STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:5512–5528
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3148

Selection of important variables and determination of functional form for continuous predictors in multivariable model building

Willi Sauerbrei[1,*,†], Patrick Royston[2] and Harald Binder[1]

[1]*Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Stefan-Meier-Strasse 26, 79104 Freiburg, Germany*
[2]*MRC, Clinical Trials Unit, 222 Euston Road, London NW1 2DA, U.K.*

SUMMARY

In developing regression models, data analysts are often faced with many predictor variables that may influence an outcome variable. After more than half a century of research, the 'best' way of selecting a multivariable model is still unresolved. It is generally agreed that subject matter knowledge, when available, should guide model building. However, such knowledge is often limited, and data-dependent model building is required. We limit the scope of the modelling exercise to selecting important predictors and choosing interpretable and transportable functions for continuous predictors. Assuming linear functions, stepwise selection and all-subset strategies are discussed; the key tuning parameters are the nominal P-value for testing a variable for inclusion and the penalty for model complexity, respectively. We argue that stepwise procedures perform better than a literature-based assessment would suggest.

Concerning selection of functional form for continuous predictors, the principal competitors are fractional polynomial functions and various types of spline techniques. We note that a rigorous selection strategy known as multivariable fractional polynomials (MFP) has been developed. No spline-based procedure for simultaneously selecting variables and functional forms has found wide acceptance. Results of FP and spline modelling are compared in two data sets. It is shown that spline modelling, while extremely flexible, can generate fitted curves with uninterpretable 'wiggles', particularly when automatic methods for choosing the smoothness are employed. We give general recommendations to practitioners for carrying out variable and function selection. While acknowledging that further research is needed, we argue why MFP is our preferred approach for multivariable model building with continuous covariates. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: regression models; variable selection; functional form; fractional polynomials; splines

1. INTRODUCTION

In developing regression models, data analysts are often faced with many predictor variables that may influence an outcome variable. In the following $x=(x_1,\ldots,x_k)$ denotes the vector of predictor variables under consideration and $g(\mathbf{x})=(\beta_1 x_1+\cdots+\beta_k x_k)$ a linear function of them. A normal

*Correspondence to: Willi Sauerbrei, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Stefan-Meier-Strasse 26, 79104 Freiburg, Germany.
†E-mail: wfs@imbi.uni-freiburg.de

Received 11 October 2007
Accepted 11 October 2007

errors multiple regression model is given by $y|\mathbf{x}=\beta_0+g(\mathbf{x})+\varepsilon$, with $\varepsilon\sim N(0,\sigma^2)$. For a binary outcome the logistic regression model, $\text{logit}\Pr(y=1|\mathbf{x})=\beta_0+g(\mathbf{x})$ is used, whereas for the Cox proportional hazards model the effect of predictors is modelled through the hazard function $\lambda(t|\mathbf{x})=\lambda_0(t)\exp[g(\mathbf{x})]$. Considering also more general functions $g(\mathbf{x})$, e.g. with non-linear components for some variables x_j, we will discuss selection of variables and of the functional form for continuous variables in these three types of models.

Two main aims should be distinguished when creating a model. The first is prediction, with little consideration of the model structure; the second is explanation, where we try to identify influential predictors and gain insight into the relationship between the predictors and the outcome through the model structure. In prediction, model fit and mean square prediction error are the main criteria for model adequacy. The area of application we have in mind is clinical epidemiology, where studies are done to investigate whether particular variables are prognostically or diagnostically important, or are associated with an increased risk. For continuous predictors, the shape of the function is often of interest, e.g. whether there is an increasing trend or a plateau at high values of x. Because disease causation is invariably multifactorial, such assessments must be done in a multivariable context. In reality, many variables may be considered as potential predictors, but only a few will have a relevant effect. The task is to identify them. Often, generalizability, transportability and practical usefulness are important components of a good model and should be kept in mind when developing a model. Consider, for example, a prognostic model comprising many variables. All constituent variables would have to be measured in an identical or at least in a similar way, even when their effects were very small. Such a model is impractical, therefore 'not clinically useful' and likely to be 'quickly forgotten' [1]. In reality, a model satisfying the second aim, although not providing an optimal predictor in the sense of minimizing mean square error (MSE) or similar criteria, will typically have only slightly inferior performance. A model fitting the current data set 'too well' may not reflect the underlying relationships adequately, and so not satisfy the second aim.

The distinction between prediction and explanation was emphasized by Copas [2], who noted that the underlying loss functions are different and stated that a good predictor 'may include variables which are not significant, exclude others which are, and may involve coefficients which are systematically biased'. Such a predictor would clearly fail to satisfy the explanatory aim of many studies. Apart from these general considerations, no clear guidance on how to develop a multivariable model appears to be available. The main aim of this paper is to discuss options and provide guidance to practitioners who develop models.

