Pashova, Ulm:

Two Survival Tree Models for Myocardial Infarction Patients

Sonderforschungsbereich 386, Paper 195 (2000)

Online unter: http://epub.ub.uni-muenchen.de/

Two Survival Tree Models for Myocardial Infarction Patients

Victoria Pashova

Kurt Ulm

Institut für Medizinische Statistik und Epidemiologie
Technische Universität München
Ismaninger Straße 22, 81675 München, Germany
E-mail: victoria.pashova@imse.med.tu-muenchen.de

Acknowledgement

This research was supported by grant UL 94/11-1 of the Deutsche Forschungsgemeinschaft.

Many thanks to Dr. Petra Barthel and Prof. Dr. Georg Schmidt for providing the data set and clearing the way through the jungle of medical terminology as well as for their input on the medical feasibility of the models.

We would also like to thank our colleague Ursula Berger for her helpful hints and suggestions on parts of the paper.

Summary

In the search of a better prognostic survival model for post- acute myocardial infarction patients, the scientists at the Technical University of Munich’s “Klinikum rechts der Isar” and the German Heart Center in Munich have developed some new parameters using 24-hour ECG (Schmidt et al 1999). A series of investigations were done using these parameters on different data sets and the Cox – PH model (Schmidt et al 1999, Ulm et al 2000). This paper is a response to the discussion paper by Ulm et al (2000), which suggests a Cox model for calculating the risk stratification of the MPIP data set patients including the predictors ejection fraction and heart rate turbulence. The current paper suggests the use of the classification and regression trees technique for survival data in order to deduct a survival stratification model for the NIRVPIP data set. Two models are compared: one contains the variables suggested by Ulm et al (2000) the other model has two additional variables, namely presence of couplets and number of extra systolic beats in the longest salvo of the patient’s 24-hour ECG. The second model is shown to be an improvement on the first one.

Introduction

A lot of research was done lately in the area of survival analysis involving a wide spectrum of models, including the beloved Cox-PH model (Harrell 1996, Schemper 1996), but also neural networks (Harbeck 2000), and classification and regression trees (Dannegger 2000, Le Blanc 1992). Each of those methods has its advantages and its unresolved problems. The classification and regression trees method is preferred here for its interpretability and its advantage in dealing with interactions.

Continuous variables are often used in medical research, however, for classification and stratification purposes categorical or dichotomized variables are often more useful. For example, for the indicator variable DIABETES it is clear that people with diabetes (DIABETES = 1) have worse prognosis than people without diabetes (DIABETES = 0) when all other factors are excluded. But how can one give such a definite risk stratification when continuous variables, such as ejection fraction or heart rate frequency are involved? The use of survival trees as a modeling method can help in this direction. Not only does a tree give optimal cut points for each variable, but it also gives a corresponding (in general different) optimal cut point for every subsequent subgroup of the data set.

This paper takes another look at the NIRVPIP data set from the perspective of survival trees. The NIRVPIP data set was gathered at the Klinikum rechts der Isar of TUM between November 1994 and November 1999 . It contains 1353 patients (95.6% censoring) with observation periods between 11 and 1751 days. All patients have survived an acute myocardial infarction and are possible candidates for the cardioverter defibrillator – an invasive prophylactic treatment which reduces the risk of sudden cardiac death. The paper discusses two survival tree models. One model in a sense reflects the Cox-PH model used by Ulm et al (2000) onto the space of survival tree models since it also contains the variables ejection fraction and heart rate turbulence as building block variables. The second tree model is an extension and an improvement of the first one as it contains some additional variables. Each of the two models divides the patients into two distinct survival strata. The patients with extremely low survival probability are good candidates for the cardioverter defibrillator.

The survival tree models were created with the use of a specialized S-plus library (Dannegger, 1997). The rest of the analysis was done with SPSS.

Statistical Methods

Classification and Regression Trees (CART) for Survival Data

CART is a prognostic system with hierarchical structure, based on recursive partitioning. (Breiman et al, 1984). The idea of CART can be extended and applied to survival data in the following way (Dannegger, 2000). A binary tree is grown from the group of predictors by recursively dividing the observation space into two disjoint subspaces while an optimal predictor and its optimal cut value are chosen at each step. The log-rank statistic is used as a criterion for goodness of split for each subsequent node. When a training sample, containing time, status, and predictor values is fed in the algorithm, the output is a collection of end-nodes which can be described by the parent-branches in terms of predictor values. Each end-node contains the number of total and censored observations falling into the current category, as well as a Kaplan-Maier estimation of the cumulative survival for the group. After an oversized tree is grown, it can be pruned using cross-validation while balancing between fit and complexity of the tree. CART has the advantage of producing a clear and interpretable classification of observations into risk groups by value of the predictors. Each group is characterized by a certain survival function, the estimate of which can be used for prediction of survival for new observations.

