Dannegger, Klinger, Ulm:

Identification of Prognostic Factors with Censored Data

Sonderforschungsbereich 386, Paper 11 (1995)

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

Projektpartner

Identification of Prognostic Factors with Censored Data

FELIX DANNEGGER[1]
Institut für Medizinische Statistik und Epidemiologie
Technische Universität München

ARTUR KLINGER[2]
Institut für Statistik
Universität München

KURT ULM[3]
Institut für Medizinische Statistik und Epidemiologie
Technische Universität München

Summary

A major issue in the analysis of diseases is the identification and assessment of prognostic factors relevant to the development of the illness. Statistical analyses within the proportional hazards framework suffer from a lack flexibility due to stringent model assumptions such as additivity and time–constancy of effects. In this paper we use tree based models and varying coefficient models to allow for detectability of prognostic factors with possibly nonadditive, nonlinear and time–varying impact on disease development. Questions concerning model and smoothing–parameter selection are addressed. An analysis of a dataset of breast cancer patients demonstrates the ability of these methods to reveal additional insight into the disease influencing mechanisms.

Keywords: Breast cancer; Nonlinear effects; Penalized likelihood; Prognostic factors; Recursive partitioning; Smoothing; Survival analysis; Tree based models; Varying coefficients.

[1]email: felix@quetzal.imse.med.tu-muenchen.de

[2]email: artur@stat.uni-muenchen.de

[3]email: kurt@quetzal.imse.med.tu-muenchen.de

1. Introduction

Identification of prognostic factors in medical settings is of high importance, in particular when trying to determine therapy or treatment. Consider oncologic diseases, where identifying patients with a high risk of relapse is necessary not only because of limited resources, but also to avoid exposing the low–risk population to the strains and dangers of an adjuvant therapy like radiation or chemotherapy. Great efforts have been undertaken and are ongoing both within the medical community as well as in biostatistics to identify factors relevant to the development of the illness. Survival analyses using the proportional hazards model proposed by Cox (1972) coupled with Kaplan–Meier survival curves have become the standard procedure to evaluate the impact of certain factors on disease development.

In practical situations however, problems arise using solely this approach. Within the proportional hazard modelling framework there is no natural way to extract subpopulations of different risks. In addition, when common assumptions such as linearity, additivity and time–constancy of effects are violated, determination and assessment of risk factors becomes unreliable. Often it may not be reasonable to assume that the impact of a factor remains constant over time or that factor size influences risk in a strictly linear fashion. Thus alternative approaches are needed to help detect more complicated relationships.

One way to address the problem of linearity and additivity of effects are tree based models, which have become a popular additional method of analysis, due to their ability to naturally and optimally stratify populations into subgroups with distinctively different prognosis, in the process automatically identifying relevant prognostic factors and possible low-order interactions. In addition, the intuitive structure of trees is a powerful tool in communicating results outside the statistical community.

More recently, models generalizing the proportional hazards model, or in discrete time situations, approaches using logistic models have been de-

veloped to allow for detection and modelling of nonlinear and time–varying effects of prognostic factors. These very flexible techniques offer insight into complicated processes influencing the disease at interest.

We wish to demonstrate that these two different modelling approaches can be combined to further the understanding of how prognostic factors influence disease development. In addition to extracting subpopulations with differing risk expectations we use results of a tree based model to feed optimally dichotomized covariates into the varying coefficient framework which in turn is used to analyze disease influencing mechanisms. As with all highly adaptive modelling approaches care has to be taken when it comes to the question of model and smoothing parameter selection. Thus, where possible we introduce data–driven methods to select these parameters.

After reviewing notation in section 2, we describe tree based models briefly in section 3 and varying coefficient models in more detail in section 4. In section 5, we then analyze a population of 315 post–operative breast cancer patients using these methods.

2. Data Notation

Survival data usually consist of failure time measurements and additional covariates which are assumed to influence time to failure. We will assume that covariate values do not depend on time, although most of the methodology discussed here can be adapted to handle time–dependent covariates. An observation is given as the triple (T, δ, X), where T denotes time under observation, δ is the indicator of failure and $X = (X_1, X_2, \ldots, X_p)$ is a vector of p covariates. Assume U to be the true and possibly unknown time to failure and let V denote the true censoring time. Then δ is defined as $\delta := I_{\{U \leq V\}}$ and the observed time is taken to be $T = \min(U, V)$. In the context of varying–coefficient models, we divide the set of covariates into covariates $z_1, \ldots, z_p$ often constructed from basic covariates and a set of metric covariates $x_{p+1}, \ldots, x_q$ called effect–modifiers.

3. Tree based models

Most of the recent developments in tree based models go back to the monograph of Breiman, Friedman, Olshen and Stone (1984). Tree based models rely upon a recursive, binary partitioning of the predictor space $\mathbf{X}$ into disjunct subspaces to either form groups of elements, called nodes, which are homogenous with respect to the response variable of interest, or to form subgroups with maximized between group heterogeneity. This process is repeated for the resulting subgroups, until it is determined that further partitioning is not warranted. Nodes which are not split again are called terminal nodes and form the final subgroups. The result of such an algorithm can be displayed in a binary tree structure.

3.1. Recursive partitioning

The main component of recursive partitioning algorithms regardless of the type of response variable is a set S of split inducing binary questions of the form 'Is $X_i \in A$?' where $i \in 1, \ldots, p$ and $A \subset \mathbf{X}$. Observe that $S = S_1 \cup S_2 \cup \ldots \cup S_p$, where each S_i is the set of binary questions concerning covariate $i = 1, \ldots, p$. For ordered covariates X_i, the set of possible questions reduces to 'Is $X_i \leq c$', with c taking on all values of covariate realizations for elements in the current node. For unordered covariates, all possible divisions of categories into two groups must be examined. Each of these questions induce a candidate split s, sending elements belonging to A to the left sibling node, others to the right.

The other components of recursive partitioning algorithms need to be adapted to the data situation at hand. They are:

- A goodness of split criterion which is evaluated for all candidate splits s to determine the best split of a node. Usually, this criterion will measure the homogeneity of the resulting subgroups of a candidate split

with respect to the response variable, choosing the split which produces the most homogenous sibling nodes. Alternatively, goodness of split criteria have been derived which maximize heterogeneity between subgroups.