If the number of independent variables is large, a parsimonious model is preferred, i.e. a subset of 'important' predictors whose regression coefficients β_j differ from 0. For this task, sequential strategies such as forward selection (FS), stepwise selection (StS) or backward elimination (BE) procedures or all-subset selection with different optimization criteria, such as Mallows' C_p, the information criteria of Akaike (AIC) or the Bayesian information criteria (BIC) [3], are often used in practical applications. Despite their importance and the enormous attention paid to them in the statistical literature, the properties of such strategies are not well understood. Comparisons are usually made on particular data sets or in simulation studies. The studies tend to concentrate on prediction and use the MSE for comparing strategies. The influence of individual variables and whether they are included in a model are often not considered.

Differences in results from various strategies and weaknesses of particular strategies are often emphasized in the literature. All too often these studies work with small data sets. Anomalous

DOI: 10.1002/sim

results reported in such studies may be why StS, BE and the like are often criticized. Nevertheless, they are widely used in practice. For example, in a large randomized trial (CHARM) with 7599 patients suffering from chronic heart failure, predictors of mortality and morbidity were chosen by using FS with $P < 0.01$ as the inclusion criterion [4]. Furthermore, small sample size has been used as an argument for modifying some of the criteria, e.g. changing AIC to AIC-c [3]. Such minor variations are confusing for practitioners and of limited value. More relevant for differences in results are the use of different nominal significance levels in selection strategies, an aspect often ignored. For example, the all-subsets strategy using AIC as the criterion includes more variables than BE with a 0.05 significance level. After about half a century of research on and use of variable selection methods, a chasm still remains between methodologically orientated papers and the use of strategies in actual medical studies.

Multivariable model building is even more difficult and controversial if continuous variables such as age, systolic blood pressure or (in cancer) tumour size are candidates for inclusion. What functional form should such variables assume in a multivariable model? Most often linearity is assumed, but non-linearity in the relationship with the response may prevail, and if allowed for, could substantially improve the fit. To clarify: in this paper, a 'linear model' for a predictor x assumes a risk score or linear predictor of the form $x\beta$, and similarly a non-linear model denotes non-linearity in x in the risk score. For example, we regard Cox models of the form $\lambda(t|x) = \lambda_0(t)\exp(\beta x)$ and $\lambda(t|x) = \lambda_0(t)\exp(\beta\sqrt{x})$ as, respectively, linear and non-linear in x.

To cope with continuous variables, the range is often divided into about four to five groups at 'suitable' cutpoints and corresponding dummy variables are used in models. In clinical studies, it is more usual to create just two groups. The use of cutpoints is generally problematic. The resulting step function is a poor approximation to the true relationship, and almost always fits the data much less well than a suitable continuous function. The practice of creating the so-called 'optimal' cutpoints, which maximize a measure of fit such as model χ^2, is particularly hazardous, since effect estimates are biased, P-values are much too small and the cutpoints are most unlikely to be reproducible in new studies [5].

The use of splines in regression models has been promoted as a method of modelling continuous covariates. There are several approaches for fitting generalized additive models [6] based on splines, see, e.g. Wood [7]. For combining fitting of smooth functions with variable selection, StS of degrees of freedom for each covariate is typically used in conjunction with a back-fitting approach [6]. When optimization in smoothing-parameter space is performed, 'infinite' smoothness leads to linear or constant terms (see, e.g. References [7, 8]). The latter is equivalent to excluding the respective covariates. All these options allow for considerable flexibility, but just as spline-based function estimation is susceptible to erroneous fitting of local noise, different approaches may lead to contradictory results.

As an alternative, fractional polynomials (FP) [9] have been proposed to model possible non-linearity in the relationship with the outcome. Multivariable FP (MFP) modelling [10] is an extension that combines the selection of FP functions for several continuous variables with BE of uninfluential variables. The aim is that an MFP model should fit the data well, and also be simple, interpretable and transportable [11].

Model selection in general remains a major challenge. Nowadays, it is straightforward with modern computing machinery and statistical sophistication to fit almost any given model to data, but finding a good model is difficult. There is agreement that subject matter knowledge should guide model building, but often this is limited or fragile and data-dependent model building is

DOI: 10.1002/sim

necessary [12]. For such cases, at least rough guidance is required for practical analyses in 'common' situations. Chatfield [13] discusses several important issues in model building. In the spirit of his introduction 'After about 35 years in statistics, it may be worthwhile to set down my approach to the subject, both to encourage and guide practitioners and also to counterbalance a literature that can be overly concerned with theoretical matters far removed from the day-to-day concerns of many working statisticians', we discuss in this paper selection of variables and functional forms.

