J. R. Statist. Soc. A (1999)
162, *Part* 1, *pp.* 71–94

Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials

W. Sauerbrei

University of Freiburg, Germany

and P. Royston

Imperial College School of Medicine, London, UK

[Received June 1997. Final revision April 1998]

Summary. To be useful to clinicians, prognostic and diagnostic indices must be derived from accurate models developed by using appropriate data sets. We show that fractional polynomials, which extend ordinary polynomials by including non-positive and fractional powers, may be used as the basis of such models. We describe how to fit fractional polynomials in several continuous covariates simultaneously, and we propose ways of ensuring that the resulting models are parsimonious and consistent with basic medical knowledge. The methods are applied to two breast cancer data sets, one from a prognostic factors study in patients with positive lymph nodes and the other from a study to diagnose malignant or benign tumours by using colour Doppler blood flow mapping. We investigate the problems of biased parameter estimates in the final model and overfitting using cross-validation calibration to estimate shrinkage factors. We adopt bootstrap resampling to assess model stability. We compare our new approach with conventional modelling methods which apply stepwise variables selection to categorized covariates. We conclude that fractional polynomial methodology can be very successful in generating simple and appropriate models.

Keywords: Breast cancer; Model stability; Nominal *P*-value; Resampling; Variable selection; Variable transformation

1. Introduction

To be useful in clinical practice, prognostic and diagnostic indices must be derived from accurate models developed by using appropriate data sets. Most such indices are symptom and/or measurement scores and contain information from several predictors. An example is the well-known Nottingham prognostic index for survival from primary operable breast cancer (Galea *et al*., 1992), which is based on tumour size, lymph node status and histological tumour grade. General statistical aspects of prognostic factor studies have been considered by Simon and Altman (1994) and Harrell *et al.* (1996). An important component of the construction of a prognostic index is the model building stage. Clinical investigators almost always record data on many variables, aiming to investigate all of them simultaneously and to identify those with prognostic ability. Difficulties associated with 'standard' procedures such as stepwise selection are overfitting and underfitting, biased estimates of the regression

Address for correspondence: W. Sauerbrei, Institute for Medical Biometry and Informatics, University of Freiburg, Stefan-Meier-Strasse 26, D-79104 Freiburg, Germany.
E-mail: wfs@imbi.uni-freiburg.de

0964–1998/99/162071

parameters of the final model and a lack of reproducibility of the regression parameters in new data. A review of these methods and of the related mathematical and statistical problems is given by Miller (1990). Although subject-matter knowledge should be used to guide selection, some variables will inevitably be chosen mainly by statistical principles — essentially, P-values for including or excluding variables. The definition of a 'best' strategy to produce a model which has good predictive properties in new data is difficult. A model which fits the current data set well may be too data driven to give adequate predictive accuracy in other settings.

A second obstacle to model building, consideration of which is the main topic of the present paper, is how to deal with non-linearity in the relationship between the outcome variable and a continuous or ordered predictor. Traditionally, such predictors are entered into stepwise selection procedures as linear terms or as dummy variables obtained after grouping. The assumption of linearity may be incorrect, leading to a misspecified final model in which a relevant variable may not be included (e.g. because the true relationship with the outcome is non-monotonic) or in which the assumed functional form differs substantially from the unknown true form. Categorization introduces problems of defining cutpoints (Altman *et al.*, 1994), overparameterization and loss of efficiency (Morgan and Elashoff, 1986; Lagakos, 1988). In any case a cutpoint model is an unrealistic way to describe a smooth relationship between a predictor and an outcome variable. An alternative approach is to keep the variable continuous and to allow some form of non-linearity. Previously, quadratic or cubic polynomials have been used, but the range of curve shapes afforded by conventional low order polynomials is limited. Box and Tidwell (1962) proposed a method of determining a power transform of a predictor. A more general family of parametric models, proposed by Royston and Altman (1994), is based on so-called fractional polynomial (FP) functions. Here, one, two or more terms of the form X^p are fitted, the exponents p being chosen from a small preselected set of integer and non-integer values. FP functions encompass conventional polynomials as a special case.

Royston and Altman (1994) dealt mainly with the case of a single predictor, but they also suggested and illustrated an algorithm for fitting FPs in multivariable models. Although FPs have only a small number of terms, even with one predictor they provide a rich class of possible functional forms leading to a reasonable fit to the data in many cases. However, the considerable flexibility that is available in multivariable models with FPs in each of several predictors may cause serious overfitting and could lead to a final model with features which conflict with current medical knowledge. Achieving consistency in this respect should be one of the central aims of the analyst (Harrell *et al.*, 1996). Here we illustrate the FP approach and demonstrate its advantages in contrast with a conventional backward selection procedure in which the variables are first categorized. The work is based around analyses of data on prognostic factors for breast cancer survival and diagnostic indicators for malignant breast tumours. We propose modifications to the multivariable FP procedure of Royston and Altman (1994) which are designed to reflect basic medical knowledge of the types of relationship to be expected between certain predictors and risk. Aspects of the stability of the final model selected by the procedure are investigated by bootstrap resampling methods (Altman and Andersen, 1989; Chen and George, 1985; Sauerbrei and Schumacher, 1992).

Nonparametric regression methods are complementary or alternative to parametric modelling of curved relationships. Loosely speaking, three main classes of model may be discerned, depending on the type of basis function used: regression splines, smoothing splines and kernel methods. All have several variants and a large literature which we do not attempt to survey here. A useful text-book covering spline and kernel regression is Eubank (1988), and well-written monographs on smoothing splines are by Hastie and Tibshirani (1990) and Green and Silverman (1994). Although originally described in terms of least squares regression with a single predictor, most of the approaches have been (or may be) extended to the generalized linear models framework and to proportional hazards (Cox) regression. The choice is somewhat more limited when models with several predictors are required. One approach is the generalized additive model (GAM), described in detail by Hastie and Tibshirani (1990). We use a simple technique based on GAMs to check the functional form of the predictors in our final FP model.

As with the usual stepwise procedures, our multivariable FP approach uses *P*-values as stopping criteria for selecting the final model. In applications the most commonly used nominal *P*-value for variables selection is 0.05, but depending on the aims of the study smaller or larger values, which result in models with respectively smaller or larger numbers of parameters, may be preferred (Sauerbrei, 1999). Here we investigate the influence of the nominal *P*-value on the complexity of the final model and the related aspect of overfitting by calculating parameterwise shrinkage factors (PWSFs) (Sauerbrei, 1999), which is an extension of the approach of Verweij and van Houwelingen (1993).

In medical studies it is often desirable to classify patients into several groups with different prognoses (e.g. poor, moderate and good) or final diagnoses (e.g. benign or malignant tumour). Using the final models from FP multivariable selection procedures and the usual backward elimination (BE) approach, we categorize the prognostic indices to compare the resulting classification schemes descriptively. In diagnostic studies the primary aim is classification and therefore a rule which assigns patients to one of (usually) two categories is required. The usefulness of the FP approach in building parsimonious models as the basis of diagnostic tests is demonstrated in data from a study to produce an index which differentiates between benign and malignant breast tumours (Sauerbrei *et al.*, 1998).

The paper is structured as follows. Section 2 describes the methods. We discuss issues of variables selection, describe FPs and their fitting procedures for models with single and multiple predictors and suggest how to find simple final models. We examine underfitting and overfitting, bias in parameter estimates and model stability, and we describe an approach based on GAMs to check the functional form. The section ends by considering prognostic and diagnostic classification schemes. Section 3 gives the results of applying the FP procedures and post-fitting investigations to two data sets which are used to motivate and illustrate the methods. The first data set, which is discussed the more extensively, concerns the development of a prognostic model for recurrence-free survival in node positive breast cancer. The second involves a diagnostic model for breast tumour malignancy. Section 4 is a discussion.

The data used in the paper can be obtained from

`http://www.blackwellpublishers.co.uk/rss/`

2. Methods

2.1. Stepwise variables selection and nominal *P*-value

Stepwise selection methods are usually the method of choice for building multivariable prognostic models. Initial decisions on covariate transformations (e.g. logarithmic, square or square root) or categorizations are made univariately and may lead to new base-line variables. Such decisions are guided by medical reasoning, current practice (e.g. categorizations that are common in the literature, based on the distribution of the values) and/or investigation of the influence of a variable on the outcome. Searching for the 'optimal' cutpoint may be seen as an extreme approach of the latter type leading to data-dependent classifications with undesirable characteristics (Altman *et al.*, 1994). Here we adopt categorizations based on medical reasoning.

We use BE as the variables selection strategy. Often BE and forward stepwise selection lead to the same model. Arguments in favour of BE are given by Mantel (1970) and by Sauerbrei (1992) from a simulation study. The predefined nominal P-value is an important feature of stepwise methods which affects the final model chosen. Simulation studies indicate that for uncorrelated factors the true type I error rate of a stepwise algorithm is only slightly higher than the nominal P-value, and that a prespecified value of $P = 0.16$ often gives results identical with selection based on all-subsets regression using Akaike's information criterion AIC (Akaike, 1969; Sauerbrei, 1992). The value 0.16 corresponds to the asymptotic significance level of the all-subsets approach for the linear model using AIC or Mallows's C_{p} (Mallows, 1973; Teräsvirta and Mellin, 1986). We used 0.01, 0.05 and 0.16 as BE nominal P-values to build relatively simple or complex models. Sauerbrei (1999) discusses this aspect in more detail.

2.2. Fractional polynomials

FPs provide a wide range of functional forms which are useful in parametric modelling. Examples of the *ad hoc* use of FPs as regression functions appear in the recent and earlier literature (e.g. Count (1942) and Wingerd (1970)). We give a brief summary of univariate and multivariable FPs here; greater detail and several examples are given by Royston and Altman (1994).

