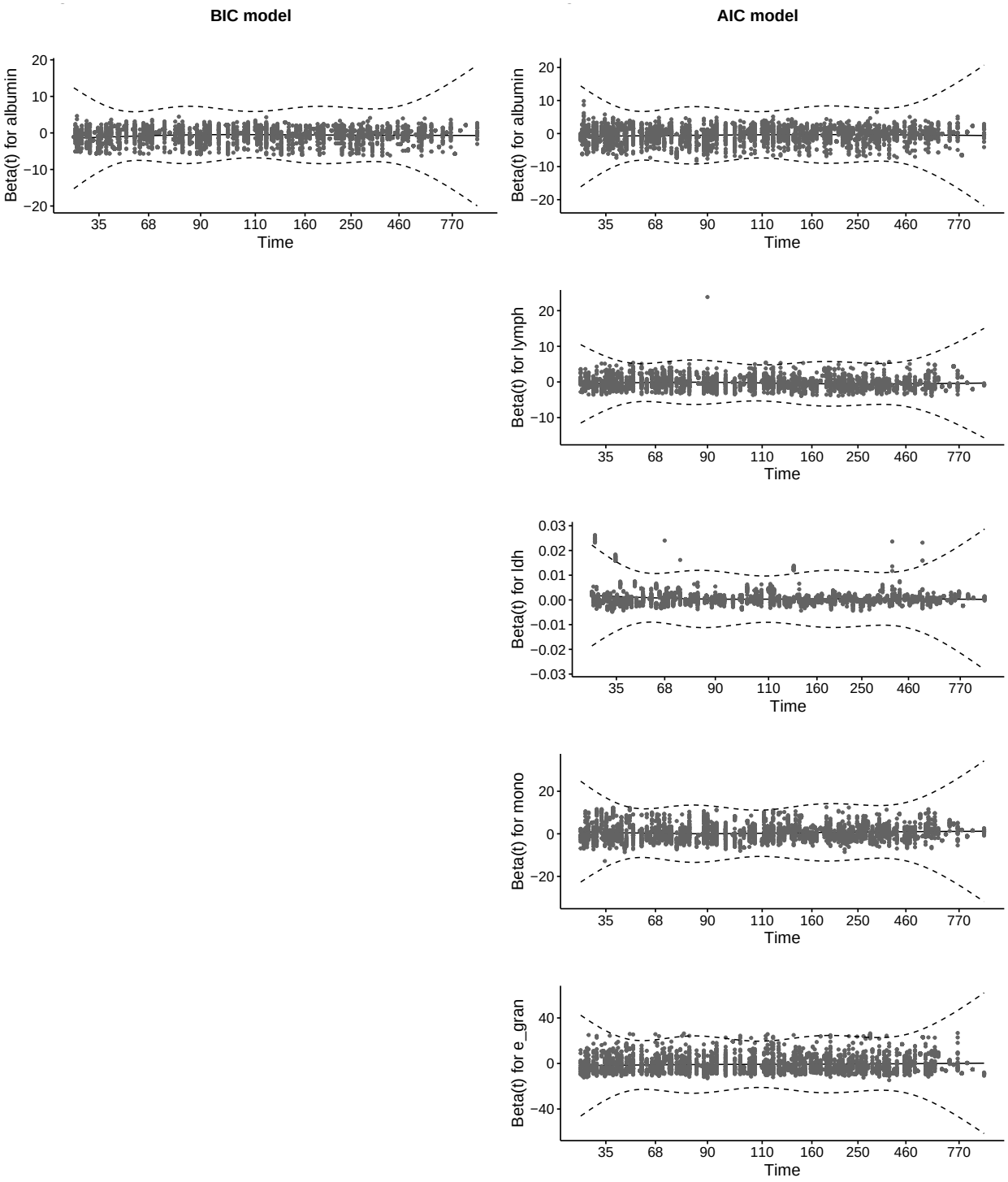
This resulted in the following two models

	BIC model		AIC model	
	Covariable	hazard ratio	Covariable	hazard ratio
albumin	-0.694	0.5	-0.501	0.606
lymphocytes			-0.332	0.718
lactate dehydrogenase			0.001	1.001
monocytes			0.528	1.696
eosinophil granulocytes			-0.842	0.431

In both cases, a higher albumin level is expected to lead to a lower risk of tumor disease 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 a little expected effect on the risk (due to the high values), 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 zero, which means that the proportional hazard assumption appears to be fulfilled.