- A method to grow right–sized trees. This normally involves enforcing a minimum node size and a minimum improvement in homogeneity. If these requirements can't be met by any of the candidate splits, the node is labelled terminal, and no further partitioning is attempted for that node. More flexible and computationally intensive methods usually referred to as 'pruning' employing cross–validation and similar in spirit to forward–selection backward–deletion methods are commonly used when trees are intended to be used as optimal predictors.

- Methods assigning estimated response values to elements of a terminal node or summary statistics describing the terminal nodes. In the classification setting for example this will be the same estimated class for each element of a terminal node.

As we intend to use recursive partitioning solely for survival data we refer to Breiman et al. (1984) for a more extensive discussion of these components in the classification or regression setting.

3.2. Adaptation to survival data

The construction of the set of candidate splits S remains unchanged in the survival analysis setting. However, to enable recursive partitioning on censored data, the goodness of split criterion, the method to grow right–sized trees and the way elements of a terminal node are characterized need to be adapted to the survival data situation. Extensions of recursive partitioning to the survival analysis setting can be found in Gordon and Olshen (1985), Ciampi, Chang, Hogg and McKinney (1987), Segal (1988), Davis and Anderson (1989), LeBlanc and Crowley (1992) and LeBlanc and Crowley (1993).

Here we use the two–sample log–rank test as a goodness–of–split criterion to maximize heterogeneity between resulting subgroups. An optimal partition is determined for each covariate, while final selection of the best overall split for a node is deferred until adjusted p–values π_{adj} have been calculated for each of the maximally selected test statistics. P–values are also used in a formulation of a stopping rule to avoid oversized trees. We use the conservative approach of declaring a node terminal if the maximized log-rank statistic is not significant at a prespecified significance level π_{max}. Taking into account the fact that we are using maximally selected test statistics, we use the following permutation techniques to derive adjusted p–values for each split (see LeBlanc (1990) for details). Let $LR_{max}(i,t)$ be the maximized log–rank statistic for covariate i at node t. To estimate corresponding p–values for the optimal split of each covariate, we draw m permutation samples of the population of node t, that is we permute the (T,δ) of the individuals with their covariate vectors X. For each of these permuted samples we maximize the log–rank statistics for all covariates $i = 1,\ldots,p$ and thus receive $LR^k_{max}(i,t)$ with $k \in 1,\ldots,m$ and $i = 1,\ldots,p$. For an estimate of the p–values of the original maximized statistics, we use

$$\pi_{adj}(i) = \frac{\sum_{k=1}^{m} \{I_{\{LR^k_{max}(i,t) \geq LR_{max}(i,t)\}}\} + 1}{m+1}. \qquad (1)$$

We then choose covariate j for which

$$\pi_{adj}(j) = \min_{i \in \{1,\ldots,p\}} \{\pi_{adj}(i)\}$$

to split node t using the cutpoint found by maximizing the log–rank test for the original population of node t. If $\pi_{adj}(j) < \pi_{max}$ the split is performed, otherwise t is declared terminal. Note that in order to receive adjusted p–value estimates of adequate resolution m must be chosen sufficiently large. The correction term in (1) assures that the estimate will be conservative and always at least equal to $1/(m+1)$. We intentionally avoid using cross–validation based pruning methods, as this computationally intensive procedure is of little gain when trying to identify covariates with prognostic impact, optimal

cutpoints and low–level interactions. In situations where trees are used as predictors, it will be desirable to combine p–value adjustments with pruning techniques.

Finally, to compare and describe the derived subpopulations we use Kaplan–Meier estimates of cumulative survival. In addition we employ the suggestion of LeBlanc and Crowley (1992) to estimate a proportionality parameter for each terminal node with respect to the overall population to compare risk expectations.

4. Varying–coefficient models

4.1. Logistic models for survival data

Frequently survival data are reported using days or months as time units. In this context we propose a time discrete survival model with possible failure failure times $T_i \in \{1, \ldots, S\}$ and identify the index s with the number of intervals since an individual has been at risk. To express survival data in terms of logistic models we introduce the risk indicator

$$r_i(s) = I\{T_i \geq s\} = \begin{cases} 1 & \text{individual } i \text{ is at risk in interval } s \\ 0 & \text{otherwise.} \end{cases}$$

and the failure indicator

$$y_i(s) = \delta_i I\{T_i = s\} = \begin{cases} 1 & \text{individual } i \text{ is at risk and fails in interval } s \\ 0 & \text{otherwise.} \end{cases}$$

for each time interval $s = 1, \ldots, S$. For an observed event of individual i during interval T_i we have $y_i(T_i) = 1$ and $r_i(T_i) = 1$ and for a censored one $y_i(T_i) = 0$ and $r_i(T_i) = 1$. Suppose that $y_i(s)$ is the outcome of a Bernoulli experiment in each interval s. It is clear that 'not at risk' ($r_i(s) = 0$) implies 'no failure' ($y_i(s) = 0$). Conditional probabilities of failure given the risk indicator and the covariates z_i, $\lambda_i(s) = \mathrm{P}(y_i(s) = 1 | r_i(s), z_i, x_i)$, are linked

to a time–varying predictor, in the following written as $\eta_{is}(z_i, x_i) = \eta_{is}$. Assuming

$$\log \frac{\mathrm{P}(y_i(s) = 1 | r_i(s), z_i, x_i)}{\mathrm{P}(y_i(s) = 0 | r_i(s), z_i, x_i)} = \eta_{is} \qquad \text{for} \quad r_i(s) = 1$$

leads to a logistic model for the time–discrete hazard function

$$\alpha_i(s) = P(T_i = s | T_i \geq s, z_i, x_i)$$

with

$$\begin{aligned} \mathrm{P}(y_i(s) = 1 | r_i(s), z_i, x_i) &= r_i(s) \tfrac{\exp(\eta_{is})}{1+\exp(\eta_{is})} \\ &= r_i(s)\alpha_i(s). \end{aligned} \tag{2}$$

The standard approach for estimating parameter effects in this model is based on likelihood inference. Arjas and Haara (1987) give general conditions in presence of censoring and time–dependent covariates where the full likelihood of model (2) has the form of a likelihood for standard logistic models. It is highly recommended to use grouped data for computation. Let $y_h^*(s)$ be the number of observed events in subpopulation h, characterized by a common covariate vector z_h, x_h, and let $r_h^*(s)$ be the corresponding number of individuals at risk in s. Then the log–likelihood can be written as

$$l(\eta) = \sum_{s=1}^{S} \sum_{h \in R_s} y_h^*(s)\eta_{hs} - r_h^*(s) \log(1 + \exp(\eta_{hs})), \tag{3}$$

where R_s is the set of distinct subpopulations at risk in interval s. Note that this likelihood is also correct in presence of tied observations.