2.2.1. Models with a single covariate

An FP function of degree $m > 0$ for an argument $X > 0$ is essentially $\beta_0 + \Sigma_{j=1}^{m} \beta_j X^{p_j}$, where the β_j are regression parameters and $X^0 \equiv \log(X)$ (Royston and Altman, 1994). The powers $p_1 < \ldots < p_m$ are positive or negative integers or fractions selected from a small predefined set, $\mathcal{P} = \{-2, -1, -0.5, 0, 0.5, 1, 2, \ldots, \max(3, m)\}$. The full definition includes possible 'repeated powers' which involve powers of $\log(X)$. For example an FP of degree $m = 3$ with powers $\mathbf{p} = (-1, -1, 2)$ is of the form $\beta_0 + \beta_1 X^{-1} + \beta_2 X^{-1} \log(X) + \beta_3 X^2$.

An FP model of degree m is considered to have $2m$ degrees of freedom (DF) (excluding the constant, β_0): 1 DF for each β and 1 DF for each power. This is a slight overestimate of the effective number of DF (Royston and Altman, 1994). The deviance is defined as minus twice the maximized log-likelihood. This is not the standard definition, but deviance differences between models, which are used for statistical inference, are unaffected.

The best first-degree FP for X is that with the smallest deviance among the eight models with one regressor ($X^{-2}, X^{-1}, \ldots, X^3$; see set $\mathcal{P}$). Similarly the best second-degree FP is the model with lowest deviance among those with all possible pairs of powers from $\mathcal{P}$: $(-2, -2), (-2, -1), \ldots, (2, 3), (3, 3)$, i.e. $(X^{-2}, X^{-2} \log(X)), (X^{-2}, X^{-1}), \ldots, (X^2, X^3), (X^3, X^3 \log(X))$.

The second-degree FP with minimal deviance is preferred at the α% level to the best first-degree FP if the deviance difference exceeds the $(100 - \alpha)$-percentile of χ^2 with 2 DF. Otherwise, the first-degree FP is preferred to a linear term if the corresponding deviance difference exceeds the $(100 - \alpha)$-percentile of χ^2 on 1 DF. Inference using these particular choices of DF is based on an argument in which FPs are seen as embedded within a more general class of non-linear models; see Royston and Altman (1994) for details. Accordingly the comparison between best fitting first- and second-degree FP models is a nested hypothesis testing problem, not a non-nested one as might at first sight appear.

2.2.2. Models with several covariates

Consider now the case of k continuous and b binary covariates as possible regressors. Royston and Altman (1994) suggested an iterative estimation algorithm, analogous to backfitting (Breiman and Friedman, 1985), to handle the problem of determining the FP with the smallest deviance for all the many possible power vectors $\mathbf{p}_1, \ldots, \mathbf{p}_k$ and corresponding degrees $m_1, \ldots, m_k$. Since it does not process all combinations, such an algorithm may not find the model with minimal deviance, but we believe it unlikely that its chosen model will have a substantially higher deviance than the 'optimal' model.

The model initially comprises a linear term for each of the $k + b$ variables. The best FP for each continuous covariate (taken in an arbitrary order) is found in turn, the choice of 'best' being based on deviance differences as explained at the end of Section 2.2.1. Included in the model are all the binary covariates and variables representing the FP functions of continuous covariates processed previously. The first cycle is complete when all the variables have been processed once. Further cycles are undertaken until the model stabilizes, i.e. when none of the FP functions changes. Convergence usually takes 2–4 cycles.

Variables selection is implemented by using a form of BE, as follows. The procedure again starts with a model in which each regressor is entered as a linear term. During each cycle, each binary covariate is tested for removal from the current model at a predefined nominal P-value α. For each continuous covariate, the best fit second-degree FP is tested at the α-level against the best fit first-degree FP, as described earlier. If this is not significant, the first-degree FP is tested against a straight line. If this test is not significant, the linear term is likewise tested and finally, if the linear term is not significant, the variable is eliminated for the remainder of the present cycle. Eliminated variables are re-entered at the next cycle and may be selected or again eliminated. The procedure is repeated until the model is stable with respect to the FP functional forms and the variables selected.

The procedure described above requires a minor modification when some of the predictors are highly correlated, as the final model may depend on the order of processing the variables. To ensure a unique final model, we initially order the covariates according to their P-values for removal singly from the linear model which includes all the covariates. The best FP transformation is found for the predictor with the smallest P-value first, then that for the predictor with the next smallest P-value, and so on.

2.3. Finding parsimonious and medically consistent final models

The simplicity (parsimony) of a final model and its consistency with medical knowledge are important criteria for its usefulness and acceptability. Here we consider the issues of monotonicity and asymptotic behaviour of model functions in the context of the multivariable FP procedure described above.

In principle, functions with an asymptote should be used to model relationships where the outcome is expected to 'level off' at high values of X. In practice such relationships are almost invariably monotonic as well as asymptotic. We propose using functions of this type to handle the incorporation of medical knowledge into our multivariable selection procedure. Of FPs only those of degree 1 are guaranteed monotonic, and then only those with negative powers (i.e. rectangular hyperbolae) have an asymptote. Such FPs do not always seem to fit the data adequately.

A suitable alternative class of functions is the negative exponentials, $\exp(-\gamma X)$ for $\gamma > 0$. We propose applying an *a priori* exponential transformation to any covariate which medical knowledge strongly indicates should be monotonically and asymptotically related to the outcome. The value of the parameter γ must be estimated from the data, and a method is described in Appendix A. Suppose that the estimate is γ_0, and let $X_* = \exp(-\gamma_0 X)$. For such variables only degree 1 FPs with powers from the reduced set {0.5, 1, 2, 3} should be used in the multivariable FP procedure. The range of available functions of X is thereby seriously reduced and the approach should be used only when the prior evidence about the type of relationship is sufficiently clear cut. As a check on model assumptions which are based on medical knowledge, a second-degree FP in X may be added to a final model which already has X_*. The reduction in deviance is tested against χ^2 on 4 DF. See Appendix A for the rationale for this test. A change in deviance that is significant, say, at the 0.1% level contradicts the original assumptions of monotonicity and/or asymptotic behaviour, or may point to problems with the data.

Finally, again to reduce overfitting, we limit FPs for ordinal covariates which have five categories or fewer to degree 1.

The principle of combining powers from $\mathcal{P}$ may be used to produce FPs of degree 3 or more. However experience suggests that degree 2 is sufficient in most medical applications. Specifically, in multivariable modelling, we recommend fitting FPs of degree no higher than 2 to reduce the chance of selecting overcomplex and 'uninterpretable' models which may be the result of substantial overfitting.

2.4. Model stability and shrinkage

Variables selection strategies are often criticized because the final regression model that they produce is 'invalid'. Two main aspects are replication stability in selecting the final model and the bias of its parameter estimates, which may be considerable (Miller, 1990). The bootstrap (Efron, 1979) may be used to investigate variation among the variables chosen in a final model. A random sample of size n drawn with replacement from the original observations is called a bootstrap replication. A large number of bootstrap replications, say M, is taken and treated as M independent samples. In each replication, the variables selected and their corresponding powers are determined by using the selection strategy summarized above. 100 replications may be regarded as sufficient for this investigation (Sauerbrei and Schumacher, 1992).

Several methods which attempt to correct for the selection bias of the final model are available. With survival data, Verweij and van Houwelingen (1993) used a cross-validated likelihood as a possible criterion for model selection. In addition they suggested a shrinkage factor based on cross-validation calibration as a way of reducing the bias in parameter estimates caused by model building. Suppose that for the ith of n patients, who has covariate vector $\mathbf{x}_i$, a prognostic index $\mathrm{PI}_i = \mathbf{x}_i\hat{\beta}$ is derived from a Cox model. Now take PI_i as the only covariate in a new Cox model. If the corresponding regression parameter is c, the log-likelihood $l(c)$ is maximized at $\hat{c} = 1$. Now consider the cross-validated index $\mathrm{PI}_{(-i)} = \mathbf{x}_i\hat{\boldsymbol{\beta}}_{(-i)}$ as the only covariate, $\hat{\boldsymbol{\beta}}_{(-i)}$ being the parameter vector estimated using all observations except the ith. The corresponding log-likelihood $l^*(c)$ is maximized by a regression estimate c^*, which is usually less than 1 and which is interpreted by Verweij and van Houwelingen (1993) as a shrinkage factor. A value of c^* close to 1 may indicate that there is hardly any overfitting, whereas a small value (say less than 0.7) indicates overfitting.

In an extension of the idea, Sauerbrei (1999) proposes PWSFs estimated in a similar fashion. The initial Cox model is fitted after standardizing each of the p regressors to have mean 0 and standard deviation 1. Standardization affects neither the overall log-likelihood nor the P-value of each regressor and is used in a related approach of Tibshirani (1996) in which selection and shrinkage are combined. Then PWSFs $\mathbf{c} = (c_1, \ldots, c_p)$ are estimated by maximizing $l^*(\mathbf{c})$ for a model whose p covariates are the p terms $x_{ij}\hat{\beta}_{(-i)j}$ which sum to $\mathrm{PI}_{(-i)}$. As with the approach of Verweij and van Houwelingen (1993) n Cox models must be fitted, but the programming is easily done with standard packages. It is necessary only to programme a loop of length n which eliminates the ith patient, estimates $\hat{\beta}_{(-i)j}$ for the remaining $n-1$ patients and stores the results. For the ith patient the p terms $x_{i1}\hat{\beta}_{(-i)1}, \ldots, x_{ip}\hat{\beta}_{(-i)p}$ are calculated. The PWSFs are the usual regression coefficients from a Cox model with the p terms as covariates.