The tree diagrams in this paper consist of the following elements: split(parent) nodes, indicating node number, split variable, and split value in a form of a question. All observations answering “yes” to the question are routed to the left. The ones which answer with “no” are routed to the right. The split nodes are circular and contain the p-value of the log-rank statistic at the given cut point. End nodes are rectangular and indicate node number, number of elements in the node (n), and number of events (i.e. number of non-censored observations) in the node.

Data Description

The NIRVPIP data set contains 1353 entries of patients who have survived an acute myocardial infarction (59 patients died during the observation time, the rest were censored). The data set contains a total of 45 characteristics of the following types: physical, background disease and/or medication, infarct history, and heart characteristics. The last group contains a large number of measurements and calculations made on the first 24 hour ECG after the last infarct of each patient. A summary of the most important variables is given in Table 1. The continuous variables were considered in addition as dichotomous, using cut-points from previous investigations of this data set. Observation time and status were recorded for each patient, where status indicated death. Variables Onset and Slope were defined as in Schmidt et al (1999) and dichotomized at 0 and 2.5 respectively. Variable HRT is the sum of the dichotomized Onset and Slope. In other words, a patient with HRT = 0 has good Slope and Onset, a patient with HRT = 2 has bad Onset and bad Slope, and a patient with HRT = 1 has either bad Slope or bad Onset. Variable Creatinkinase had 152 missing values in the original data set, variables Onset and Slope – 396 each, the variable containing the number of extra systolic beats in the longest salvo – six, heart rate frequency – one, and heart rate variability – nine missing values. Multivariate linear and logit regression models were created for the first four variables from the rest of the predictors in order to estimate their missing values. The missing value in heart rate frequency was replaced by the mean and for heart rate variability – by the median value (since the distribution of heart rate variability is skewed).

Table 1: NIRVPIP data set

Variable / type	mean (SD) or number (%)	Variable / type	mean (SD) or number (%)
Physical		Heart	
Age	61.02 (12.07)	Creatinkinase	892 (971)
Sex (male)	1029 (76.1%)	Ejection Fraction	53.82 (12.95)
		Heart rate frequency	64.78 (10.82)
Infarct history			
More than one infarct	196 (14.5%)	Number of Couplets	4.26 (39.83)
Number blood vessels affected	1.96 (.86)	Number of Salvos	.31 (3.96)
		Number of extra systolic beats in the longest salvo	.58 (2.34)
Background disease/medication			
Diabetes	227 (16.8%)	Heart rate variability	28.17 (13.24)
Nicotine	690 (51%)	Onset	-.009 (.024)
Femana	315 (23.3%)	Slope	8.97 (8.23)
β - blocker	1240 (91.6%)	HRT (1)	300 (22.2%)
Nitrate	173 (12.8%)	HRT (2)	130 (9.6%)
Diuretic	557 (41.2%)		

Results

It has already been shown using the Cox model (Ulm et al, 2000) that patients with ejection fraction greater than 50 have good prognosis regardless of the level of HRT. However, different survival curves are estimated for patients with ejection fraction (EF) less than 50 but with different levels of HRT. In support of that, a CART model for survival data was built using the variables EF and HRT. Figure 1 shows the resulting tree (Tree 1), pruned and 10-fold cross-validated. Patients with low ejection fraction (EF < 24.2) have the worse survival rate – 55% cumulative 2-year survival rate (node 2). Patients with high ejection fraction (EF > 45.95) have the best prognosis – 98% cumulative 2-year survival rate (node 5). Patients with medium ejection fraction are further distinguished according to their heart rate turbulence. Patients with bad heart rate turbulence (HRT = 2) have a cumulative 2-year survival rate of 76% (node 7) and patients with better heart rate turbulence ($HRT \neq 2$) have 92% cumulative 2-year survival rate (node 6). The groups of patients in end nodes 5 and 6 have a 2-year cumulative survival of over 92% and hold 92.8% of all patients in the NIRVPIP data set. The remaining 97 patients are in node 2 and 7 (Table 2). Figure 2 shows the Kaplan-Maier estimates of cumulative survival for each end-node.

Table 2: Survival and censoring of Tree 1.

	Cumulative Survival		**Percent**
nodes	***2 years***	***4 years***	**Censoring**
2	.55	.28	57.58
5	.98	.98	98.51
6	.92	.90	93.93
7	.76	.67	76.56

Figure 1: NIRVPIP data set – optimal tree with EF and HRT.