4.2. Varying coefficients

The generalized linear model approach assumes the predictor to be a linear function of the covariates

$$\eta_h = \beta_0 + \sum_{j=1}^{p} \beta_j z_{hj}. \tag{4}$$

Parameter estimates are obtained by maximizing (3) over $\beta_0, \ldots, \beta_p$. Dropping the time constancy assumption in (4) leads to a dynamic generalized

linear model (cf. Fahrmeir and Tutz (1994) ch.8, 9), where coefficients are allowed to vary over time. In the simplest form we have a semiparametric time–discrete survival model

$$\eta_{hs} = \beta_0(s) + \sum_{j=1}^{p} \beta_j z_{hj}, \tag{5}$$

where the baseline effect is assumed to be a smooth function and estimated simultaneously. Extensions to time–varying coefficient models of the form

$$\eta_{hs} = \beta_0(s) + \sum_{j=1}^{p} \beta_j(s) z_{hj}.$$

where the relative risk of failure for a certain subpopulation depends on the basic time scale are straightforward. Since we are not able to assume specific functional forms for continuous covariates $x_1 \dots x_q$, like the concentration of hormones, we drop the linearity assumption in other directions than time, too. To avoid the 'curse of dimensionality' let us assume additivity of the varying coefficients

$$\eta_{hs} = \beta_0(s) + \sum_{j=1}^{p} \beta_j(x_{hj}).$$

Additive models of this structure are discussed in detail in Hastie and Tibshirani (1990). Combining these two extensions leads to the flexible framework of varying–coefficient models, introduced by Hastie and Tibshirani (1993). These models extend the predictor (4) to

$$\eta_{hs} = \beta_0(s) + \sum_{j=1}^{p} z_{hj}\beta_j(s) + \sum_{j=p+1}^{p+q} \beta_j(x_{hj}) z_{hj}, \tag{6}$$

where effects are assumed as a function varying smoothly over the effect–modifiers time and x_j. One may interpret the coefficients in (6) as interactions between covariates and time or between categorical and metrical covariates.

The functions β_j are estimated by maximizing a penalized log–likelihood criterion

$$j(\beta_1, \dots, \beta_{p+q}) = l(\eta) - \sum_{j=1}^{p+q} \lambda_j J(\beta_j), \tag{7}$$

where $J(\beta_j)$ is a roughness penalty, penalizing deviations from smooth functions. For convenience we include time in the set of effect–modifiers and continue to write $\beta_j(x)$ when referring to a time–varying effect.

It is well known (Green and Silverman (1994)) that the maximizer of j using the integrated squared curvature

$$J(\beta_j) = \int (\beta_j(x)'')^2 dx \tag{8}$$

as penalty function is a natural cubic spline. Though this smoother is very popular, it is not adequate in a variable selection procedure, because it assumes the two–parametric familiy of all linear functions as 'smoothest'. Consequently at least two parameters are used to describe an effect varying or not varying. Thus as an alternative we use first order splines with roughness penalty

$$J(\beta_j) = \sum_{s=2}^{T} \frac{(\beta_j(x_s) - \beta_j(x_{s-1}))^2}{x_s - x_{s-1}}, \tag{9}$$

for observation points $x_1 < \ldots < x_S$ to penalize deviations from (time-) constant β_j. This penalty allows us to include semiparametric models (5) automatically in the model choice. Following Wahba (1990), Wahba, Wang, Gu, Klein and Klein (1994) the maximizer of (7) exists and is unique as soon as the common maximum likelihood estimator restricted to $J(\beta_0), \ldots, J(\beta_{p+q}) = 0$ can be determined uniquely. For the two smoothers proposed here $J(\beta_j) = 0$ corresponds to a non–varying effect β_j when a first order penalty (9) is used whereas for the second order penalty (8), $J(\beta_j) = 0$ leads to a coefficient linear in time or x_j. In the context of survival models stronger smoothness restrictions often become appropriate as time proceeds and only a few individuals are left in the riskset. Introducing a monotonous time transformation $g(x)$ for the first order penalty (9) yields

$$J(\beta_j) = \sum_{s=2}^{T} \frac{(\beta_j(x_s) - \beta_j(x_{s-1}))^2}{g(x_s) - g(x_{s-1})}$$

and the amount of smoothing is controlled by the slope of $g(x)$ determing the differences $g(x_s) - g(x_{s-1})$.

Maximizing the penalized likelihood criterion (7) is done iteratively by a Fisher scoring algorithm which can be written as reweighted penalized least squares estimation. The penalized least squares problems are again solved iteratively by a Gauss–Seidel or backfitting algorithm. This procedure reduces the initial problem to penalized weighted least squares problems

$$\hat{\beta}_j = \arg\min_{\beta_j} \sum_{s=1}^{S} \sum_{h \in R_s} w_{hs}(\tilde{y}_h(x_s) - z_{hj}\beta_j(x_s))^2 + J(\beta_j) \qquad (10)$$

for single functions β_j. Let S_j be a linear smoothing operator or hat matrix derived from a penalty (8) or (9), which maps the working observation vector $(\tilde{y}_1(x_1) \ldots \tilde{y}_{R_S}(x_S))'$ to the 'smoothed' estimates $z_{hj}\hat{\beta}_j$ corresponding to (10). Backfitting iterates these operators $S_0, \ldots, S_{p+q}$ on certain working observations up to convergence of the solutions. The algorithm is described in detail in Hastie and Tibshirani (1993) or in Fahrmeir and Klinger (1995) in the context of event history analysis.