Although restricting model building to the two issues, the topic is huge and the literature is controversial. Of necessity, we will present a personal view of a particularly complex problem, and thereby inevitably ignore much of the literature. Nevertheless, we believe that our approach will help to guide model building in many cases. We will discuss key issues of selection strategies. For continuous variables, we will emphasize the need for a systematic search for possible non-linear effects. Aiming to develop interpretable and practical models, we will put forward arguments that MFP is often a useful approach for multivariable model building with continuous covariates. Several issues will be illustrated in real data sets. In Section 2 we will summarize relevant issues when building a regression model. In Section 3 we discuss procedures for selection of variables and the functional form for continuous variables. Section 4 discusses key issues with the main topics in three examples. Sections 5 contains practical recommendations. Section 6 is a discussion.

An extended version of this paper with additional references is available from the first author as a Technical Report [14]. A monograph [15] on the topic will be published in 2008.

2. ISSUES IN BUILDING REGRESSION MODELS

Model building starts with many implicit or explicit assumptions and decisions, including how to deal with missing values and which model class to choose. In our experience, the normal errors, logistic and Cox proportional hazards models are the most important regression models, at least in clinical epidemiology. Most of the issues apply to other classes of models, such as parametric survival models or other generalized linear models, but none will be considered specifically. We do not consider several other scenarios, including hierarchical models, Bayesian approaches, robust statistical methods, model averaging or machine learning methods. We also ignore model building for microarray data in which the number of predictors exceeds the number of observations. These (and additional) extensions must be considered in a framework wider than is possible here. Even with the restriction on just these three types of regression models, many important issues of model building remain. The most important issues from our point of view are listed in Table I.

For an informed discussion of our two main issues of interest, variable selection and determination of the functional form for continuous variables, we also need some assumptions about the other issues. Some of them are restrictive, e.g. no missing data and no interaction, others are less strict, e.g. the number of variables may be 40, or 9 observations per variable are acceptable. The strict assumptions simplify the assessment of the main questions, but do not seriously limit the scope of the assessment and recommendations (see Section 5).

The assumption about sample size is considered in the third paragraph in the Discussion. A smaller sample size may be sufficient if interest lies only in deriving a predictor.

DOI: 10.1002/sim

Table I. Issues in building regression models when the aim is to identify influential variables and to determine the functional form for continuous variables.

Issue	Assumption in the paper (unless stated otherwise)	Reason for the assumption
Subject matter knowledge	No knowledge	Subject matter knowledge should always be incorporated in the model building process or should even guide an analysis. However, often it is limited or non-existent, and data-dependent model building is required. This situation will be considered here
Number of variables	About 5–30	With a smaller number of variables selection may not be required. With many more variables (e.g. high-dimensional data) the approaches may no longer be feasible or will require (substantial) modifications
Correlation structure	Correlations are not 'very' strong (e.g. correlation coefficient below 0.7)	Stronger correlations often appear in fields such as econometrics, less commonly in medicine. For large correlations, non-statistical criteria may be used to select a variable. Alternatively, a 'representative', e.g. a linear combination of the correlated variables, may be chosen
Sample size	At least 10 observations per variable	With a (much) smaller sample size, selection bias and model instability become major issues. An otherwise satisfactory approach to variable and/or function selection may fail, or may require extension (e.g. shrinkage to correct for selection bias)
Completeness of data	No missing data	Particularly with multivariable data, missing covariate data introduce many additional problems. Not to be considered here
Variable selection procedure	Only sequential and all-subsets selection strategies are considered	Stepwise and all-subsets procedures are the main types used in practice. Backward elimination and an appropriate choice of significance level gives results similar to all-subsets selection
Functional form of continuous covariates	Full information from the covariate is used	Categorizing continuous variables should be avoided. A linear function is often justifiable, but sometimes may not fit the data. Check with FPs or splines whether non-linear functions markedly improve the fit
Interaction between covariates	No interactions	Investigation of interactions complicates multivariable model building. Investigation of interactions should take subject matter knowledge into account

DOI: 10.1002/sim

3. PROCEDURES FOR MODEL BUILDING

Basic modelling assumptions, such as proportional hazards in the Cox model, will be taken for granted. As stated in Table I, we will assume that subject matter knowledge is not available and data-driven methods are required to derive a model.

3.1. Variable inclusion/exclusion as the only issue

In this section, a linear effect will be assumed for all continuous variables, perhaps following a preliminary transformation such as $\log x$. Model selection then reduces to a decision to include or exclude each variable. For k candidate variables, there are 2^k possible models.

Many procedures for selecting variables have been proposed. Often they do not lead to the same solution when applied to the same problem. There seems to be no consensus among modellers as to the advantages and disadvantages of the various procedures. All procedures are criticized, but for different reasons. In practical terms, however, either the full model (all available predictors) must be fitted or one of the strategies for selecting variables must be adopted.