2.5. *Checking the functional form*

Several researchers propose the use of nonparametric methods for model criticism, either informally (graphically) or by the application of formal goodness-of-fit tests based on nonparametric models. Here we use GAMs (Hastie and Tibshirani, 1990) to check that important features of the data have not been neglected in our final FP models. GAMs are typically implemented with cubic interpolation splines as the smoothing 'engine', though other choices are possible. Models are fitted by maximizing the likelihood subject to a roughness penalty which controls the smoothness of the final fitted curve (see also Green and Silverman (1994)). The penalty and hence the degree of smoothness are conveniently expressed as 'equivalent degrees of freedom' (EDF). The EDF and the penalized log-likelihood have a role in approximate significance testing between nested models, though unfortunately the asymptotic distribution of a difference in penalized log-likelihoods is unknown.

Hastie and Tibshirani (1990) suggested curves with 4 EDF for general smoothing. We fit a GAM with 4 EDF separately to each continuous covariate in a final FP model, keeping the functional form of the remaining predictors unchanged. In other words, each model has just one nonparametric non-linear term, the rest being FPs or linear. We plot the resulting smooth (partial prognostic index) and its pointwise 95% confidence interval against the covariate and include in the plot the corresponding curve from the FP model. The fitted values for both curves are standardized to have zero mean. We look for consistency between the nonparametric and parametric fits, informally guided by the nonparametric confidence intervals. Although the (penalized) deviance from each GAM is available, it is unclear how to compare it formally with the deviance from the FP model, since the two models are non-nested and the fitting criteria differ.

2.6. *Prognostic and diagnostic classification schemes*

In survival or logistic regression analysis, the estimated PI or diagnostic index DI for a given patient may be used to estimate relative risk (hazard ratio or odds ratio). In the first example (in the survival context) our interest is in identifying subgroups of patients with different risks of failure. The aim is to define groups which are well separated and sufficiently substantial to be useful in a clinical setting. We use the quantiles of the distribution of PI to define three groups of about equal size. Survival rates are estimated for each subgroup. The estimated relative risks in a Cox model with two dummy variables (indicating group affiliation) are used to quantify intergroup differences.

In the second example (using logistic regression) DI is converted to the estimated probability of having cancer by using the 'antilogit' function $\exp(\mathrm{DI})/\{1+\exp(\mathrm{DI})\}$. Each patient may be classified as belonging to group C (cancer) if her probability exceeds a chosen cutpoint, or otherwise as belonging to group B (benign disease). The sensitivity (the percentage of correct classifications in group C) and specificity (the percentage of correct classifications in group B) are estimated by comparing the classification based on DI with the final diagnosis. An increase in the cutpoint leads to an increase in the specificity but to a decrease in the sensitivity. The behaviour of all possible cutpoints is conventionally represented as a receiver operating characteristic (ROC) curve (Hanley and McNeil, 1982), which is a plot of sensitivity *versus* 1 minus specificity. We compare the diagnostic abilities of different indices by presenting ROC curves, by giving several values of sensitivity and specificity, as well as the area under the ROC curves, a measure of overall discrimination.

2.7. Approach to analysis

In the next section, we compare three approaches to the modelling of the cancer data sets. Initially we apply categorizations to the continuous variables (Byar, 1984) and use conventional methods of analysis, including stepwise variables selection in the multivariable analyses. Secondly we retain the continuous variables as continuous and fit univariate and multivariable FP models as described in Sections 2.2.1 and 2.2.2 respectively. At this stage we ignore possible inconsistencies in the final model with medical knowledge and the need to incorporate monotonic and/or asymptotic relationships as discussed in Section 2.3. We then consider improvements which lead to a final FP model with properties that are consistent with medical knowledge. For the node positive breast cancer data, we check components of the final FP model for possible underfitting against GAMs, consider possible overestimation caused by model building and examine the stability of the various models obtained. Finally we compare the prognostic or diagnostic abilities of the various models.

3. Results

3.1. Prognostic factors in node positive breast cancer

From July 1984 to December 1989, the German Breast Cancer Study Group recruited 720 patients with primary node positive breast cancer into the Comprehensive Cohort Study (Schmoor *et al.*, 1996). Randomized and non-randomized patients were eligible, and about two-thirds were entered into the randomized part. The effectiveness of three *versus* six cycles of chemotherapy and of additional hormonal treatment with tamoxifen were investigated in a 2 × 2 design. After a median follow-up time of nearly 5 years, 312 patients had had at least one recurrence of the disease or died. The recurrence-free survival time of the 686 patients (with 299 events) who had complete data for the standard factors age, tumour size, number of positive lymph nodes, progesterone and oestrogen receptor status, menopausal status and tumour grade is analysed in this paper. In the randomized part of the study, the number of cycles of chemotherapy has no influence on recurrence-free survival (Schumacher *et al.*, 1994) and is not considered here in the analysis of prognostic factors. All analyses are adjusted for hormonal treatment, but estimates of the treatment effect itself are not of interest for the present purposes.

The medically based categorizations given in Table 1 were used by Schumacher *et al.* (1994) for analysis of the randomized trial. The tumour size, number of lymph nodes and tumour grade are coded to reflect their expected ordinal relationship with survival, whereas age is coded as a categorical variable with patients of age 45 years and under as the base-line group. The latter decision was based on discussion in the literature of a possible non-linear relationship between age and survival. For all the analyses reported below, hormonal

Table 1. Parameter estimates with standard errors SE for univariate and multivariable Cox regression models of categorical and categorized covariates, all adjusted for hormonal treatment†

Variable	*Name*	*Categorization group*	*Univariate models*		*Multivariable models*			
			$\hat{\beta}$	*SE*	*Full*		*Model I*	
					$\hat{\beta}$	*SE*	$\hat{\beta}$	*SE*
Age (years), X_1	—	$\leqslant 45$	0.00	—	0.00	—	—	—
	X_{1a}	45–60	−0.24	0.14	−0.40	0.18	Out	—
	X_{1b}	>60	−0.11	0.16	−0.38	0.23	Out	—
Menopausal status, X_2	—	Pre	0.00	—	0.00	—	—	—
	X_2	Post	0.15	0.12	0.27	0.17	Out	—
Tumour size (mm), X_3	—	$\leqslant 20$	0.00	—	0.00	—	—	—
	X_{3a}	>20	0.30	0.15	0.21	0.15	Out	—
	X_{3b}	>30	0.25	0.13	0.06	0.13	Out	—
Tumour grade, X_4	—	1	0.00	—	0.00	—	0.00	—
	X_{4a}	$\geqslant 2$	0.87	0.25	0.55	0.25	0.55	0.25
	X_{4b}	3	0.25	0.13	0.01	0.14	Out	—
Number of positive lymph nodes, X_5	—	$\leqslant 3$	0.00	—	0.00	—	0.00	—
	X_{5a}	>3	0.77	0.13	0.68	0.14	0.73	0.13
	X_{5b}	>9	0.61	0.15	0.58	0.15	0.57	0.15
Progesterone receptor (fmol), X_6	—	$\geqslant 20$	0.00	—	0.00	—	0.00	—
	X_{6a}	< 20	0.77	0.12	0.61	0.14	0.64	0.12
Oestrogen receptor (fmol), X_7	—	$\geqslant 20$	0.00	—	0.00	—	—	—
	X_{7a}	< 20	0.42	0.12	0.01	0.14	Out	—

†Model I was chosen by BE with level $\alpha = 0.05$. 'Out' for a given variable indicates exclusion from the relevant model.

treatment status is included in every model. The prognostic factors comprise five continuous, one binary and one ordinal variable (tumour grade). For the ordinal variable we use two dummy variables, resulting in a total of $b = 3$ binary and $k = 5$ continuous variables.

3.1.1. Univariate analysis

The left-hand portion of Table 1 shows the results of fitting Cox proportional hazards regression models for recurrence-free survival using each of the categorized and categorical covariates singly. The binary variables X_{1a}, X_{1b} etc. are coded 1 if an individual's value falls into the relevant category and 0 otherwise. For ordinal variables, the chosen method of coding (see Table 1) is preferable if stepwise selection procedures are used. In the univariate models all variables except age and menopausal status are significant at the 1% level. There is a major difference in prognosis for the ordinal variable grade (X_4) between grade I and grade II/III (X_{4a}), but the difference between grade II/III and III (X_{4b}, the effect of grade III compared with grade II) is much less pronounced. There are large differences between all three categories for tumour size (X_3) and number of positive lymph nodes (X_5). These results are consistent with findings in the medical literature.

Table 2 shows the results of fitting univariate FP models to each of the continuous covariates. The difference in deviance for each preferred FP model (using nominal P-value $\alpha = 0.05$) is shown in bold type. All variables show a highly significant influence on recurrence-free survival. A linear model is chosen once (for X_3), an $m = 1$ model twice (X_6, X_7) and an $m = 2$ model twice (X_1, X_5). In contrast with the analysis based on categorized covariates, age is identified as an important predictor for recurrence-free survival. The deviances for the best fitting FPs suggest that non-linear relationships are a feature of the

Table 2. Univariate FP analyses for continuous covariates, all adjusted for hormonal treatment†

Variable	*Deviance difference (DF)*				*Powers*	
	Linear versus none (1)	*m =1 versus linear (1)*	*m =2 versus m =1 (2)*	*m =2 versus none (4)*	*m =1*	*m =2*
X_1	0.01	4.09	**13.96**	18.06	−2	−1, −1
X_3	**16.14**	2.95	1.40	20.49	−0.5	−1, 3
X_5	49.65	24.53	**9.35**	83.53	0	−2, −1
X_6	32.91	**16.29**	2.01	51.21	0	−0.5, 0
X_7	3.46	**13.00**	5.16	21.62	0	−2, −1

†The deviance difference for the preferred model is shown in bold.

data. We defer the question of possible monotonic and/or asymptotic relationships discussed in Section 2.3 until Section 3.1.3.