Other survival models which are connected to generalized linear models like the grouped Cox model or the piecewise exponential model can be handeled within the same framework. See Klinger (1993) or Fahrmeir and Klinger (1995) for details.

4.3. Smoothing parameters and variable selection

Generally both choice of smoothing parameters and variable selection may be carried out by optimizing one global criterion estimating the prediction error. For varying–coefficient models criteria like generalized cross–validation (GCV) or Akaike's Information Criterion (AIC) require the trace of the hat matrix in the last iteration step, see Hastie and Tibshirani (1990) and Wahba et al. (1994). Due to the high dimension of the involved matrix inversions computation is still very time–consuming. For a global optimization this quantity has to be computed frequently. To overcome computational burdens we use fast algorithms based on simple heuristics.

Smoothing parameters

Hastie and Tibshirani (1990) propose to use the traces of the smoothing matrices S_j in the final iteration step as 'effective number of parameters' or 'degrees of freedom' of the smoother and select the smoothing parameters $\lambda_0, \ldots, \lambda_{p+q}$ according to a given 'number of parameters'. Using our penalties, λ_j tunes the degrees of freedom from 1, respectively 2 up to the number of distinct time intervals or datapoints of x_j. In order to obtain a procedure for variable selection, one needs a fast automatic algorithm to choose smoothing parameters. Our proposal is an iterative algorithm based on AIC,

$$\text{AIC} = -2l(\eta) + 2\nu, \tag{11}$$

where ν are the degrees of freedom for the model.

The proposed algorithm mimics a statistician who tunes the effective number of parameters for each coefficient β_j seperately. An 'optimal' smoothing parameter is found by trading off the goodness of fit measured by the negative log–likelihood with the degrees of freedom. To estimate only β_j, consider $\beta_0, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_{p+q}$ as 'known coefficients' and let η^{-j} be the predictor composed by those 'known coefficients', then the algorithm switches between

1 Optimization of λ_j:

$$\lambda_j = \arg\min\{-2l(\eta^{-j} + \hat{\beta}_j(x_j, \lambda_j))z_{hj} + 2\text{tr}(S_j(\lambda_j))\}, \tag{12}$$

with $\text{tr}(S_j(\lambda_j))$ denoting the trace of the smoothing operator used in the final Fisher scoring iteration, where only $\hat{\beta}_j(, x_j\lambda_j)$ is estimated and η^{-j} is assumed to be known;

and

2 Updating of the coefficients by estimating all parameters in the entire model simultaneously using the smoothing parameter λ_j computed in step 1.

These two steps are repeated for $j = 0, \ldots, p+q, 0, \ldots, p+q, 0, \ldots$ until the traces of the smoothing operators do not change any more. Optimization in step (1) is carried out by a golden–section search algorithm applied to the Fisher–scoring procedure. By initializing step (2) with the estimate of step (1) the algorithm reaches it's target soon. In our experience this procedure gives stable results for smoothing parameters in connection with AIC. A detailed description of this 'switching' algorithm and simulation studies are given in Klinger (1993).

Alternatively, criteria derived from the idea of cross–validation, such as the generalized cross–validated deviance described in Hastie and Tibshirani (1990) can be used to estimate smoothing parameters. However in the context of survival models the 'leaving–one–out' heuristic for this procedure would be censoring of an individual or a subpopulation at only one time interval. This is not reasonable because the information that the individual is at risk just before and after the censoring is still in the data. Indeed we observed that the more the data are grouped, the smoother the estimates become. Within a parameter selection procedure this involves different smoothing parameters depending on the ability to group the actual model. AIC is not sensitive to grouping of data and uses only likelihood criteria which are also plausible in context of survival data.

Variable selection

Selection between different models is done by AIC, too. Computing the trace of the hat–matrix in the last Fisher scoring iteration to obtain degrees of freedom for the entire model is too time consuming. Instead we proceed as before and use the traces of the smoothing operators $\mathrm{tr}(S_j(\lambda_j))$ to approximate the effective number of parameters by the sum of these traces. Our selection criterion is

$$\mathrm{AIC}_M = -2l(\eta) + 2\left(\sum_{j=0}^{p+q} \mathrm{tr}(S_j(\lambda_j))\right), \tag{13}$$

where we choose the model with minimal AIC_M. One may extend (13) to

$$\mathrm{AIC}_{MX} = -2l(\eta) + \gamma \left(\sum_{j=0}^{p+q} \mathrm{tr}(S_j(\lambda_j)) + \rho_j \right)$$

where for example γ is chosen such, that the selection criterion corresponds to BIC. This extension allows to trade off between complexity in the coefficients (smoothness) and complexity of the model (number of covariates included). The term ρ_j is used, if the corresponding covariate z_j results from an optimized split of the tree based model described previously. For such a split we set $\rho_j = 1$ to incorporate the additional degrees of freedom due to the estimation of the cut–off point.

The selection first proceeds stepwise in a forward manner. One starts with only the baseline included and estimates a supermodel including one covariate more and so on. AIC_M is computed from the estimation result for each candidate covariate. Since the type of the smoother influences the result, this is done for the first order smoothing spline using at least one degree of freedom and for the cubic smoothing spline using two or more degrees of freedom seperately. Interactions are only considered between included covariates. The forward selection stops, when no supermodel reaches a lower AIC_{MX}. Now all possible models excluding one coefficient are computed, if one of these submodels has a lower AIC_{MX} we start a stepwise backward deletion. The backward deletion is repeated until no submodel can be preferred in the sense of the AIC_{MX} criterion.

5. Breast cancer study

In 1987, a prospective study on $n = 315$ post-operative breast cancer patients was initiated at the department of obstetrics and gynecology of the Technische Universität München in order to reveal and assess prognostic factors related with relapse. During the course of the study 102 patients experienced a relapse. Follow-up time ranged from 1 to 88 months with a

median follow–up of 47 months. In order to identify subpopulations with different risk expectations, new factors such as the urokinase–type plasminogen activator (uPA) and it's inhibitor (PAI-1) in the following referred to as PAI were investigated in addition to classical factors such as age of patient, number of removed positive lymph–nodes, tumor size and hormone receptor status. A dichotomized version of the number of removed positive lymph nodes — 'lymph node status' was included, indicating absence or presence of any positive lymph nodes. A complete listing of covariates included in the analysis can be found in table 1.