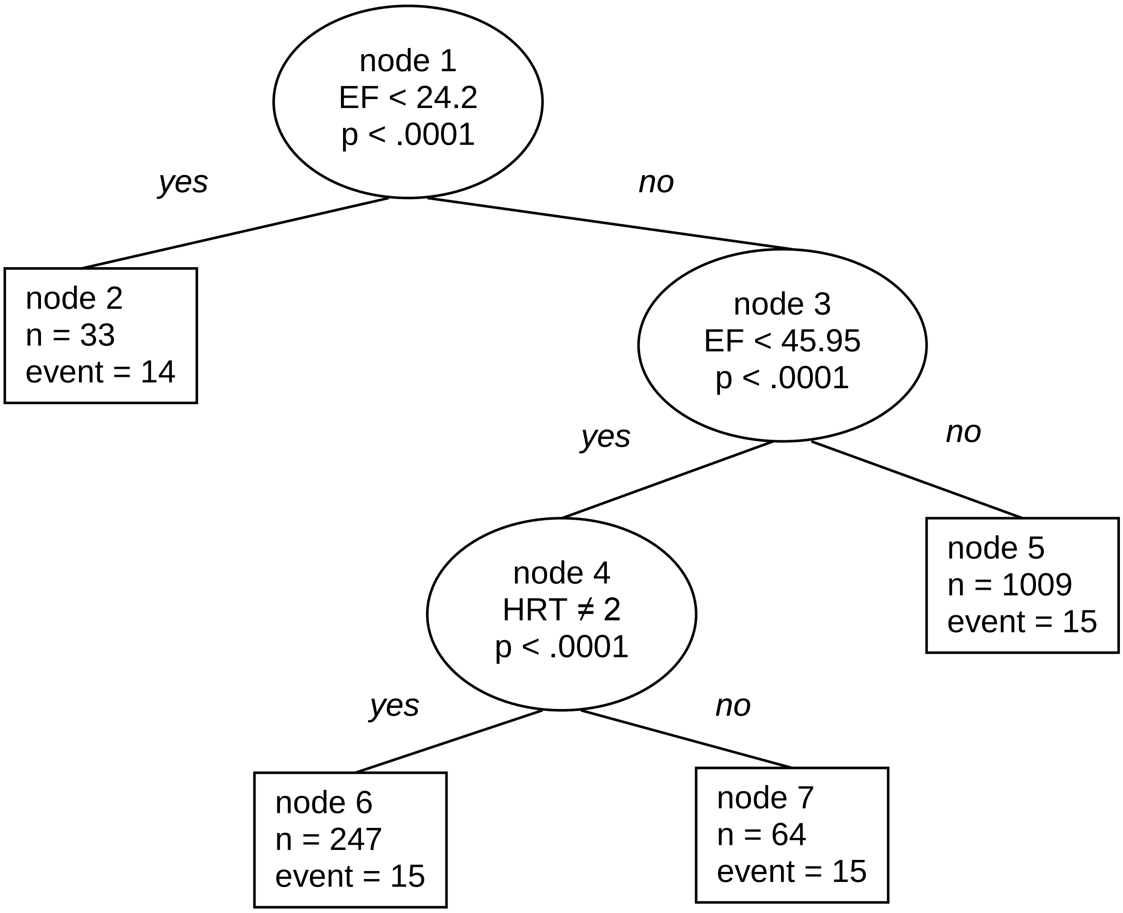
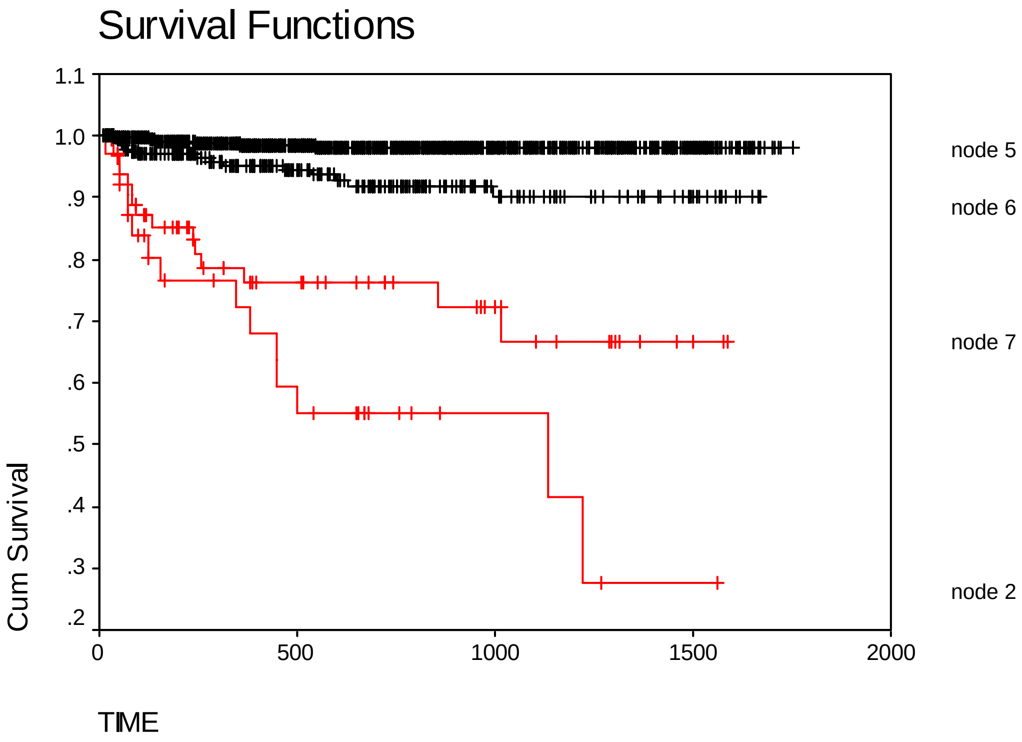