3.1.2. *Multivariable analysis*

Table 1 also shows a Cox regression analysis of the full model comprising all categorical and categorized covariates (which has deviance 3433.24) and the results of a BE procedure with nominal P-value $\alpha = 0.05$. If a likelihood ratio test is applied to the sets of dummy variables which jointly represent each categorized variable (e.g. with 2 DF for X_1), only X_5 and X_6 are significant at the 5% level in the full model. If each dummy variable is considered separately then X_{1a} and X_{4a} are also significant. The BE procedure results in a model which we call model I comprising X_{4a}, X_{5a}, X_{5b} and X_{6a}. Its deviance is 3441.56.

The results of applying the multivariable FP fitting algorithm (Section 2.2) are shown in Table 3. The final model (model II) comprises X_{4a} and FPs for X_1, X_5 and X_6 and has deviance 3420.72. For each of the four variables included in model II, the differences in deviance reported in Table 3 are obtained from multivariable models which include the other three variables. The deviance differences for variables that are not in model II are calculated after adding each corresponding term to model II. Note the inclusion of age (X_1), which (as in the univariate analysis) is excluded from the categorized model I but which FP analysis reveals to be significantly related to the risk of recurrence. If we were to allow only linear

Table 3. Multivariable FP models for recurrence-free survival time

Variable	*Deviance difference*				*Model II: powers or exclusion*
	Linear versus none	*m =1 versus linear*	*m =2 versus m =1*	*m =2 versus none*	
Continuous					
X_1	0.02	3.09	**16.23**	19.34	−2, −0.5
X_3	1.96	3.02	0.33	5.31	Out
X_5	43.07	23.77	**7.30**	74.14	−2, −1
X_6	24.81	**6.98**	1.33	33.12	0.5
X_7	0.76	1.15	0.25	2.16	Out
Binary					
X_2	0.21				Out
X_{4a}	**4.59**				1
X_{4b}	0.14				Out

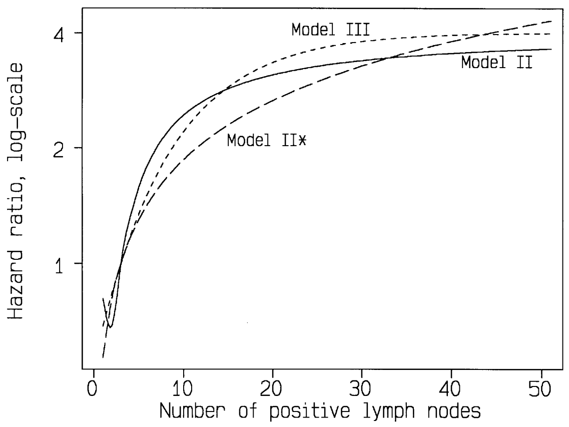


Fig. 1. Three models for the effect of the number of positive lymph nodes on the hazard of tumour recurrence

relationships or simple transformations, the strong influence of age on recurrence-free survival would be missed.

3.1.3. *Incorporation of medical knowledge*

In model II, a non-monotonic second-degree FP is chosen for age (X_1) and the number of lymph nodes (X_5), the latter with powers $(-2, -1)$. The full curve in Fig. 1 represents the X_5-component of the fitted log-hazard-ratio (i.e. $\mathrm{PI}_{\mathrm{nodes}}$) from this model. The quantity actually plotted is $\hat{\beta}_{51}(X_5^{-2} - 3^{-2}) + \hat{\beta}_{52}(X_5^{-1} - 3^{-1})$, making three positive lymph nodes (the sample median) the reference point for the curve. For completeness, Fig. 1 shows the curve from the best fitting first-degree FP model (called model II*) which has power 0 (i.e. $\log(X_5)$), also adjusted to 0 at $X_5 = 3$. The second-degree FP indicates a decreasing risk with an increase in the number of positive lymph nodes from 1 to 2, which seems biologically implausible since increasing lymph node involvement is known to be associated with a poorer prognosis. The anomalous effect may be caused by overfitting the data. The first-degree FP does behave appropriately but is a somewhat worse fit; the deviance is increased by 7.30 (P-value $\alpha = 0.03$); see Table 3. To produce a model which better accords with medical knowledge, we use the modified FP approach described in Section 2.3.

The preliminary exponential transformation is applied to X_5 after univariate estimation of γ_0 by grid search (see Appendix A), which gives $\gamma_0 = 0.12$. FPs of the transformed variable $X_{5*} = \exp(-0.12X_5)$ are restricted to degree 1 and their powers are limited to the set $\mathcal{P}_+ = \{0.5\ 1, 2, 3\}$, implying a range of uncertainty for γ of 0.06–0.24. A new model (III) is produced by the modified multivariable FP algorithm. Except for X_5 its functional form is the same as model II. Details of both models are given in Table 4.

Model III has deviance 18 lower than that of model I and 2.5 higher than that of model II but with one term fewer than for model II and, more importantly, a functional form for X_5 which is consistent with medical knowledge. Addition of the best fitting second-degree FP in X_5 to model III reduces the deviance by 3.72 on 4 DF (P-value $\alpha = 0.4$), so the model passes the check for underfitting described in Section 2.3. A 95% profile-likelihood-based confidence interval for γ within model III is (0.06, 0.26), which is similar to the range of γ-values searched in the FP analysis of X_{5*}.

Fig. 1 shows also the fitted log-hazard-ratio curve for X_5 from model III, again adjusted to

Table 4. Parameter estimates for models II and III for recurrence-free survival†

Variable	*Transformation*	*Model II*		*Model III*	
		$\hat{\beta}$	*SE*	$\hat{\beta}$	*SE*
Age	$(X_1/50)^{-2}$	1.79	0.33	1.74	0.33
	$(X_1/50)^{-0.5}$	−8.02	1.75	−7.82	1.75
Tumour grade ⩾ 2	X_{4a}	0.50	0.25	0.52	0.25
Number of positive lymph nodes	X_5^{-2}	3.88	0.77	—	—
	X_5^{-1}	−5.49	0.86	—	—
	$\exp(-0.12X_5)$	—	—	−1.98	0.23
Progesterone receptors	$(X_6+1)^{0.5}$	−0.057	0.011	−0.058	0.011

†Age is divided by 50 before transformation to improve the scaling of the regression coefficients.

0 at $X_5 = 3$. The anomalous effect for a small number of positive nodes is eliminated. The relative risk from the new model rises more slowly than the second-degree FP in model II, but more quickly than the simple log-transformation (model II*). The steep increase in the hazard ratio up to about 15 positive nodes and the flattening for higher numbers seem consistent with medical knowledge.

3.1.4. *Model stability*

The stability of model III was investigated by bootstrap resampling. The FP selection algorithm was applied to each of 200 bootstrap samples, with restrictions placed on permitted FP degrees as before. X_{5*} was used in each replication without recalculating the preliminary exponential transformation of X_5. The variables selected and their powers are noted. The results for the first eight replicates are given in Table 5 and demonstrate that all eight models differ somewhat.

Bootstrap inclusion frequencies and other summaries for each variable are given in Table 6. The inclusion proportions for the variables X_1, X_{5*} and X_6 are 94%, 100% and 98.5% respectively, confirming their importance. X_{4a} entered in 59% showing that it appears to have weak prognostic influence. The need to include it in a parsimonious final model is in doubt. Among the variables not entering model III, the inclusion proportion for X_3 is 42.5%, showing that it also may have prognostic relevance. The other variables have inclusion proportions that are less than 20%, indicating that they are at most weakly prognostic and are appropriately excluded from the final model. An interpretation of the summary of selected powers is somewhat problematic. The most common powers for X_1, namely

Table 5. Stability of variable selection and FP functions in model III for recurrence-free survival

Variable	*Model III powers*	*Powers for the following bootstrap replicate numbers:*							
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
X_1	−2, −0.5	−1, −1	−1, −0.5	Out	−2, −2	−2, −2	−2, −2	−2, −1	−2, −1
X_2	Out	Out	Out	Out	Out	Out	Out	Out	Out
X_3	Out	−0.5, −0.5	Out	Out	Out	−2	Out	Out	Out
X_{4a}	1	Out	Out	1	1	Out	Out	1	1
X_{4b}	Out	Out	Out	Out	Out	Out	Out	Out	Out
X_{5*}	1	1	1	1	1	1	1	1	1
X_6	0.5	0.5	1	1	0.5	0	−1, −0.5	0	0.5
X_7	Out	Out	Out	Out	Out	Out	Out	Out	1

Table 6. Inclusion percentages of variables for 200 bootstrap replicates and numbers of different types of model chosen

Variables	*Bootstrap inclusion frequency (%)*	*Frequency (%) of models chosen*			*Most common model chosen*	
		Linear	$m = 1$	$m = 2$	*Powers*	*Frequency (%)*
Continuous						
X_1	94	0.5	14.5	79	−2, −2	29
X_3	42.5	14	24	4.5	1	14
X_{5*}	100	96	4	—	1	96
X_6	98.5	23	68.5	7	0	35.5
X_7	13	1.5	5	6.5	−2, −2	5
Binary						
X_2	17.5					
X_{4a}	59					
X_{4b}	9					

(−2, −2), are similar but not the same as those in model III; nevertheless, a second-degree FP is chosen in 79% of replicates. The actual powers of (−2, −0.5) for X_1 in model III are chosen only 17 times, and model III itself does not appear at all among the 200 replicates. A power of 1 is chosen for X_{5*} in 96% of replicates.