5.1. Tree based model

A survival tree was grown on the data using all covariates available to assess impact on the response 'time to relapse'. Parameters of the algorithm were such that a node was declared terminal if no candidate split resulted in an adjusted log–rank test statistic significant at the $\pi_{max} = .05$ level or if one of the resulting sibling nodes contained less than 10 individuals. P–value adjustment was performed using the permutation technique described with $m = 2000$ permutations for each split.

Figure 1 shows the resulting tree with 8 splits and 9 terminal nodes. The node number, covariate used to split the node, corresponding cutpoint and adjusted p–value are recorded beneath each split. For every terminal node, the node number, the number of individuals and events and an estimate of relative risk are recorded. Chosen covariates and cutpoints for each node are also recorded in table 2. For binary covariates, no cutpoints are given, instead factor level 0 individuals are sent to the left, level 1 individuals to the right.

The partitioning process begins by splitting up the entire population according to whether the number of removed positive lymph nodes is less than or greater than 6.5. This result confirms the well known fact that lymph node status is the factor with the highest prognostic impact on relapse. Before

Covariate	Description	Range
AGE	Age of patient at surgery in years	$27.3 - 88.6$
LYPO	Number of removed, positive lymph nodes	$0 - 40$
TUMOR	Tumor size in cm	$0.5 - 15$
DHORM	Hormone receptor status	$0 =$ positive $1 =$ negative
DPR	Progesteron receptor status	$0 =$ positive $1 =$ negative
DER	Estrogen receptor status	$0 =$ positive $1 =$ negative
UPA	urokinase–type plasminogen activator (ng/mg protein)	$0.04 - 15.17$
PAI	plasminogen activator inhibitor (ng/mg protein)	$0.06 - 248.8$
MENOP	menopausal status	$1 =$ premenopausal $2 =$ postmenopausal $3 =$ perimenopausal
DLYP	Lymph node status	$0 =$ node negative $1 =$ node positive

Table 1: Covariates in breast cancer study.

Node	Covariate	Cutpoint
1	LYPO	6.5
2	PAI	27.5
3	DPR	binary
4	PAI	14.8
6	LYPO	11.5
8	DLYP	binary
14	PAI	8.2
17	PAI	10.4

Table 2: Split data.

p–value adjustment, PAI seems to be the factor with the second highest impact followed by the binary covariate lymph node status DLYP, tumor size and progesteron receptor status. The situation changes somewhat when the adjusted p–values are used to rank factor importance. Now both lymph node status covariates are ahead of progesteron and estrogen receptor status, while PAI und tumor size drop several ranks. Details can be found in table 3. Observe that it is not possible to decide between LYPO and DLYP on the basis of the adjusted p–values alone, as the number of permutations chosen was not large enough. In such cases one has the option of increasing the number of permutation samples or reverting back to the original p–values. Here, due to the great difference in original p–values, LYPO with it's cutpoint of 6.5 is finally selected as the best root level split. Figure 2 shows the resulting Kaplan–Meier survival curves for nodes 2 and 3 in contrast to the entire population.

Optimized cutpoints at node 1 were used to construct dichotomized versions of covariates to be used in conjunction with the varying coefficient modelling framework of section 5.2..

Continuing downward from node 3 the algorithm seperates a group of

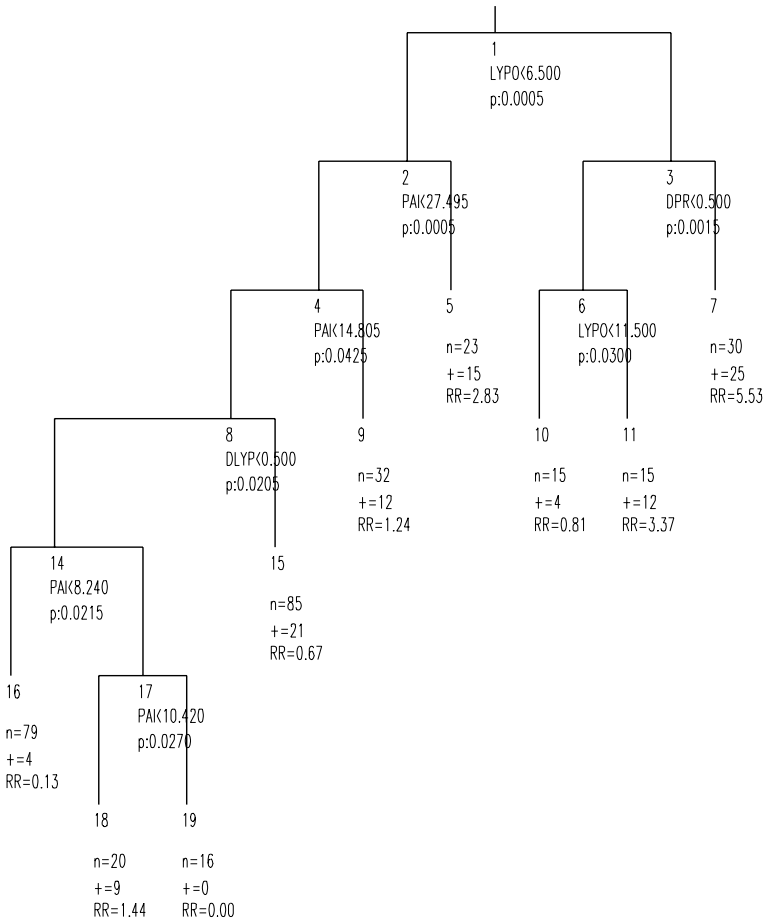

Figure 1: Graph of tree based model

		Goodness of split			
		before adjustment		after adjustment	
Covariate	cutpoint	rank	p–value	rank	p–value
LYPO	6.5	1	$< 10^{-16}$	1	< 0.0005
PAI	27.5	2	$< 10^{-9}$	5	0.0005
DLYP	binary	3	$< 10^{-7}$	2	< 0.0005
TUMOR	8.25	4	$< 10^{-4}$	6	0.0010
DPR	binary	5	$< 10^{-4}$	3	< 0.0005
UPA	4.4	6	0.00014	8	0.0120
DER	binary	7	0.00023	4	< 0.0005
AGE	62.2	8	0.00104	9	0.04450
DHORM	binary	9	0.00355	7	0.0035
MENOP	categorical	10	0.23293	10	0.2870

Table 3: Optimized candidate splits for node 1.