There are two main types of strategies for variable selection. First, sequential strategies, such as FS, StS or BE procedures, are based on a sequence of tests of whether a given variable should be added to the current model or removed from it. A nominal significance level α for each of these tests is chosen in advance and largely determines how many variables will end up in the model. Some computer programs use a stepwise strategy and combine it with AIC or BIC as selection criterion. Mantel [16] gives good arguments for advantageous properties of BE in comparison with StS. In the following, BE(α) and StS(α) denote variable selection with StS or BE and a pre-specified significance level α for inclusion or elimination. Often an excluded variable will be permitted to re-enter or an already included one may be dropped. In the second type, all-subsets strategies, all 2^k possible models are fitted and the best model is chosen by optimizing an information criterion derived from the likelihood function. AIC and BIC are often used. They compare models based on goodness of fit penalized by the complexity of the model. In the normal errors model the residual sum of squares is taken as a measure of goodness of fit and the penalty term is the number of variables in the model multiplied by a constant. The penalty constant of $\log n$ for BIC is larger than that of 2 for AIC, generally resulting in models with a smaller number of predictors. See e.g. Reference [3] for modifications of AIC and BIC relevant to smaller sample sizes.

Because of mathematical intractability, the true type I error probability of stepwise procedures is unknown. However, Sauerbrei [17] concluded from simulation studies that, at least in the normal errors model, the true type I errors of stepwise procedures are only slightly higher than the nominal α. The importance of the value chosen for α is often ignored in textbooks. Arguing from asymptotic and simulation results on the significance level for the all-subsets criterion and simulation results for stepwise approaches, Sauerbrei [18] stated that BE (0.157) may be used as a proxy for all-subsets procedures with Mallows' C_p or AIC. The latter methods have an asymptotic significance level of 0.157 for the inclusion of one additional variable.

An important advantage of stepwise procedures is the ability to choose α in line with the main aim of the study, e.g. $\alpha=0.20$ when selecting confounders in an epidemiological study or $\alpha=0.01$ for variables in a diagnostic study [18]. By applying the procedures to different data sets, we will illustrate below that variation in α is the main driver for selecting different models. If the same α or an α corresponding to the asymptotic significance level of AIC or BIC is used, the resulting

DOI: 10.1002/sim

models will be similar or identical to those derived using all-subsets methods with AIC or BIC as the stopping criterion.

3.2. Functional form for continuous variables

3.2.1. Traditional methods. An important issue is how to deal with non-linearity in the relationship between the outcome variable and a continuous predictor. Traditionally, such predictors are entered into the full model or in selection procedures as linear terms or as dummy variables obtained after grouping. If the assumption of linearity is incorrect, a misspecified model will be obtained. An influential variable may not be included or the assumed functional form will differ substantially from the unknown 'truth'. Categorization introduces problems of defining cutpoint(s), over-parametrization and loss of efficiency [5]. In any case, a cutpoint model is unrealistic for describing a possibly smooth relationship between a predictor and an outcome variable. An alternative approach is to keep the variable continuous and allow some form of non-linearity. Quadratic or less frequently cubic polynomials have been used in the past, but the range of curve shapes afforded by polynomials is limited [9]. Two classes of flexible functions are more popular nowadays.

3.2.2. Splines. When discussing spline techniques, one must distinguish between techniques for fitting functions of a single variable and those that allow for multivariate models. For the former, the general principle is that the domain of a covariate is covered by knots and local polynomial pieces are anchored at these knots. The many available techniques differ in the number of knots used, the approach for determining knot positions and in the way the parameters for the polynomial pieces are estimated. Smoothing splines, as used, e.g. in [6], essentially place one knot at each unique covariate value and use a roughness penalty for parameter estimation. With regression splines, only a small number of knots is used, with the advantage that parameter estimation can be performed by standard regression methods. One prominent approach (illustrated, e.g. in Reference [12]) is based on restricted cubic splines, which ensure linearity in the tails of the distribution of x, thus avoiding unrealistic 'end effects' of the fitted functions. Alternatively, regression splines can be extended by using a large number of knots (e.g. equally spaced) combined with penalized parameter estimation [7], to render the exact knot position less important. Such procedures are closely related to smoothing splines and share the advantage of the latter that only a single smoothing parameter must be selected per covariate.

When splines are applied in multivariable models, such as generalized additive models [6], additive effects of covariates are typically assumed. For estimation, classical back fitting [6], treatment as a mixed model [19] or simultaneous estimation of all smooth functions with optimization in smoothing-parameter space [7, 8] can be used. We will use the latter in the following.