The coefficients scatter evenly around zero in both models (except at the edges). However, in the score test with the null hypothesis that there is no slope of the covariates against time, all covariates except monocytes are significant in both cases (in some cases even strongly significant), so that the proportionality assumption may not apply.

Cutoff

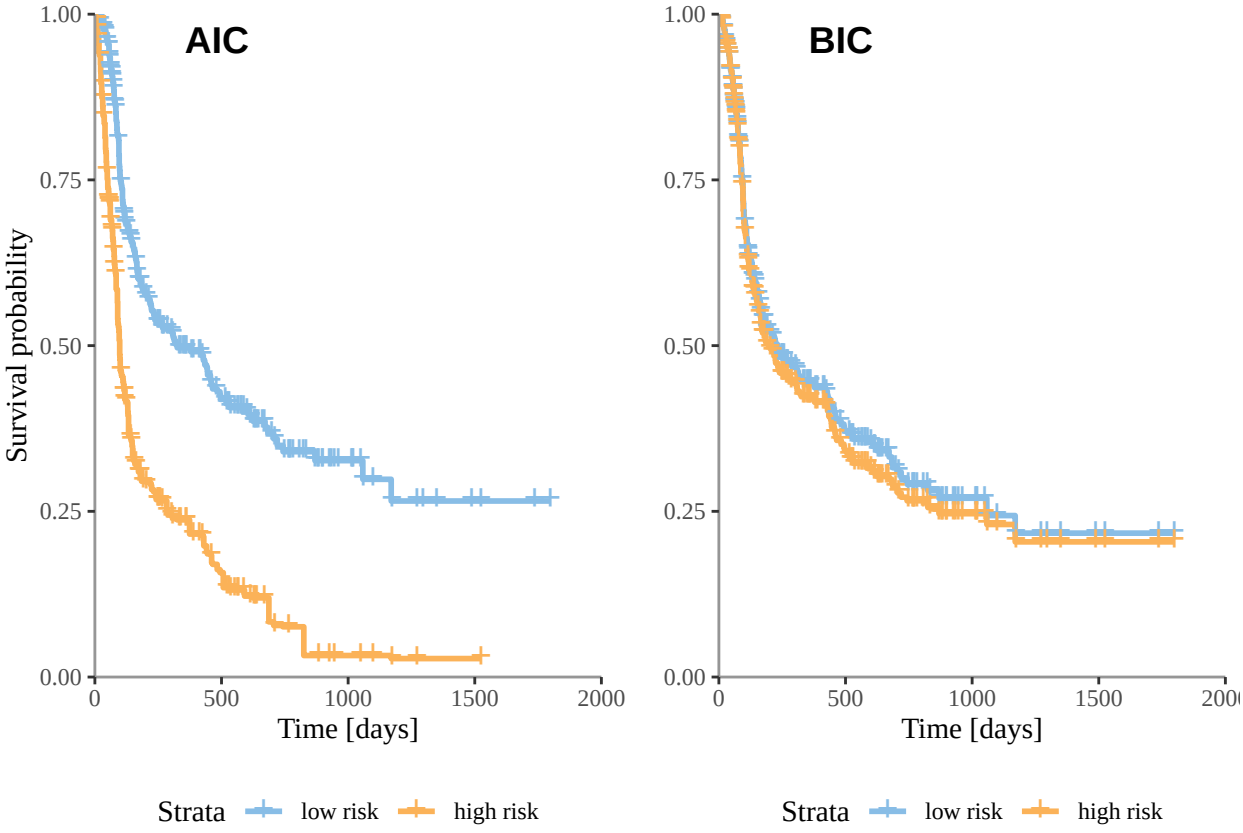
In order to divide patients into two groups, low risk and high risk, this division was made on the basis of the logrank statistics.

The linear predictor for the above 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 groups.

After all logrank statistics were calculated, the classification that had the maximum logrank statistic was used. Thus, the optimal cutoff of the linear predictor is the mean between the highest linear predictor of the ordered lower group and the lowest linear predictor of the ordered upper group. 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:



In the case of the AIC model, a good separation between the two groups can be recognized, while in the BIC model the separation is barely visible. However, it must be expressly warned that these may be too optimistic. This is because the same data is used to estimate both the model and the cut-off values, which can lead to overfitting.