Figure 2: Kaplan-Maier curves for the end-nodes of Tree 1



Another classification is achieved if additional variables are included in the CART model. Using all predictors in the model building process, the optimal, 10-fold cross-validated tree (Tree 2), extracts only the predictors ejection fraction (EF), heart rate turbulence (HRT), presence of couplets (TR.CPL, a binary indicator), and number of extra systolic beats in the longest salvo of the patient's 24-hour ECG (VESS.NMA) as shown in Figure 3. In particular, Tree 2 finds further discrimination criteria which separate the groups of Tree 1 with medium survival rate (nodes 6 and 7) into groups with lower and groups with higher survival rate. Namely, patients without couplets and with good heart rate turbulence ($HRT \neq 2$) in node 8 and patients with repetitive arrhythmia but less than 5 number of extra systolic beats in the longest salvo and good heart rate turbulence ($HRT = 0$) in node 12 have very high survival rates. On the other hand, worse survival rates are predicted for patients without couplets but with bad heart rate turbulence (node 9) and patients with repetitive arrhythmia with at least one salvo of length greater than five (node 13), even though they have good heart rate turbulence ($HRT = 0$). Another category of patients with low survival rates is the one with couplets and bad heart rate turbulence ($HRT \neq 0$) in node 11. With the classification in Tree 2, the patients in end-nodes 5, 8, and 12 have cumulative 2-year survival greater than 97% and are 89.3% of all patients. The remaining 145 patients have survival rates between 55% and 83%. Please refer to Figure 3 for the optimal tree, Figure 4 for the Kaplan-Maier curves of the end nodes, and Table 3 for the 2- and 4-year survival rates and percent censoring in the end-nodes of both trees.

A note on bootstrapping:

Since at node 4 variable TR.CPL had just a slightly better p-value of the log-rank statistic than variable HRT ($1.17 \cdot 10^{-5}$ vs. $2.39 \cdot 10^{-5}$), 10 bootstrap samples were taken at that split and 10 different sub-trees were created. Four out of ten had variable TR.CPL at that split. Since HRT appeared more often, the best tree with HRT at node 4 was also briefly considered. It differed from Tree 2 in only two end nodes, and that did not change the general stratification into low and high risk patients. Therefore, the original split at node 4 was kept.

Figure 3: NIRVPIP data set – optimal tree with EF, HRT, TR.CPL, and VESS.NMA.