3.2.3. Fractional polynomials. FP functions are a flexible family of parametric models [9]. Here, one, two or more power transformations of the form x^p are fitted, the exponent(s) p being chosen from a small, preselected set $S=\{-2,-1,-0.5,0,0.5,1,2,3\}$ where x^0 denotes $\log x$. An FP function with two terms (FP2) is a model $\beta_1 x^{p_1}+\beta_2 x^{p_2}$ with exponents p_1 and p_2. For $p_1=p_2=p$ ('repeated powers'), FP2 is defined as $\beta_1 x^p+\beta_2 x^p \log x$. This gives eight FP1 functions (including linear) and 36 FP2 functions. FP functions with m terms encompass conventional polynomials of degree m as a special case. More than two terms are rarely required in practical applications. Sauerbrei and Royston [10] and Ambler and Royston [20] describe in detail the procedure for selecting FP functions. For multivariable model building MFP was proposed. It combines variable

DOI: 10.1002/sim

selection by BE with determination of functional form for continuous predictors by the FP function selection procedure [10, 11].

3.3. Practical aspects

Methods for variable selection are usually developed in the normal errors model and transferred by analogy to more general models, such as generalized linear models or models for censored survival data. No consequent difficulties arise with the use of stepwise procedures. For some of the all-subsets procedures, problems may arise with modelling censored survival data. When selecting variables and modelling the functional form with the MFP procedure (see Section 3.2.3), computer programs for the situations mentioned are available [21]. While a generalized linear model framework is available for most of the (additive) multivariable spline-based techniques (see Section 3.2.2), the censored survival case is more problematic. In particular, estimation by maximizing a penalized likelihood function (e.g. [7]) requires special adaptation and thus is not easily transferable. In this respect, test-based sequential procedures have an advantage.

4. EXAMPLES

Initially assuming linearity, we will compare several variable selection strategies in a data set on brain cancer (glioma). We will illustrate the importance of the significance level α in stepwise strategies. A study on oral cancer will be used to discuss determination of the functional form for one predictor. In a breast cancer study (German Breast Cancer Study Group (GBSG)), we will consider the simultaneous selection of variables and functional form when some covariates are continuous.

4.1. Comparing selection strategies

Two chemotherapy regimens were compared in a randomized trial of patients with malignant glioma. We will consider the prognostic value of 15 variables in a complete case analysis on 411 patients (274 deaths). All predictors other than age (x_5), for which a linear effect is assumed, are binary. For further details of the study and a comparison of several methods for investigating the effects of prognostic factors, see Reference [22].

Table II gives the models selected by BE(α) and StS(α) at significance levels $\alpha = 0.01$, 0.05, 0.10, 0.157, and from the all-subsets procedure using the AIC criterion.

All models include the four 'strong' predictors x_3, x_5, x_6, x_8. We will call this model M_B. If α is increased, BE(α) adds further variables to M_B. Except at $\alpha = 0.10$, StS(α) selects the same variables as BE(α). The StS(0.10) model seems slightly implausible since the additional variable x_1 is included in no other model and is eliminated in the StS(0.157) model. Apart from one minor difference (x_{13} being included instead of x_{14}), the AIC model is the same as the BE(0.157) and StS(0.157) models.

4.2. Functional form for one variable

A total of 1065 incident cases of primary oral cancer were enrolled in a large population-based case–control study of the US National Cancer Institute. Controls were frequency matched to cases by age, sex and race. As in Rosenberg *et al.* [23], the analysis presented here is restricted

DOI: 10.1002/sim

Table II. Glioma example.

Procedure	Significance level	Model selected
BE	0.01	M_B
StS	0.01	M_B
BE	0.05	M_B+x_{12}
StS	0.05	M_B+x_{12}
BE	0.10	$M_B+x_{12}+x_4+x_{11}+x_{14}$
StS	0.10	$M_B+x_{12}+x_1$
BE	0.157	$M_B+x_{12}+x_4+x_{11}+x_{14}+x_9$
StS	0.157	$M_B+x_{12}+x_4+x_{11}+x_{14}+x_9$
AIC	—	$M_B+x_{12}+x_4+x_{11}+x_9+x_{13}$

Variables selected with AIC and with backward elimination (BE) and stepwise selection (StS) at different significance levels. All models selected include the factors x_3, x_5, x_6 and x_8. The corresponding model is denoted M_B.

to the African-American arm of the study (194 cases and 203 controls). The primary exposure measure is the usual number of 1 oz ethanol-equivalent drinks consumed per week as inferred from questionnaire data. All analyses presented are in principle univariate, with confounder adjustment for sex, quartiles of age (21–48, 49–56, 57–65 and 66–80 years of age) and recent cigarette smoking (non-smoker, 1–19, 20–39 and 40+ cigarettes per day).