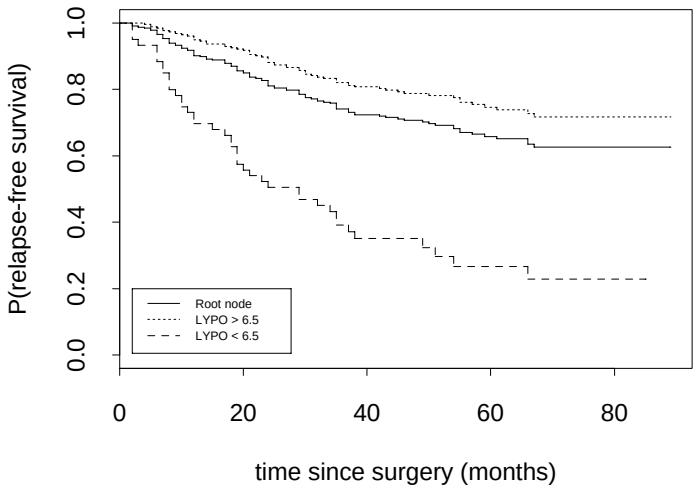

Figure 2: Partition resulting from root level split.

high relative risk (RR) patients with negative progesteron receptor status (RR=5.5) while a further split on LYPO at node 6 reveals a small group of patients with marginally lower risk (RR=0.83) as compared to the whole population. On the branch of the tree originating at node 2, where a large group of lower risk patients was produced by the first split on LYPO, PAI apparently plays an important role in further determining prognosis. Two small subpopulations with elevated PAI concentrations and accordingly increased risks are seperated before remaining patients are split based on their lymph node status at node 8. For a large group of node negative patients with PAI concentrations less than 8.2, the tree predicts a very low risk of relapse (RR=0.13). The split at node 17 is difficult to explain since it appears to show a break in the monotony of the relationship between PAI concentration and risk expectation. This becomes less of a problem when considering the small size of the originating node. Still, the impact of PAI on the risk of relapse appears to be a smooth one and the tree is not able to handle it well by repeatedly splitting off very small subpopulations.

5.2. Varying–coefficient model

A common way to interpret the hazard function in this application is not linear in time. More naturally — taking into account decreased prediction accuracy as time progresses — one looks at hazards per month in the first year, per quarter in the second and third year and per year later on. In this context, we use a transformed time grid

$$\{1,\ldots,12,13,\ldots,36,37,\ldots,88\} \mapsto \{1,\ldots,12,12.33,\ldots,20,20.08,\ldots,24.33\}.$$

This grid also helps overcome boundary problems on the right–hand side of the time axis where the riskset is small. Here the transformation causes stronger smoothness restrictions. Estimation results using the transformed time–scale are compared to estimates based on original time–scale by the AIC_M criterion.

We start by including dichotomized covariates obtained from the canditate splits of the root node in addition to the original covariates in the selection process. Table 4 shows results of the first selection step where the component (LYPO<7) is chosen constant over time. To take into account that the cut–off point for this covariate was estimated by recursive partitioning we add one degree of freedom. The resulting AIC_{MX}, 1167.4, is still the lowest and thus (LYPO<7) is selected as first covariate.

The variable selection is continued in table 5. Some coefficients, those with $\mathrm{tr}(S_j)$=1, are included in a (semi–) parametric manner. The selection stopped after 5 steps. Again, since the covariate (UPA>4.4) results from an estimated cut–off point, we increase the AIC_M by 2. Neither UPA nor any other covariate or interaction decreased the selection criterion computed from the 5 coefficient model. The model choice proceeds with the backward deletion starting with the model,

$$\begin{aligned}\eta &= \beta_0(s) + \beta_1(s)(\mathrm{LYPO}\in\{1,\ldots,6\}) + \beta_2(\mathrm{PAI}) \\ &\quad +\beta_3(s)(\mathrm{AGE}< 62.25) + \beta_4(s)\mathrm{DPR} + \beta_5(s)(\mathrm{LYPO}= 0).\end{aligned}$$

Variable	loglikelihood	AIC_M	$tr(S_0)$	$tr(S_1)$	smoother
β_1(AGE)	-599.9	1211.2	3.2282	2.4780	cubic
β_1(PAI)	-585.8	1190.1	3.1292	4.7165	cubic
β_1(UPA)	-597.6	1208.3	3.2731	3.3341	first order
$\beta_1(s)$(PAI> 27.5)	-589.9	1188.2	3.1838	1.0000	first order
$\beta_1(s)$(UPA> 4.4)	-596.9	1202.4	3.2723	1.0000	first order
$\beta_1(s)$DPR	-590.5	1192.0	3.4613	2.0000	cubic
$\beta_1(s)$DER	-592.8	1196.6	3.4609	2.0000	cubic
$\beta_1(s)$DHORM	-593.5	1198.1	3.5859	2.0000	cubic
$\beta_1(s)$(LYPO= 0)	-587.5	1187.1	2.9819	3.0740	cubic
$\underline{\beta_1(s)(\text{LYPO}< 7)}$	-578.5	1165.4	3.2245	1.0000	first order
$\beta_1(s)$(AGE< 62.25)	-597.4	1203.2	3.2267	1.0000	first order

Table 4: First step of variable selection using the optimal first order or cubic spline.

As can be seen from table 6 none of the submodels indicate better prediction through their AIC_M, so the model selection process terminates.