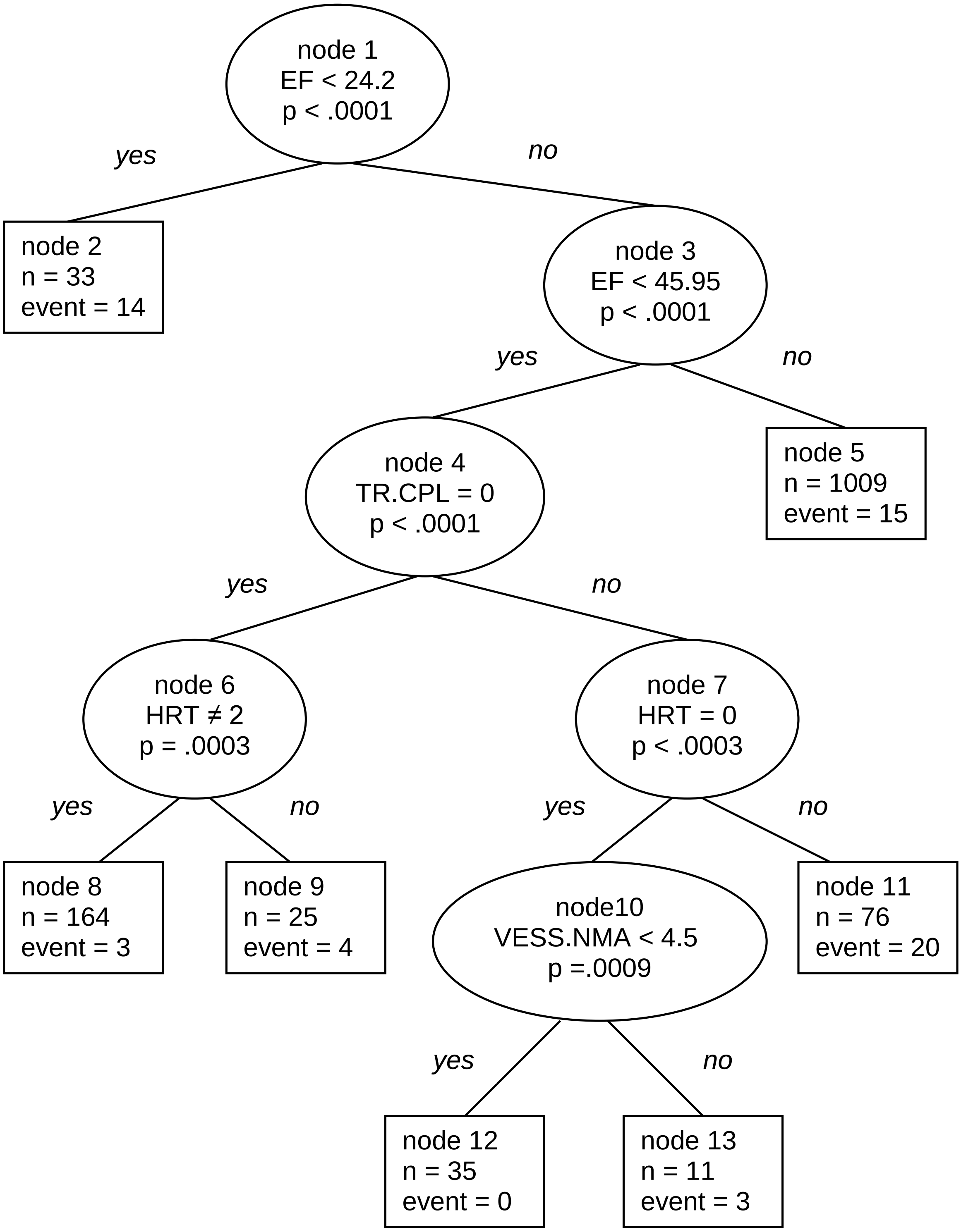


Figure 4: Kaplan-Maier curves for the end-nodes of Tree 2

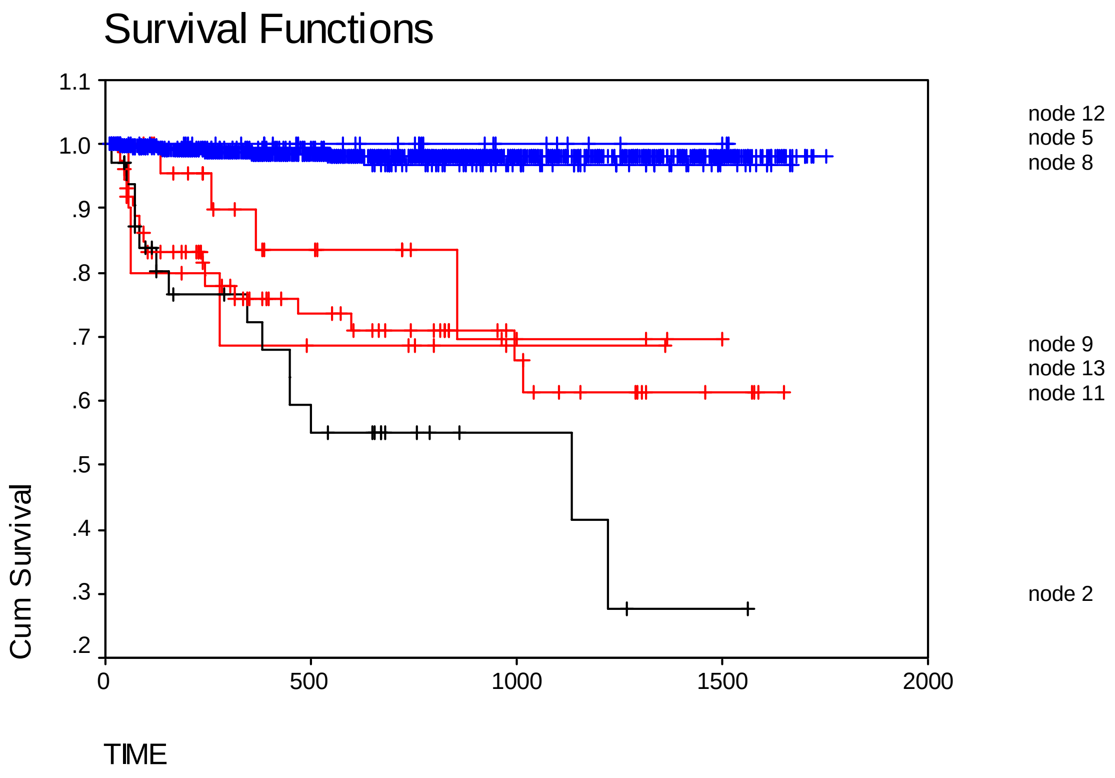


Table 3: Survival and censoring of Tree 2.

	Cumulative Survival		**Percent**
nodes	***2 years***	***4 years***	**Censoring**
2	.55	.28	57.58
5	.98	.98	98.51
8	.97	.97	98.17
9	.83	.70	84.00
11	.71	.61	73.68
12	1.00	1.00	100.00
13	.68	.68	72.73

Two simplified classification survival models (Model 1 and Model 2) were created using the two corresponding trees by grouping the nodes with good and the nodes with worse survival rates as shown in Figure 5 and Figure 6. In Model 1, the higher survival rate group contains 92.8% of all patients and has 97% cum. 2-year survival rate. There are only 97 patients in the lower survival group and their predicted cum. 2-year survival rate is 68%. The groups of patients in Model 2 have similar survival rates, however, more patients are in the lower survival group. In other words, Model 2 extracts some patients from the higher survival group of Model 1 according to additional (repetitive arrhythmia) characteristics. Some model diagnostics are shown in Table 4 for comparison of the two models.

Table 4: Comparison of Model 1 and Model 2.

	Model 1	**Model 2**
Sensitivity	.71	.69
1 – Specificity	.1182	.0804
Log-rank statistic	180.34	261.76
AIC	185.38	249.76
Error rate	.44	.17

Model 2 has a slightly worse sensitivity, but a better specificity, better AIC and error rate than Model 1. Figure 7 shows a plot of the deviance throughout the cross-validation procedure and the residual deviance for trees of different sizes. The deviance of Tree 2 is slightly better than the deviance of Tree 1 as both trees were grown (residual deviance). Although the cross-validation deviance of Tree 2 is only slightly better than that of Tree 1 (193.10 vs. 194.91), it confirms that Tree 2 fits the NIRVPIP data better than Tree 1.