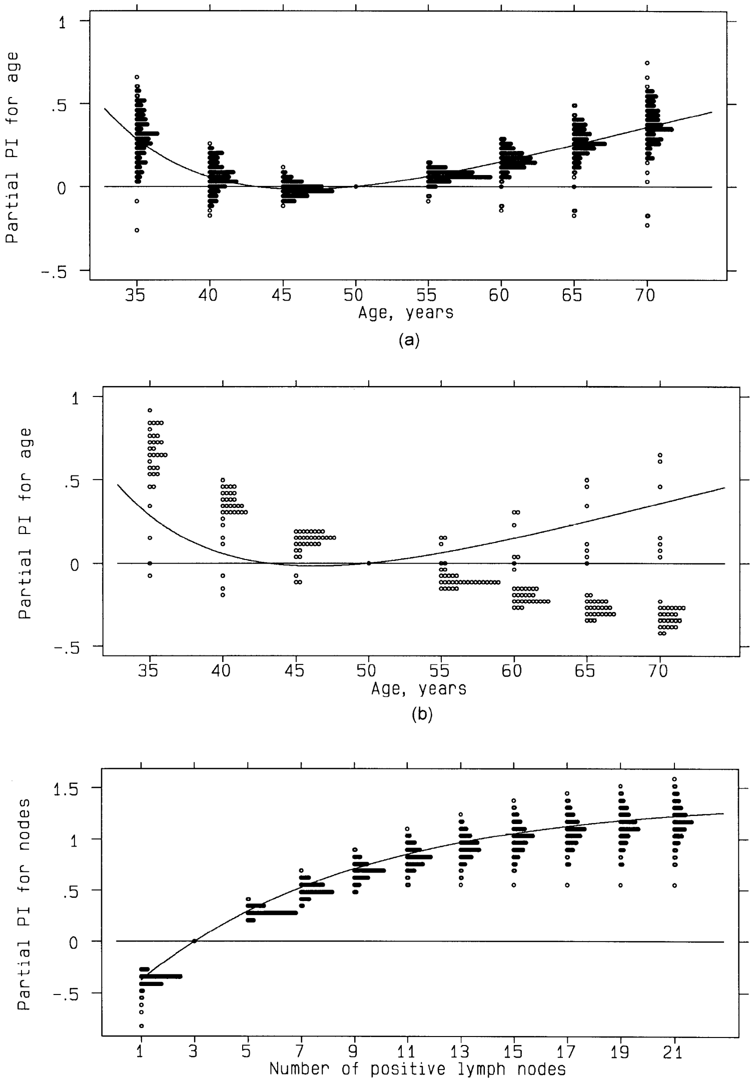
The stability of the effects of age (X_1) and number of positive lymph nodes (X_5) are further investigated as follows. For each of the 188 bootstrap replicates in which X_1 is selected, the age component of PI is adjusted to 0 for a particular age in the same way as described in Section 2.3 for the number of positive lymph nodes. This gives an adjusted partial index, $\mathrm{PI}_{\mathrm{age}}$ say. We chose to adjust to age 50 years. Fig. 2 shows dotplots of $\mathrm{PI}_{\mathrm{age}}$ for ages between 35 and 70 years in 5-year increments.

Fig. 2(a) shows the 155 out of 200 replicates in which age but not menopausal status (X_2) is included in the model, whereas in Fig. 2(b) age and menopausal status are both included (in 33 replicates). Because of the strong correlation the distribution of $\mathrm{PI}_{\mathrm{age}}$ depends on whether or not menopausal status is also selected. The difficulty of interpreting regression estimates in models obtained by variables selection procedures is well known and is further illustrated when interpreting the results of bootstrap resampling of such procedures applied to correlated covariates. Analogous results for X_5, with $\mathrm{PI}_{\mathrm{nodes}}$ adjusted to 0 at three nodes as before, are shown in Fig. 2(c). The effect of X_5 increases strongly with the number of positive nodes. No peculiarities in the distribution of $\mathrm{PI}_{\mathrm{nodes}}$ are evident.

We conclude that, according to the limited criteria described here, model III, which excludes menopausal status and for which Fig. 2(a) shows the relevant functional form for age, is reasonably stable.

3.1.5. *Model checking with generalized additive models*

Fig. 3 shows the results of fitting a GAM with 4 EDF to each continuous variable in model III, as described in Section 2.5. The impression is of generally close agreement between the FP and GAM fits. For patients of age less than 30 years the FP model indicates a larger effect than the GAM, and for those near 60 years the GAM appears to show a small peak in the hazard ratio (see Fig. 3(a)). A third-degree FP (which may have up to two turning-points) produces a curve somewhat resembling the GAM for older ages, but the reduction in deviance is not significant. For more than about 25 positive lymph nodes the GAM fit shows



(a)

(b)

(c)

Fig. 2. Adjusted partial prognostic index from model III as a continuous curve in each plot and bootstrap distributions (200 replications) of adjusted partial prognostic indices for selected covariate values: (a) age but not menopausal status included in the model (155 replications); (b) age and menopausal status both included (33 replications); (c) number of positive lymph nodes (200 replications) (the curves for age and number of positive lymph nodes are adjusted to 50 years and three nodes respectively)

a decreasing risk which is strongly against medical knowledge. It may be caused by the sparseness of the data in this region. As may be seen in the more detailed Fig. 3(d), the FP model has a much steeper decrease than the GAM for very low progesterone receptor values. This supports the approach that is common in the medical literature in which a binary variable based on a cutpoint between 5 and 20 fmol l^{-1} is used to define receptor negativity. The differences in deviance (model III minus GAM) for X_1, X_5 and X_6 are 3.74, 3.24 and -1.56 respectively.

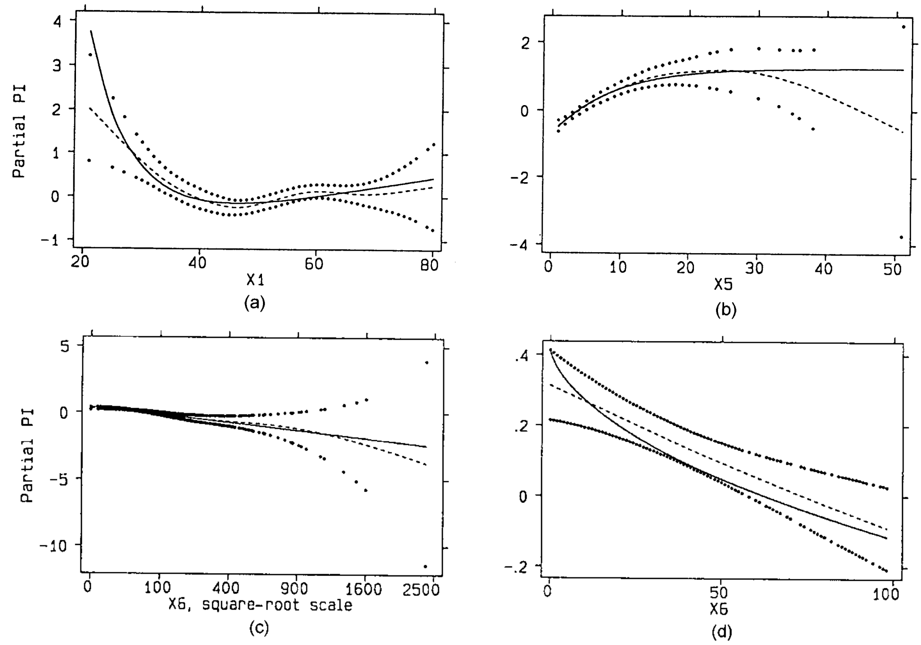


Fig. 3. Comparison between adjusted partial prognostic indices for FP and GAM models for each continuous covariate in model III, conditional on the rest of model III (———, FP fits; - - - -, GAM fits; ·········, 95% confidence bands for GAMs): (a) age; (b) number of positive lymph nodes; (c) progesterone receptor concentration; (d) as for (c) but restricted to the range [0, 100) fmol l^{-1}

3.1.6. Nominal P-value and shrinkage

Models I and III were initially constructed by using the nominal P-value $\alpha = 0.05$. The model selection procedures were repeated using $\alpha = 0.01$ and $\alpha = 0.16$. For $\alpha = 0.01$, X_{4a} was removed from models I and III. Model III (0.01) includes X_1 with powers $(-2, -1)$, X_{5*} with power 1 and X_6 with power 0. For $\alpha = 0.16$, model I was augmented with X_{1a}, X_{1b}, X_2 and X_{3a}, resulting in model I (0.16) which includes all seven candidate variables except for X_7. Model III is unchanged. The deviances for models I (0.16) and III (0.01) are 3433.45 and 3427.88 respectively.

Excepting X_{4a}, our final models I and III seem to include only strong factors because the decrease in the nominal P-value to 0.01 has virtually no other influence on these models. The increase in nominal P-value has no influence on FP model III, but model I (selected by BE) now includes several extra variables whose prognostic relevance seems weak. This interpretation is confirmed by the pattern of the shrinkage factors reported in Table 7.

The global shrinkage factors are nearly identical for all the models (between 0.92 and 0.96). However, the PWSFs clearly indicate that the effects of variables included only in model I (0.16) because of the high nominal P-value of 0.16 seem to be substantially overestimated. The PWSFs for the two age variables (X_{1a}, X_{1b}) and the correlated menopausal status variable (X_2) lie between 0.09 and 0.41. The effects of X_{3a} and X_{4a} seem to be overestimated, but all other PWSFs are between 0.85 and 1.00, indicating that the corresponding estimates of the regresssion coefficients may be close to the unknown true values.