Validation

Cross-validation was implemented to validate the models 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

For validation, a 5-fold simple cross-validation was repeated ten times with 20-fold multiple imputation. Outside the cross-validation, the original data set was divided into test and training data, while within the cross-validation the test and training data were imputed. Variables were selected on the imputed training data using AIC and BIC (taking into account factor $\frac{M}{1-f}$ described above). The Concordance Index (discrimination) and the integrated Brier Score (calibration) were calculated on each imputed test data set of the twenty-fold imputation using the partial test data and the mean value of the Concordance Index and Brier Score was estimated over the twenty imputations.

The mean value and standard deviation of the concordance index and integrated brier score across all repetitions and the cross-validation can be seen in the following table.

	AIC model	BIC model
Concordance Index	0.583	0.579
Integrated brier Score	0.176	0.181

Cutoff

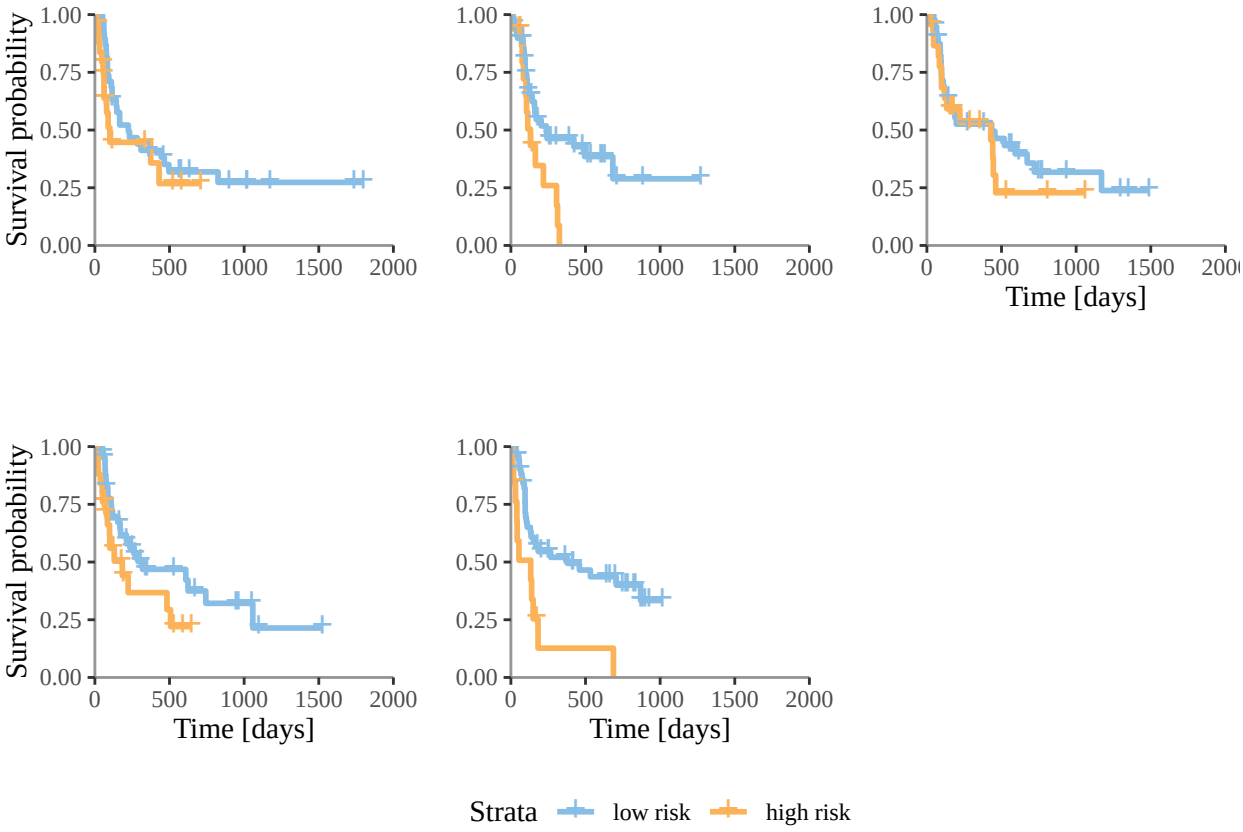
The validity of the cut-off values was analysed separately. For this purpose, before cross-validation without repetition, the original data set was divided into test and training data and in the five-fold cross-validation, the training data was 20-fold and the test data singly multiply imputed. Variables were again selected using the training data with AIC and BIC, and the group division was determined using the maximum logrank statistic. Based on the previously calculated cut-off value of the linear predictor of the maximum logrank statistic, the test data was divided into “low risk” and “high risk” groups and separate survival curves were estimated. The resulting cut-off values for each cross-validation iteration are shown below.