Furthermore we checked, whether the used transformation of the time axis improved the model. Results using smoothers gained by the 'optimal' choice from table 4 are given in table 7 in the column 'optimal, yes'. The next column lists estimates using only first order splines whereas the last two columns indicate that the fit becomes worse without the time transformation. Except β_4 all estimates recognized the same effects as time–constant and distributed the degrees of freedom similarly to the included variables. The time–constant effects ($tr(S_j)$=1) all have approximately the same value in each of the four estimates. It seems, that the type of the smoother does not have a big influence on the 'parametric' part of the model.

Figure 3 shows the coefficients for the final model using the transformed time axis and 'optimal' smoothers. If one is interested in examining how one

Variable	smoother	loglikelihood	AIC_M	$\text{tr}(S_j)$
$\beta_1(s)$(LYPO<7)	first order	-578.5	1165.4	1.0000
β_2(PAI)	cubic	-561.4	1140.3	4.7512
$\beta_3(s)$(AGE<62.25)	first order	-553.1	1125.8	1.0000
$\beta_4(s)$DPR	cubic	-546.2	1116.9	2.0000
$\underline{\beta_5(s)(\text{LYPO}=0)}$	cubic	-541.8	1113.0	3.0464
$\beta_6(s)$(UPA>4.425)	first order	-540.9	1112.8	1.0000

Table 5: Stepwise forward variable selection, chosen covariates and inclusion criteria.

Variable	loglikelihood	AIC_M
$\beta_1(s)(\text{LYPO}\in\{1,\ldots,6\})$	-556.9	1141.2
β_2(PAI)	-555.3	1131.0
$\beta_3(s)$(AGE<62.25)	-550.2	1128.1
$\beta_4(s)$DPR	-544.6	1122.0
$\beta_5(s)$(LYPO$=0$)	-546.2	1116.9

Table 6: First step of stepwise backward deletion, exclusion criteria of deleted coefficients.

Smoothers	optimal		first order		optimal		first order	
Transform.	yes				no			
Variable	$\text{tr}(S_j)$	$\hat{\beta}_j$	$\text{tr}(S_j)$	$\hat{\beta}_j$	$\text{tr}(S_j)$	$\hat{\beta}_j$	$\text{tr}(S_j)$	$\hat{\beta}_j$
$\beta_0(s)$	2.979	-4.933	1.000	-5.003	2.976	-4.959	1.000	-4.941
$\beta_1(s)$	1.000	-1.357	1.000	-1.388	1.000	-1.363	1.000	-1.388
β_2(PAI)	4.704	0.527	6.544	0.537	4.651	0.542	6.615	0.476
$\beta_3(s)$	1.000	0.926	1.000	0.912	1.000	0.912	1.000	0.911
$\beta_4(s)$	2.000	0.338	1.000	0.602	2.000	0.245	1.000	0.602
$\beta_5(s)$	2.967	-2.216	3.705	-2.146	3.037	-2.395	3.836	-2.227
$l(\hat{\eta})$	-541.8		-543.8		-544.1		-544.0	
AIC_M	1113.0		1116.1		1117.5		1116.9	

Table 7: Effective number of parameters and mean of estimated coefficients by using different smoothing operators.

covariate affects the risk when other factors are fixed, log odds ratios may be calculated as linear combinations of coefficients. Within the final model the covariate LYPO is divided into three categories: node negative patients, patients with fewer than 7 positive lymph nodes and patients having 7 or more positive lymph nodes. Patients with fewer than 7 positive lymph nodes have a distinctly lower time–constant relative risk. For node negative patients the risk of experiencing a relapse is still lower, although this phenomena is clearly time–dependent. Having no positive lymph nodes has an obvious risk decreasing effect for the first two years, after which it does not seem to matter much, whether a patient is node negative or has just a few (less than 7) positive lymph nodes (see figure 3 (b) for details). One interpretation could be that for node–positive patients with less than 7 positive lymph nodes the beginning benificial effects of an adjuvant therapy have the effect of slowly letting risk expectations for these two groups converge. The third entry into the model, AGE was chosen as time constant with patients younger than