3.1.7. Prognostic classification scheme

No continuous variable appears in model I, and the corresponding prognostic index has only

Table 7. Global factors and PWSFs of variables in four final models for recurrence-free survival†

Variable	*Results for the following models:*			
	I	*I (0.16)*	*II*	*III*
Global	0.96	0.92	0.94	0.95
X_{1a}		0.41		
X_{1b}		0.13		
X_1 (−2)			0.86	0.88
X_1 (−0.5)			0.85	0.86
X_2		0.09		
X_{3a}		0.65		
X_{4a}	0.83	0.82	0.82	0.82
X_{5a}	0.98	1.00		
X_{5b}	0.94	0.94		
X_5 (−2)			0.93	
X_5 (−1)			0.94	
X_{5*}				0.98
X_{6a}	0.98	1.00		
X_6 (0.5)			0.98	0.97

†Variables in model I (0.16) are chosen with selection level 0.16. X_7 does not enter any of the models.

27 distinct values for the 686 patients. A classification based on the exact 33% and 67% quantiles of PI is therefore not possible. Instead three groups comprising the 257 patients with lowest PI (best prognosis), the 185 with median PI and the 244 with largest PI (worst prognosis) are chosen. Since the separation between the groups depends on the size of the two extreme groups, we ensure that the sizes of the prognostic groups for the classification schemes based on models II and III are the same as for model I.

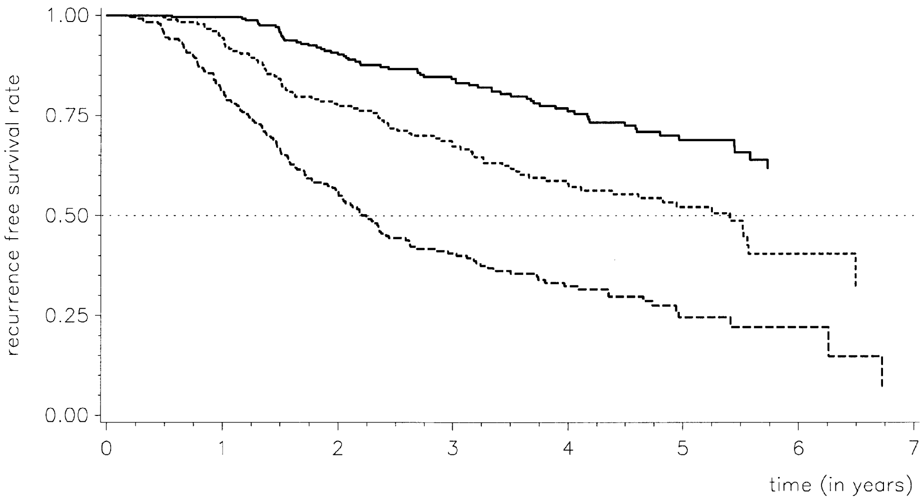
The estimates of recurrence-free survival for model III are given in Fig. 4. The corresponding estimates based on models I and II are similar. The coefficients from the Cox model for the group indicators relative to the group with the best prognosis are given in Table 8.

The prognostic ability of the classification schemes based on the FP models (II and III) is slightly better than for model I. Moving from model III with individual prognostic indices to that based on the classification scheme with prognostic index groups leads to a loss of information which in deviance terms is substantial (an increase of 40.87). The loss is similar in principle to that incurred by using a categorized variable instead of the original continuous variable, perhaps with FP transformation.

3.2. Diagnosis of malignant breast tumours

Between July 1992 and February 1994, 458 consecutive unselected women, referred for surgical biopsy with palpable, mammographic or sonographic findings or with clinical symptoms, presented at the Gynaecology Department of the University of Freiburg. In addition to their usual management, all patients received a colour Doppler examination which was carried out 'blind' to the mammographic findings. The main purpose was not to detect additional lesions but to derive rules based on differential vascularization with which to distinguish benign growths from malignant tumours.

The number of tumour arteries detected by colour flow mapping was counted, followed by a systematic duplex measurement of the peak systolic and diastolic flow velocity in all tumour



Best	257	243	207	164	116	62	22	1
Median	185	169	133	100	72	43	11	2
Worst	244	190	118	67	40	16	3	0

Fig. 4. Prognostic classification schemes for model III: Kaplan–Meier estimates for three prognostic groups based on PI (———, best; - - - - -, median; – – –, worst)

Table 8. Parameter estimates for the group indicator from the three models

Prognostic group	*Model I*		*Model II*		*Model III*	
	$\hat{\beta}$	*SE*	$\hat{\beta}$	*SE*	$\hat{\beta}$	*SE*
Best	0	—	0	—	0	—
Median	0.51	0.17	0.56	0.17	0.71	0.17
Worst	1.33	0.14	1.45	0.15	1.50	0.15

vessels. In each tumour the maximum $V_{\max}$, average V_{av} and sum V_{sum} of all peak systolic flow velocities were calculated. Tumours with no vessels were taken to have zero flow velocities. For comparisons between sides the number of 'contralateral arteries' (the number of arteries in the symmetric region of the contralateral breast) was also investigated. For more details of the study population and the measurement methods see Sauerbrei *et al.* (1998).

We develop diagnostic indices for differentiating between benign and malignant tumours. According to the final histological or cytological diagnosis, 133 (29%) of the patients have cancer (group C) and 325 have a benign tumour (group B). The women in group B were approximately 10 years younger than those in group C (medians 49 and 59 years respectively; the 10% and 90% quantiles are 30 and 65 years in group B and 40 and 76 years in group C). Except for the number of contralateral arteries, major differences in the distributions of all the covariates between the two groups are present (Sauerbrei *et al.*, 1998). The distribution of the number of arteries in the tumour according to diagnostic group is shown in Fig. 5. This variable on its own (model I) is strongly discriminatory and suggests a cancer diagnosis when the number of arteries is high. Nevertheless the distributions overlap.

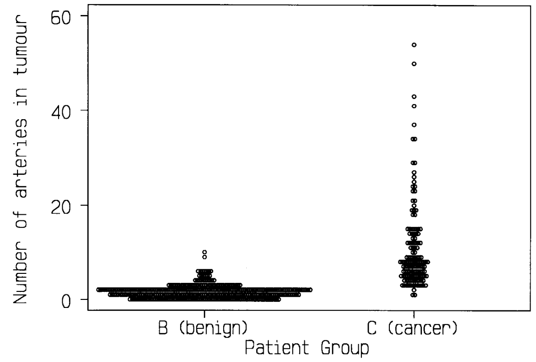


Fig. 5. Distribution of the number of tumour arteries in the benign and cancer groups

Starting in the usual way with all variables untransformed, BE with nominal P-value 0.05 produces a model with age and number of tumour arteries (model II). The sensitivity and specificity are only slightly improved. Sauerbrei *et al.* (1998) examined whether the proportion of correct classifications may be improved by using a multivariable logistic regression model based on transformations of each variable. They first investigated several 'usual' transformations of each variable in a univariate model and then developed the final model (model III) using BE with nominal P-value 0.05 for the transformed variables. Model III contains only age (linear), number of tumour arteries (log-transformed after adding 1 to avoid 0s) and number of contralateral arteries (linear). All the standardized regression coefficients are greater than 3 in absolute value. The velocity measures V_{av}, $V_{\max}$ and V_{sum} are highly correlated with the number of tumour arteries and are eliminated by the selection procedure. Using cutpoints for the probability of cancer based on the diagnostic index from model III, the proportions of correctly classified patients are higher than those from the first two approaches.

We wish to see whether the discrimination may be further improved by the use of FPs in a multivariable model. A sensible assumption is to expect the incremental effect of the number of tumour arteries (X_2) to decrease monotonically to an asymptote, so we apply a negative exponential transformation and use FPs of first degree as described in Section 2.3. The value of γ_0 is found to be 0.21, so $X_{2*} = \exp(-0.21X_2)$. The deviance is substantially smaller than that before transformation but is almost identical with that following log-transformation. The number of contralateral arteries (X_6) has only five distinct values, so FPs in X_6 are restricted to degree 1. Also, 1 is added to X_6 to avoid 0s. Application of the multivariable FP algorithm yields model IV, details of which are given in Table 9.

The usual BE approach and the FP algorithm select the same variables. The deviances for models III and IV are similar (155.6 and 151.2). In terms of deviance, both models are dominated by X_2. The diagnostic indices are

$$-11.71 + 0.0684X_1 + 5.323 \log(X_2 + 1) - 0.685X_6$$

for model III and

$$2.35 + 0.0686X_1 - 13.07 \exp(-0.21X_2) - 0.036(X_6 + 1)^3$$

for model IV. The FP approach confirms that age (X_1) should be included linearly and its

Table 9. Multivariable FP model for the probability of cancer†

Variable	*Deviance difference*				*Model IV: powers or exclusion*
	Linear versus none	*m = 1 versus linear*	*m = 2 versus m = 1*	*m = 2 versus none*	
Age, X_1	19.81	2.56	0.18	22.55	1
Tumour arteries, X_{2*}	336.50	0.00	0.14	336.64	1
Average velocity, X_3	0.61	0.34	1.44	2.39	Out
Sum of velocities, X_4	0.01	0.33	1.56	1.90	Out
Maximum velocity, X_5	0.74	0.56	1.44	2.74	Out
Contralateral arteries, X_6	10.86	4.06	—	—	3

†The asterisk in X_{2*} indicates exponential transformation.

coefficient is nearly identical in each model. The effect of X_2 is as expected, showing an increase in the probability of cancer with an increase in the number of arteries. Initially surprising is the negative association with X_6: the effect of an increase in X_6, conditionally on fixing the other covariates, is actually to reduce the risk of cancer. Thus the risk associated with a large number of tumour arteries is lowered if the contralateral breast also has more arteries than average. This may reflect a tendency for some women to have generally greater breast vascularization than others. In a sense, X_6 may act as a standardization for X_2.