Figure 1 shows the estimated effect of drinks per week on the risk of developing oral cancer, according to different models. Figure 1 (left panel) shows two fitted step functions, one with two cutpoints at the tertiles (solid horizontal lines), see also Rosenberg *et al.* [23], and one with four cutpoints at the quintiles (dashed horizontal lines). Both step functions give only a very crude approximation. Using the original two cutpoints gives three groups, indicating a small jump in the risk at $x=3$ and a large jump at $x=32$. For $32<x\leqslant 140$ the risk is estimated to be constant. If four cutpoints (five groups) are used, the risk function changes markedly. It more closely resembles the fit of the FP1 (power 0.5) function selected with the FP approach (adjusted for the other covariates), as indicated by the solid curve. The step function can be seen as a rough approximation to the FP function.

Figure 1 (right panel) presents the linear (thick solid line) and cubic (dash-dotted lines) regression spline functions from the original paper. For the latter, the fits with two segments (thick line) and with eight segments (thin line) are given, which correspond to (local) minima of AIC identified by Rosenberg *et al.* [23]. The thin solid line indicates a fit with penalized regression splines (anchored at 0) with automatic selection of the penalty parameter [7]. The thick dotted line indicates a restricted cubic spline fit using four knots, corresponding to three parameters.

The fit of the linear regression spline is roughly similar to that of the FP1 function and also indicates a continuing increase of risk with large amounts of alcohol consumption. The cubic fit with two segments indicates an implausible decrease of risk for large amounts, while the penalized regression spline fit indicates a plateau. The restricted cubic spline fit features a similar shape, but the increase for small values is somewhat larger. In general, the pointwise confidence bands (not shown) of the penalized regression spline fit caution against interpretation of such minor features of the fitted functions. Only for very small amounts of consumption do the linear regression spline fit and the cubic regression spline fit with two segments lie outside these bands, indicating a sharper increase of risk compared with the penalized regression spline fit. The cubic regression spline fit with eight segments that corresponds to the local minimum identified by Rosenberg *et al.* [23],

DOI: 10.1002/sim

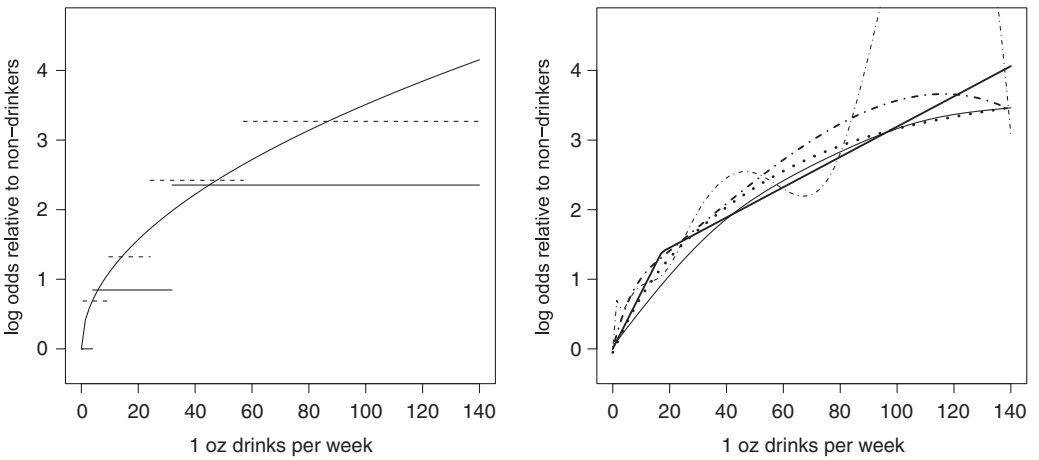

Figure 1. Oral cancer example. Estimating the influence of alcohol on the risk of developing oral cancer by different approaches. Left panel: two step-function approaches and the FP function. Right panel: five spline fits (for details see text).

with its 'wiggly' behaviour, is distinctly different from all other spline fits and obviously overfits the data.

4.3. Selection of variables and functional forms

In a study in patients with primary node positive breast cancer by the GBSG we will investigate the effect of the seven standard factors age (`age`), tumour size (`size`), number of positive lymph nodes (`nodes`), progesterone (`pgr`) and oestrogen (`er`) receptor status, menopausal status (`meno`) and tumour grade (`grade`) on recurrence-free survival (RFS) time of 686 patients (299 events) with complete data. All analyses are adjusted for hormonal treatment (`hormon`), see also [10].

Figure 2 shows the estimated functional form for age according to several methods in a multivariable model, including the treatment variable `hormon`, the binary variable `grade` (dummy variable distinguishing grade 1 from grades 2 and 3 combined) and three continuous covariates `age`, `nodes` and `pgr`. This model was derived with MFP [10]. A model with these variables will be denoted as M_V.