Figure 5: Kaplan-Maier curves for the end-nodes of Model 1

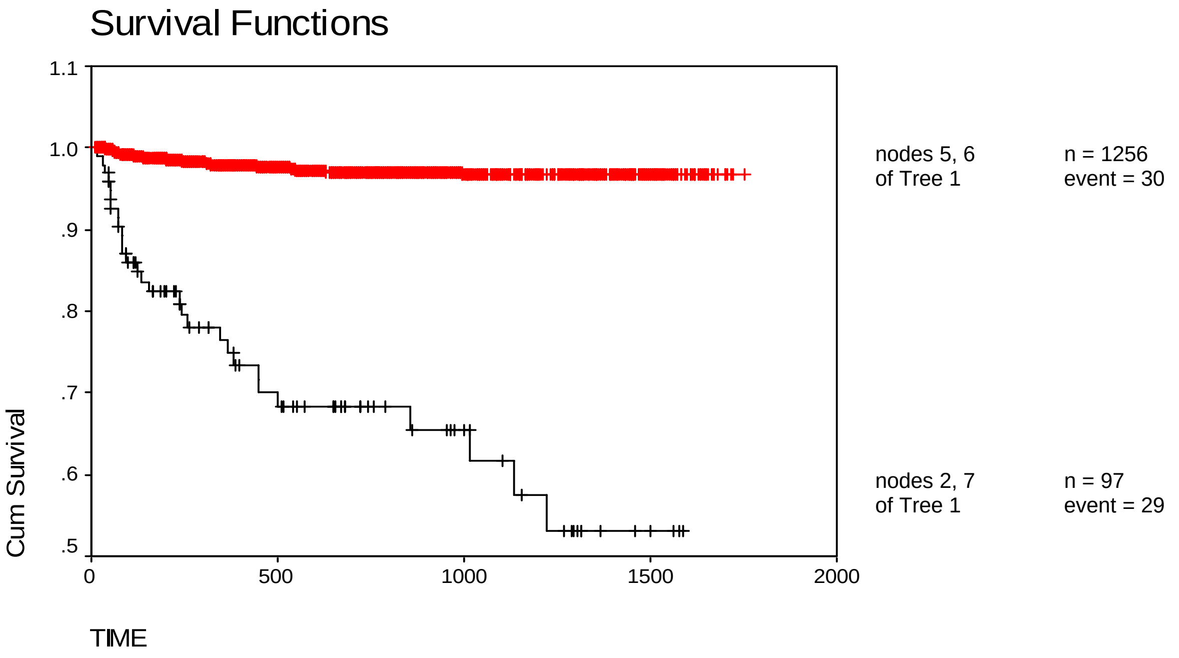


Figure 6: Kaplan-Maier curves for the end-nodes of Model 2

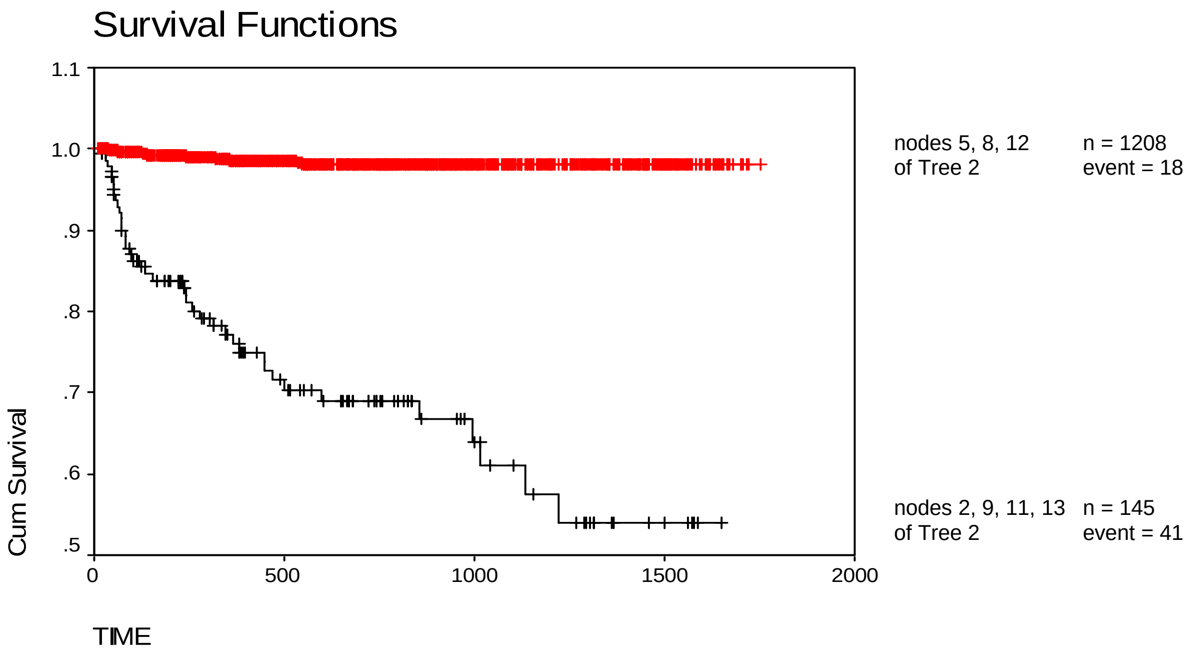
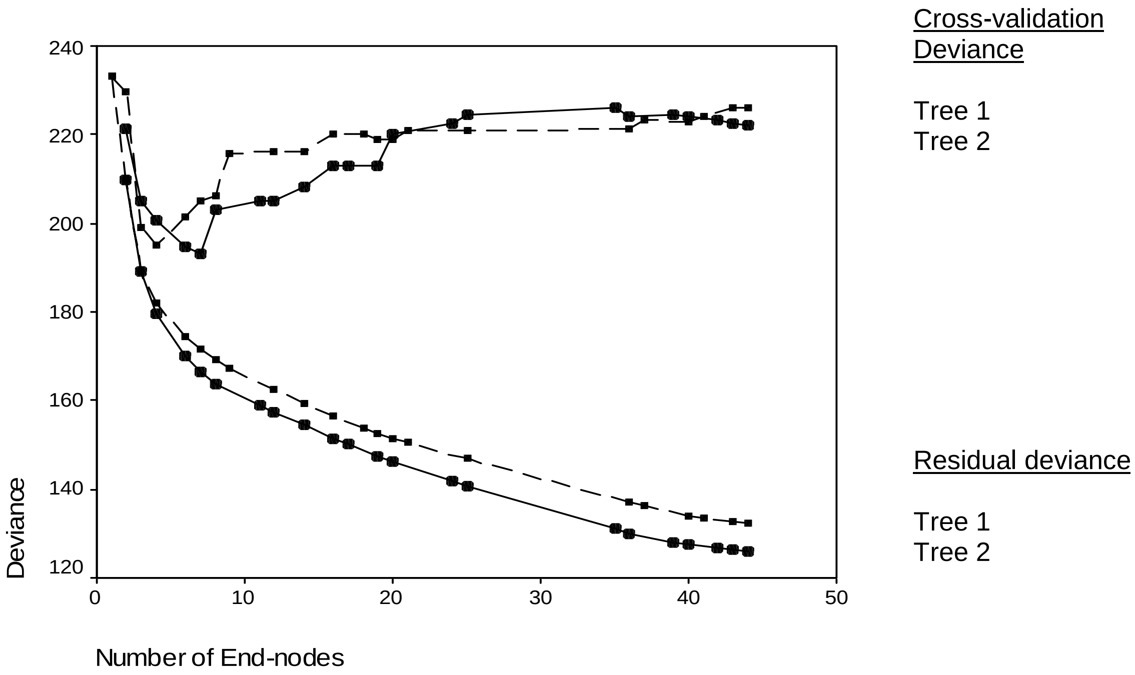