The ROC curves for models III and IV are almost identical and both indicate somewhat better discrimination than with the arteries-only model (Fig. 6). Model II yields a curve (not shown) lying between those for models I and III. The areas under each curve are 0.965, 0.972, 0.977 and 0.977 for models I–IV respectively, and the sensitivities for a given specificity of 92.3% (corresponding to a cutpoint of four arteries in model I) are 88.7%, 91.7%, 95.5% and 95.5%.

A transformation of X_2 has no effect on the ROC curve for the arteries-only model but does affect the deviance, which is 200.3 before transformation and 191.5 afterwards. It also affects the estimated individual probability of having cancer.

4. Discussion

We have shown that the multivariable FP procedure described by Royston and Altman (1994) may be used to construct simple parametric prognostic or diagnostic models. The procedure overcomes the serious problem of arbitrary categorization and in the example data sets has sufficient power to detect some quite strong non-linearities which should be accommodated by the final model. The need to ensure consistency with basic medical knowledge led to refinements of the original strategy which may be seen as improvements. For the breast cancer prognostic study, our final model agrees well with medical knowledge and has a lower deviance than does the conventional model obtained by applying BE to categorized variables. The model is quite simple, a prerequisite for being statistically stable and useful in practical applications.

For the breast cancer diagnosis data set, the multivariable FP approach confirms the model obtained in the original analysis in which only certain standard transformations (logarithm, exponential, square root, reciprocal and square) were considered in an initial step preceding BE of variables. The logarithmic transformation on its own produces a substantial improvement over a linear model. With BE all velocity measurements are omitted, giving a simple model which can easily be used in a clinical setting.

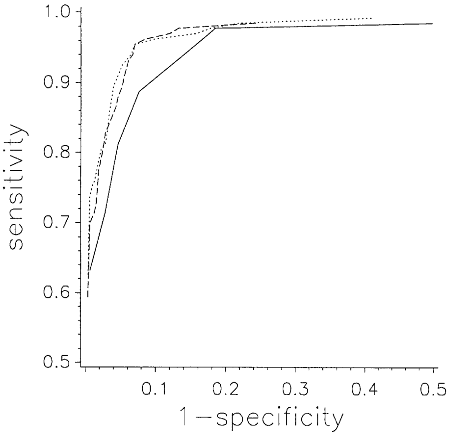


Fig. 6. ROC curves in breast cancer diagnosis, with false positive rates (1− specificity) less than or equal to 0.5: ———, model I; - - - -, model III; ··········, model IV

The FP procedure combines BE with an adaptive algorithm which is analogous to backfitting, which selects the best FP transformation for each variable in turn. The small set $\mathcal{P}$ of powers which defines the possible transformations provides a flexible class of models. To reduce the risk of overfitting that may result in overcomplex or inappropriate models, we have restricted ourselves to FPs of degree 2 or lower. Such functions have at most one turning-point. Even so, we rejected our initial FP model II for the node positive breast cancer data because the non-monotonic behaviour of the fitted function for the number of affected lymph nodes was contrary to medical knowledge. Two distinct requirements may be discerned: one for monotonic functions and another for functions with an asymptote, though often both are required. This type of medical knowledge ought to be incorporated in the statistical analysis of medical studies, but usually it is not. The first requirement is satisfied by fitting only first-degree FPs (i.e. simple power or log-transformations). For the second our approach of using FPs after negative exponential transformation, described in Section 2.3, may be effective. In the prognostic factors study the fit of the exponentially transformed lymph nodes variable is not significantly worse (in deviance terms) than the fit of a second-degree FP of the untransformed variable. The approach produces a more parsimonious model (model III) which fits as well as model II but whose behaviour is more plausible. Although FPs of first degree predominate in model III, it is nevertheless important to include the possibility of choosing second-degree FPs; otherwise statistically important and medically relevant non-monotonicities, such as the effect of age, may be missed.

'Medical knowledge' may of course vary (or at least may be interpreted differently) depending on the disease in question and the researchers analysing the study. In the prognostic factors study we finally decided on a monotonic function with an asymptote only for the number of positive lymph nodes (X_5). The decision was arrived at because the variable clearly dominates in virtually all studies of prognostic factors in breast cancer. There is no clear evidence for a monotonic function for the other continuous variables and we therefore decided on the more flexible, unrestricted FP approach to investigate their functional relationships. However, we did reanalyse the data by using an initial exponential transformation for the variables X_3, X_5, X_6 and X_7, which may be seen as a sensitivity analysis of our decision to use medical knowledge. The most important change was the elimination of

X_{4a}. The variable X_{6*} entered the model with a power of 1 and X_1 with powers $(-2, -1)$. The dominating variable X_{5*} again entered with power 1 and the deviance increased slightly. Generally the sensitivity analysis demonstrated the stability of our final model. Blettner and Sauerbrei (1993) used sensitivity analysis in the analysis of case–control studies to investigate several aspects of multivariable model building. They found a much stronger influence of some of their decisions on the final model than we did here.

Certain technical issues surrounding the FP algorithm need further exploration. We have suggested a method of ordering the variables which guarantees a unique final model once the algorithm has converged, but that model may not be the model with the smallest deviance. Other orderings are possible, e.g. according to the relative strength of the FP relationship between the outcome and each covariate separately, or of each marginally in a full FP model. It would be interesting to find the global minimum deviance by an exhaustive search of all possible FP functions for a few data sets and to compare the results with those from the present iterative procedure and variants of it. However, this would not deal with the variable selection case where sequential processing of variables is intrinsic to the method. Furthermore we have no proof that the algorithm converges in every case, so conceivably it could become stuck, oscillating between two or more models. We have not seen such behaviour, however.

We used resampling methods to investigate stability and possible overfitting. As previously reported for the simpler situation in which stepwise selection is used with predefined functional forms for the predictors, the final model virtually never appears among the bootstrap replications (Chen and George, 1985; Altman and Andersen, 1989; Sauerbrei and Schumacher, 1992). Nevertheless, as Fig. 2 shows, our final FP model in the first data set is reasonably stable. The dominating influence on recurrence-free survival of age, progesterone receptor status and the number of positive lymph nodes is confirmed in 200 bootstrap samples. Tumour grade and tumour size are identified as variables with possible prognostic value. The effect of the number of positive lymph nodes always has a functional form which is similar to that in our final model (see Fig. 2(c)). For age we must distinguish two situations depending on menopausal status, which is correlated with age (Figs 2(a) and 2(b)). In most replications (and as in our final model) menopausal status does not enter the model, and most of the estimated prognostic index functions for age are similar to the age effect in the final model. The inclusion of menopausal status of course influences the functional relationship with age.

By extending the cross-validation approach of Verweij and van Houwelingen (1993) to PWSFs, we showed that selection bias does not seem strongly to influence the parameter estimates in our final model. The investigation also demonstrates serious bias in parameter estimates for factors which are only included in a BE model after changing the selection level to 0.16. We conclude from the resampling and cross-validation results that our final FP model III includes only strongly prognostic factors, with at most one weakly prognostic factor (tumour size) excluded.

For the prognostic study, we divided the patients into three groups of about the same size to examine the separation of classification schemes based on several final models. The number of groups and the sample size in each group are somewhat arbitrary, and many other choices are possible. We know of no general recommendations for implementing this important step. Choosing smaller extreme groups in the three-group case would result in better separation, but at the cost of exaggerating what is already an overoptimistic assessment of prognostic ability. The overoptimism problem has been recognized for a long time (Efron, 1983; Gong, 1986; Phillips *et al.*, 1990; van Houwelingen and le Cessie, 1990; van

Houwelingen and Thorogood, 1995). The results demonstrate that prognostic indices which include different variables and transformations may have similar prognostic abilities, though with a slight advantage for the FP-based approaches.

As a substantive approach, multivariable additive models based on splines or other nonparametric smoothers offer flexibility of functional form. A small number of more or less general spline-based systems of model building has been suggested in the literature. Hastie and Tibshirani (1990) (chapter 9) discussed several ways of selecting variables in the final model while simultaneously determining the amount of smoothing for each term. Their 'backward and forward stepwise selection technique' (p. 260) has some similarities to our FP algorithm. For a given covariate they choose between 0 EDF (term excluded), 1 EDF (linear) and 4 EDF (non-linear smoothing). It is beyond the scope of the present paper to compare these systems with our FP approach, partly because there is a lack of any standard software. Several drawbacks of a fully nonparametric approach may be identified. We are unconvinced that any of the suggested systems of model building has been investigated sufficiently (particularly in medical applications) to encourage one to adopt it as a 'gold standard' for comparison with parametric models. There is some published experience with the GAM approach of Hastie and Tibshirani (1990) (e.g. Hastie *et al.* (1992)). With smoothing splines (and to a lesser extent with regression splines) a concise expression for the final model is not available. Although this may not matter for prediction and graphical display, it makes communication and clinical application difficult. Also, it is unclear how to incorporate medical knowledge in spline models. For example, although a spline curve can be forced to be monotone (Ramsay, 1988) it cannot be forced to have an asymptote. Instabilities in the fitted curve, which are common with splines, are a barrier to interpretation as it is usually unclear whether they are real or artefactual. For example the FP fit for X_{5*} (the negative exponentially transformed version of X_5) in the prognostic study differs considerably from the GAM fit for X_5 (Fig. 3(b)) in that it does not exhibit the medically implausible dip for $X_5 > 30$.