First, we present different approaches to assessing the functional form for age. Assuming a linear function results in a small parameter estimate, the P-value being 0.9. Graphically, the function can hardly be distinguished from the x-axis. Using the two predefined cutpoints of 45 and 60 years from the original analysis, the step function indicates a very weak effect of age with a P-value of 0.102. In the same figure we present another step function analysis, using 38 years as the only cutpoint. This cutpoint was chosen because of the result from the FP analysis with a steep increase in risk for very young patients, also presented in Figure 2 (left panel). The MFP analysis indicates a highly significant prognostic effect of age ($P<0.001$), whereas the P-value of the step function with one cutpoint is 0.019. In contrast to the step function with a jump at 38, the FP function is biologically interpretable.

Figure 2 (right panel) shows the fitted functions for age from a restricted cubic spline model (thick dotted curve) with four knots per continuous covariate and from several models fitted by

DOI: 10.1002/sim

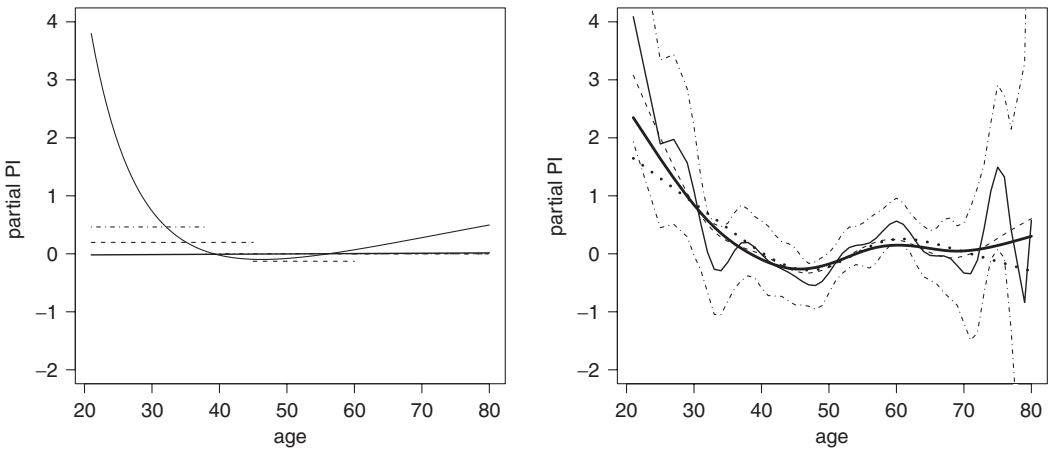

Figure 2. GBSG example. Estimation of the age function in multivariable models (see text for details). Left panel: linear function (nearly overlap with x-axes), two-step functions and the FP function. Right panel: restricted cubic spline fit, three functions derived with penalized B-splines and pointwise confidence band for the B-spline selected with AIC criterion (for details see text).

penalized B-splines. The thick solid line is from a simple model with variables corresponding to M_V, but the three continuous covariates are modelled as smooth terms with fixed degrees of freedom (df=4) for each covariate. The broken line shows the fitted function for `age` from a similar model, the only difference being that the df were chosen by optimizing the AIC. Finally, the thin line shows the fit from a more complex model (together with pointwise confidence bands—the dash-dotted lines) that additionally contains the binary variable `meno` and the two continuous predictors `size` and `er`, modelled as smooth terms, where all the degrees of freedom are chosen simultaneously by AIC.

It can be seen that with the switch from fixed df to AIC as well as with the increase in the number of predictors, additional variability is introduced that manifests as a more wiggly estimate. This example illustrates that automatic selection by using the AIC criteria might fail. The derived function is certainly not interpretable. Nevertheless, all three fits lie within the pointwise confidence bands of the more complex fit. The confidence bands suggest that a cautious interpretation of the wiggly fit is needed. Apart from a less steep function for very young age and a slight hook at about age 60, the smoothest spline function (df=4) agrees well with the FP function. Only the restricted cubic spline fit shows a different pattern, with a decrease from age 60. This behaviour is unrealistic since death from any cause was an event for RFS.

On applying the MFP procedure with a nominal significance level of $\alpha=0.05$, the continuous variables `age`, `nodes` and `pgr`, all with non-linear FP functions, and the binary variable `grade` are selected. Using BE(0.05) with linear functions or step functions (original two cutpoints for `age`) for continuous variables, the same factors except `age` are selected. However, the estimated functions are very different; see, for example, the `nodes` function in Figure 3.