[h]

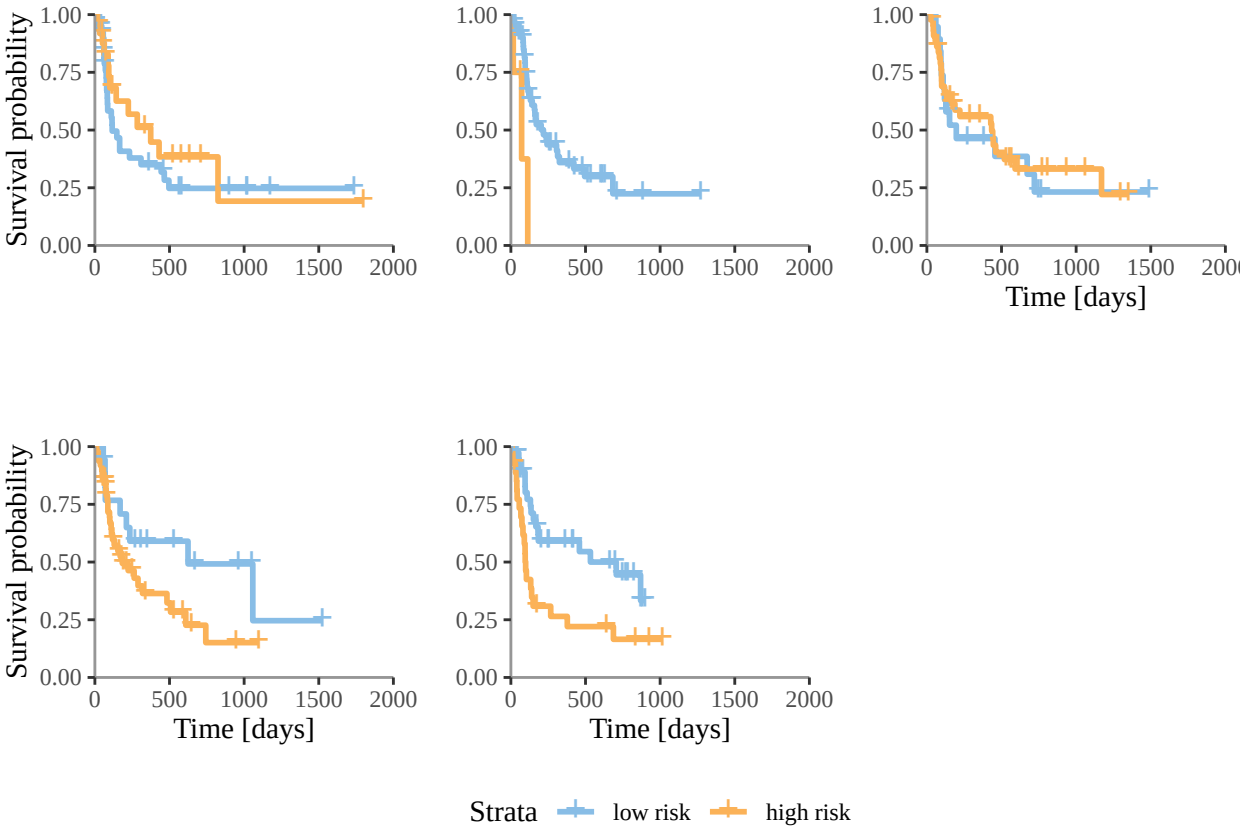
Table 1: Optimal cutoff Validation

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

The resulting survival time curves based on the test data for the AIC model



and the BIC model



are shown. It can be seen that the validity of the AIC model is generally better, since the survival time curves do not differ as much as in the BIC model.

Bibliography

Cheng, Weiting, Roman Hornung, Kai Xu, Cai hong Yang, and Jian Li. 2021. "Complement C3 Identified as a Unique Risk Factor for Disease Severity Among Young COVID-19 Patients in Wuhan, China." *Scientific Reports* 11 (1): 7857. https://doi.org/10.1038/s41598-021-82810-3.

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.

Wood, Angela M., Ian R. White, and Patrick Royston. 2008. "How Should Variable Selection Be Performed with Multiply Imputed Data?" *Statistics in Medicine* 27 (17): 3227–46. https://doi.org/https://doi.org/10.1002/sim.3177.