More than 5% of the patients in the diagnostic study are incorrectly classified by the final model III of Sauerbrei *et al.* (1998), which was obtained by BE after inspecting the results of applying a few standard transformations in univariate models. As a result the number of arteries (X_1) and the three velocity measures (X_3, X_4 and X_5) were logarithmically transformed, with no apparent need to transform age or the number of contralateral arteries. Nevertheless the suspicion remained that an important transformation might have been missed, or that a multivariable model with a lower misclassification rate might have been constructed by using more complex transformations or a different combination of them. The results with the FP approach imply that a substantial improvement in discrimination is improbable, even within a very flexible class of transformations in a multivariable setting, and that further logistic regression modelling is unlikely to be profitable.

In conclusion, we believe that it is important to extract as much information as possible from predictors and responses, while trying to avoid overfitting. Retaining continuous variables as continuous in the final model is essential to avoid a loss of information. Furthermore it is imperative that the functional form of the final model should be consistent with medical knowledge. Although this has a Bayesian flavour, we do not explicitly adopt a Bayesian approach to modelling. However, we believe that good statistical practice demands the incorporation of background knowledge, even within a classical approach to analysis. Similar principles are stated by Harrell *et al.* (1996) as 'preliminary steps' in the construction of multivariable prognostic models. The FP approaches described here are a worthwhile step towards realizing these aims.

Appendix A: Parameter estimation for exponentially transformed variables

Ideally, the values of γ for several exponentially transformed variables in a final multivariable model would be estimated jointly by maximum likelihood. This is impractical because an awkward non-linear optimization problem involving a multidimensional likelihood function arises. In univariate models estimation is straightforward, e.g. by a grid search for the maximum likelihood solution over a suitable range of γ-values. For a given predictor determined *a priori* to need exponential transformation, let γ_0 be the maximum likelihood estimate (MLE) of γ in the univariate model with covariate $\exp(-\gamma X)$. The value of γ_0 needs to be accurate only to one or two significant figures and may easily be found by trial and error.

We may exploit the FP approach to provide an approximate solution in the multivariable case as follows. Let $X_* = \exp(-\gamma_0 X)$ denote an exponentially transformed predictor, with γ_0 estimated univariately as just described. The multivariable FP algorithm of Section 2.2.2 is applied to all predictors, but for those of type X_* only degree 1 FPs with positive powers {0.5, 1, 2, 3} from the set $\mathcal{P}$ are used. When an FP in an X_* is selected, a term of the form $X_*^{(p)} = \exp(-p\gamma_0 X)$ appears in the final model. Provided that the exact MLE of γ in the final model lies in or near the interval $[0.5\gamma_0, 3\gamma_0]$, $p\gamma_0$ will be an adequate approximation to it. The condition will hold when X is not too closely related in a functional sense to other predictors. The situation is entirely analogous to that for FPs in X where the final estimated powers may be viewed as approximations to the MLEs of powers that lie on a continuous scale.

The rationale for testing the adequacy of a first-degree FP in an X_* by adding a second-degree FP in X to the final model is as follows. As indicated in Section 2.3, a second-degree FP in X will provide a reasonable fit to most relationships that are found in medical data, where relationships tend to be obscured by ‘noise’. We wish to see whether the model involving X_* can be substantially improved by fitting a different function of X instead. If so, there would be a strong argument against our medically based assumption of a monotonic effect with an asymptote. In general a non-nested hypothesis testing procedure is required. A simple approach is to use a so-called ‘encompassing test’ (Mizon and Richard, 1986). If the addition of an FP in X to the model significantly improves the fit, there is evidence that the FP in X_* is an inadequate predictor compared with the predictor obtained by combining X and X_*. The interpretation is that the FP in X is a significantly better fit than that in X_*.

Appendix B: Software

All the analyses described in this paper were carried out using Stata 5.0 (StataCorp, 1996), an inexpensive general statistics package for Windows, Macintosh and Unix platforms. The command `fracpoly` was used for univariate FP model fitting. The multivariable FP algorithm used here is implemented as an add-in program (‘ado-file’) called `mfracpol` (Royston and Ambler, 1998). The multivariable algorithm will be implemented in the packages SAS and S-PLUS in due course.

References

Akaike, H. (1969) Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, **21**, 243–247.

Altman, D. G. and Andersen, P. K. (1989) Bootstrap investigation of the stability of a Cox regression model. *Statist. Med.*, **8**, 771–783.

Altman, D. G., Lausen, B., Sauerbrei, W. and Schumacher, M. (1994) The dangers of using ‘optimal’ cutpoints in the evaluation of prognostic factors. *J. Natn. Cancer Inst.*, **86**, 829–835.

Blettner, M. and Sauerbrei, W. (1993) Influence of model-building strategies on the results of a case-control study. *Statist. Med.*, **12**, 1325–1338.

Box, G. E. P. and Tidwell, P. W. (1962) Transformation of the independent variables. *Technometrics*, **4**, 531–550.

Breiman, L. and Friedman, J. H. (1985) Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Am. Statist. Ass.*, **80**, 580–619.

Byar, D. P. (1984) Identification of prognostic factors. In *Cancer Clinical Trials—Methods and Practice* (eds M. E. Buyse, M. J. Staquet and R. J. Sylvester), pp. 423–443. Oxford: Oxford University Press.

Chen, C. and George, S. L. (1985) The bootstrap and identification of prognostic factors via Cox’s proportional hazards regression model. *Statist. Med.*, **4**, 39–46.

Count, E. W. (1942) A quantitative analysis of growth in certain human skull dimensions. *Hum. Biol.*, **14**, 143–165.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1–26.

——(1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Statist. Ass.*, **78**, 316–331.

Eubank, R. L. (1988) *Spline Smoothing and Nonparametric Regression*. New York: Dekker.

Galea, M. H., Blamey, R. W., Elston, C. E. and Ellis, I. O. (1992) The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res. Treatmnt*, **22**, 207–219.

Gong, G. (1986) Cross-validation, the jackknife and the bootstrap: excess error estimation in forward logistic regression. *J. Am. Statist. Ass.*, **81**, 108–113.

Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. London: Chapman and Hall.

Hanley, J. A. and McNeil, B. J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and accuracy, and measuring and reducing errors. *Statist. Med.*, **15**, 361–387.

Hastie, T. J., Sleeper, L. and Tibshirani, R. J. (1992) Flexible covariate effects in the proportional hazards model. *Breast Cancer Res. Treatmnt*, **22**, 241–250.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. New York: Chapman and Hall.

van Houwelingen, J. C. and le Cessie, S. (1990) Predictive value of statistical models. *Statist. Med.*, **9**, 1303–1325.

van Houwelingen, J. C. and Thorogood, J. (1995) Construction, validation and updating of a prognostic model for kidney graft survival. *Statist. Med.*, **14**, 1999–2008.

Lagakos, S. W. (1988) Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statist. Med.*, **7**, 257–274.

Mallows, C. L. (1973) Some comments on C_p. *Technometrics*, **15**, 661–675.

Mantel, N. (1970) Why stepdown procedures in variable selection? *Technometrics*, **12**, 621–625.

Miller, A. J. (1990) *Subset Selection in Regression*. New York: Chapman and Hall.

Mizon, G. E. and Richard, J. (1986) The encompassing principle and its application to testing non-nested hypotheses. *Econometrica*, **54**, 657–678.

Morgan, T. M. and Elashoff, R. M. (1986) Effect of categorizing a continuous covariate on the comparison of survival time. *J. Am. Statist. Ass.*, **81**, 917–921.

Phillips, A. N., Thompson, S. G. and Pocock, S. J. (1990) Prognostic scores for detecting a high risk group: estimating the sensitivity when applied to new data. *Statist. Med.*, **9**, 1189–1198.

Ramsay, J. (1988) Monotone splines in action (with discussion). *Statist. Sci.*, **3**, 425–461.

Royston, P. and Altman, D. G. (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Appl. Statist.*, **43**, 429–467.

Royston, P. and Ambler, G. (1998) Multivariable fractional polynomials. *Stata Tech. Bull.*, **43**, 24–32.

Sauerbrei, W. (1992) Variablenselektion in Regressionsmodellen unter besonderer Berücksichtigung medizinischer Fragestellungen (Variable selection in regression models with special reference to application in medical research). *PhD Dissertation*. University of Dortmund, Dortmund.

——(1999) The use of resampling methods to simplify regression models in medical statistics. *Appl. Statist.*, **48**, in the press.

Sauerbrei, W., Madjar, H. and Prömpeler, H. J. (1998) Differentiation of benign and malignant breast tumors by logistic regression and a classification tree using Doppler flow signals. *Meth. Inform. Med.*, **37**, 226–234.

Sauerbrei, W. and Schumacher, M. (1992) A bootstrap resampling procedure for model building: application to the Cox regression model. *Statist. Med.*, **11**, 2093–2109.

Schmoor, C., Olschewski, M. and Schumacher, M. (1996) Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Statist. Med.*, **15**, 263–271.

Schumacher, M., Bastert, G., Bojar, H., Hübner, K., Olschweski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Newmann, R. L. A. and Rauschecker, H. F. (1994) Randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *J. Clin. Oncol.*, **12**, 2086–2093.

Simon, R. and Altman, D. G. (1994) Statistical aspects of prognostic factor studies in oncology. *Br. J. Cancer*, **69**, 979–985.

StataCorp (1996) *Stata Reference Manual, Version 5.0*. College Station: StataPress.

Teräsvirta, T. and Mellin, I. (1986) Model selection criteria and model selection tests in regression models. *Scand. J. Statist.*, **13**, 159–171.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Verweij, P. J. M. and van Houwelingen, J. C. (1993) Cross-validation in survival analysis. *Statist. Med.*, **12**, 2305–2315.

Wingerd, J. (1970) The relation of growth from birth to two years to sex, parental size and other factors, using Rao's method of the transformed time scale. *Hum. Biol.*, **42**, 105–131.