Figure 7: Residual and cross-validation deviance of Tree 1 and Tree 2.



Discussion

The above given information shows that in certain aspects the model with more parameters is better than the model containing only ejection fraction and heart rate turbulence. In particular, more patients at high risk are detected: 41 out of a total of 59 deaths are in the high risk group, which contains 145 out of all 1353 patients. The model shows a better fit through a variety of statistics, including an error rate of 14%.

A model with more variables has, of course, its negative sides – the complexity of the model is higher and it contains more variables which have to be additionally gathered. Note however, that the variables needed can be extracted fast and easily from the 24 hour ECG's of the patients as they are already needed for calculating the heart rate turbulence in the simpler model. The patients need to be characterized by just two additional variables. The above obstacles bring the benefit of increasing the number of candidates for a cardioverter-defibrillator (the low survival rate patients) substantially from 7.2% to 10.7% of all acute myocardial infarction patients.

Given the fact that the censored cases in the NIRVPIP data set are 95.6% of all cases, it is impossible to obtain a model with perfect prediction. The above presented model with four predictors is optimal, given the limitations of censoring.

Bibliography

Breiman, L. et al. Classification and Regression Trees. New York: Chapman & Hall, 1993.

Dannegger, F. “Tree stability diagnostics and some remedies for instability.” Statistics in Medicine 19 (2000): 475-491.

Dannegger, F. Felixtree. Computer software (S-plus library for UNIX). Technical University – Munich, 1997.

Harbeck, N. et al. “Neural network analysis of follow-up data in primary breast cancer.” International Journal of Biological Markers 15(1) (January – March 2000): 116-122.

Harrell, F. E. Jr., Lee, K. L., Mark, D. B. “Multivariable Prognostic Models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.” Statistics in Medicine 15 (1996): 361-387.

Le Blanc, M., Crowley, J. “Relative risk trees for censored survival data.” Biometrics 48 (June 1992): 411-425.

Schmidt, G. et al. “Heart-rate Turbulence after ventricular premature beats as a predictor of mortality after acute myocardial infarction.” The Lancet, 14 April 1999: 1390-1396.

Ulm, K. et al. “A New Statistical Model for Risk Stratification of the Basis of Left Ventricular Ejection Fraction and Heart Rate Turbulence.” Discussion paper. SFB 386 – LMU, München. (in print, 2000)

Schember, M., Stare, J. “Explained variation in survival analysis.” Statistics in Medicine 15 (1996): 1999-2012.