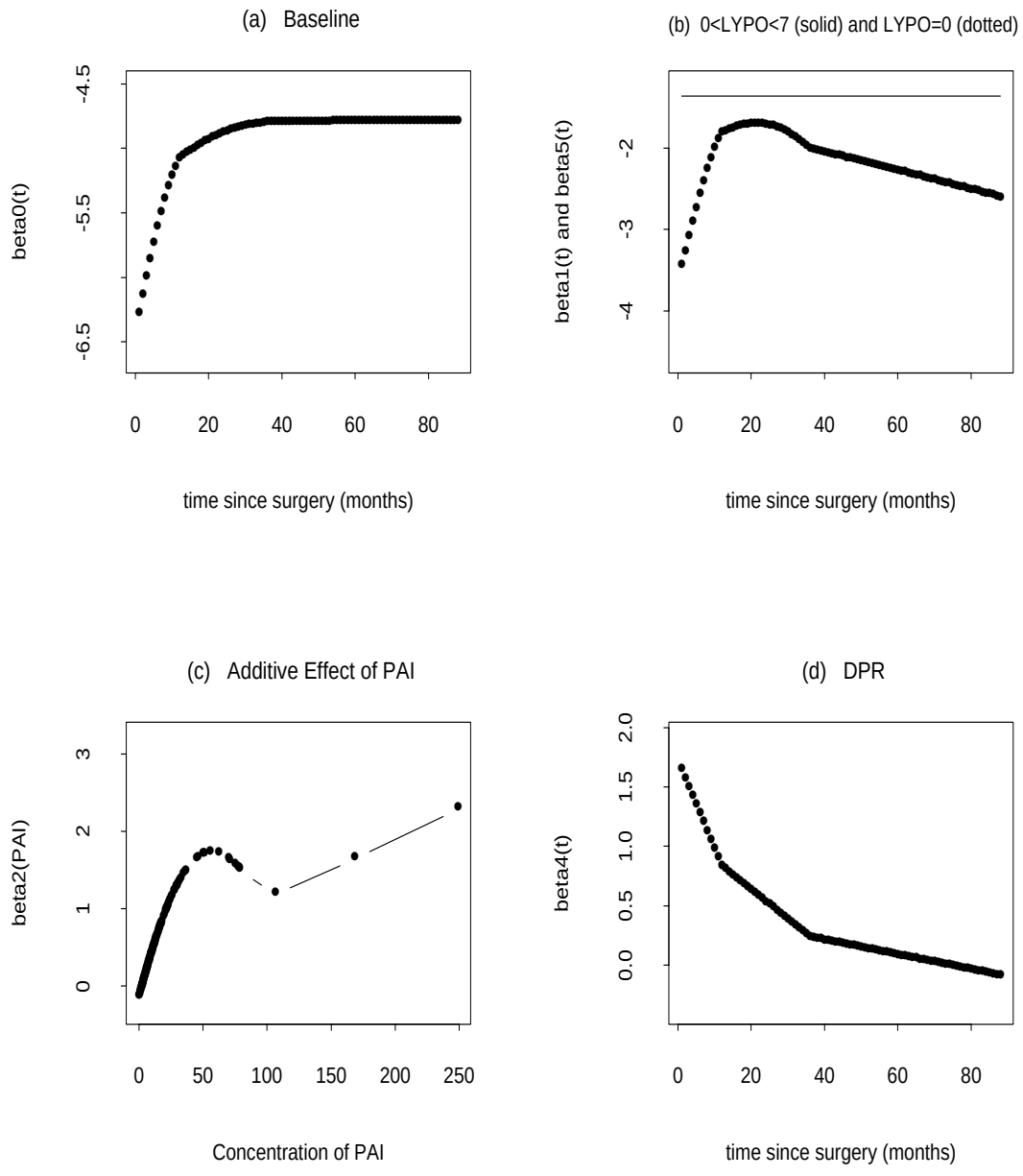

Figure 3: Varying coefficients of the final logistic model. The constant coefficient for AGE $\beta_3 = 0.926$ is not displayed.

62.25 years having a relative risk of 2.5 compared to older patients.

As suspected by looking at the tree model the varying coefficient model confirms that the impact of PAI-1 on the risk of relapse is complicated, which is reflected in it's entry into the model as a nonlinear term. The risk of relapse increases sharply up to PAI-1 concentrations of about 50 ng/mg protein remaining constant at a high level thereafter.

Progesteron receptor status enters the model as a time–varying effect, signifying a relative risk of about 5 immediately after surgery for patients with a negative receptor status. This effect continuously declines until it disappears about 3-4 years after surgery.

6. Conclusions

Currently, there is a discussion going on in the medical community about the impact of these analyses on clinical and treatment decisions. Studying variation over time of the risk associated with these and other factors may give important insights into their role in tumor cell biology.

Our findings may still be well short of changing clinical practice at the moment. Lymph node status together with the number of positive nodes have to be considered first in evaluating the risk of getting a relapse, but in addition, the absence of steroid hormone receptors and high PAI-1 tumor levels can be said to be indicators of early disease recurrence. Accordingly, patients fitting this profile could be enrolled in a tight follow–up schedule during the first years after primary treatment. Later on, for hormone receptor–negative patients remaining disease–free during this early period, a less frequent follow–up might be possible as recurrences tend to be rare. Detailed knowledge of time–dependent and non–linear risk profiles of prognostic factors will eventually enable clinicians to better predict disease recurrence and survival and to individualize follow–up and therapy.

As illustrated, the two methods proposed offer flexible extensions to the more conventional survival analysis framework. Tree based models are well

equipped to detect interactions and their results can immediately be used to stratify patients into different risk groups. On the other hand, allowing for time–varying effects and nonlinear associations enables precise and accurate explanations of influencing mechanisms using models of simple structure. This attempt to combine the advantages of the two methods can be seen as first step towards a more refined assessment of prognostic factors. Further work and practical experience is needed to solve problems of identifiability and stability when trying to combine these methods into a more tightly woven framework.

References

Arjas, E. and Haara, P. (1987). A logistic regression model for hazard. Asymptotic results., *Scand. J. Statist.* **14**, 1–18.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Chapman and Hall, New York.

Ciampi, A., Chang, C. H., Hogg, S. and McKinney, S. (1987). Recursive Partition: A versatile method for exploratory data analysis, *Festschrift in Honor of Professor V.M. Joshis's 70th Birthday*, Vol. V: Biostatistics, D. Reidel Publishing Company, Dordrecht.

Cox, D. R. (1972). Regression Models and Life Tables, *J.R. Statist. Soc. B* **34**, 187–220.

Davis, R. B. and Anderson, J. R. (1989). Exponential Survival Trees, *Statistics in Medicine* **8**, 947–961.

Fahrmeir, L. and Klinger, A. (1995). A nonparametric multiplicative hazard model for event history analysis, *Discussion paper 12*, SFB 386, Munich.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models*, Springer, New York.

Gordon, L. and Olshen, R. A. (1985). Tree-structured survival analysis, *Cancer Treatment Reports* **69**, 1065–1068.

Green, P. and Silverman, B. (1994). *Nonparametric regression and generalized linear models*, Chapman and Hall, London.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*, Chapman and Hall, London.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient Models, *J.R. Statist. Soc. B* **55**, 757–796.

KLINGER, A. (1993). *Spline–Glättung in zeitdiskreten Verweildauermodellen*, Diploma thesis, Ludwig Maximilians Universität München, Institut für Statistik.

LEBLANC, M. (1990). *Recursive partitioning for censored survival data*, Dissertation, University of Washington, Department of Statistics.

LEBLANC, M. AND CROWLEY, J. (1992). Relative Risk Trees for Censored Data, *BIOMETRICS* **48**, 411–425.

LEBLANC, M. AND CROWLEY, J. (1993). Survival Trees by Goodness of Split, *JASA* **88**, 457–467.

SEGAL, M. R. (1988). Regression Trees for Censored Data, *BIOMETRICS* **44**, 35–47.

WAHBA, G. (1990). Spline Models for Observational Data, *CBMS-NSF Regional Conference Series in Applied Mathematics*, Vol. 59, SIAM, Philadelphia.

WAHBA, G., WANG, Y., GU, C., KLEIN, R. AND KLEIN, B. (1994). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy, *Technical Report 940*, Department of Statistics, University of Wisconsin, Madison. To appear, Ann. Statist, 1996.