Medical knowledge indicates that the effect of the strong prognostic factor `nodes` should be modelled with a non-linear function. This is supported by the results from the MFP analysis (the function from Model III in [10] is shown) and the analysis with step functions in Figure 3 (left panel). The effect of `nodes` is still highly significant when a linear function is fitted ($P<0.001$),

DOI: 10.1002/sim

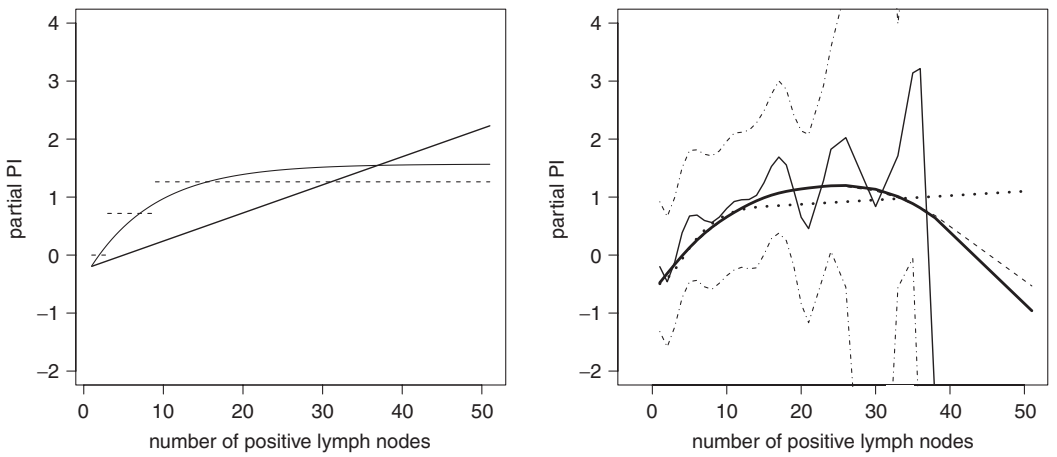

Figure 3. GBSG example. Estimates of the nodes function in multivariable models. Left panel: linear function, step function and FP function. Right panel: restricted cubic spline fit, three functions derived with penalized B-splines (for details see text), and pointwise confidence band for the B-spline selected with the AIC criterion.

but the function underestimates the risk for a low number of positive nodes and overestimates it for a large number. In Figure 3 (right panel) we give the functions from the spline approaches described above. The thick dotted curve shows a restricted cubic spline fit (using four knots per covariate). The dashed curve shows a fit with fixed degrees of freedom (df = 4), the solid curve with degrees of freedom selected by AIC and the thin solid curve (with corresponding pointwise confidence bands given by the dash-dotted curves) is a fit from a larger model where for all components the degrees of freedom have been selected by AIC. The functions indicate a decrease in risk for a large number of `nodes`, which certainly contradicts medical knowledge. This unsatisfactory type of function is partly a result of influential observations. The restricted cubic spline fit seems not to be affected by influential points, because the right-most knot is already located at the value of 16. Clear signs of overfitting are seen for the penalized B-spline fits with their large number of knots. Automatic penalty selection seems to have failed in this instance. The pointwise confidence bands caution against interpretation of minor features of the curves.

5. TOWARDS RECOMMENDATIONS

Due to limited space here, an expanded version of this section is given in [14]. Issues stated in the next three paragraphs are relevant for our recommendations.

Variable selection introduces a selection and omission bias [24]. The former occurs mainly with weakly influential variables, but is less relevant for variables with a stronger influence. Shrinkage factors may help to improve prediction and reduce selection bias. Methods combining variable selection and shrinkage, e.g. Tibshirani's LASSO [25], look promising but need improvements and further investigation of properties.

Applying a model selection procedure to a set of candidate variables produces a single model. The variables chosen may depend on the characteristics of a small number of observations. If

DOI: 10.1002/sim

the data are slightly altered, a different model may be selected. Such instability is an issue for all selection procedures. Determining the functional form of continuous variables substantially increases instability. Stability investigations should be done more often. For the breast cancer data, Royston and Sauerbrei [26] presented an analysis of models selected in bootstrap replications.

Influential points and extreme values of continuous covariates may affect the results of modelling with FP and spline functions. To counter this, Royston and Sauerbrei [27] proposed a two-stage covariate transformation to smoothly 'pull in' extreme observations and then shift the origin away from zero. Other approaches are possible.

Here we assume that subject matter knowledge is so limited that it does not influence model building. For data-dependent multivariable model building with continuous covariates, guidance is urgently needed for practitioners. Even for the simpler task, assuming a linear effect for continuous variables, literature statements conflict and are not helpful. For a detailed discussion, see [14]. From a theoretical point of view, every procedure and every single step of a model building strategy can be criticized. Nevertheless, clients expect and need an answer from the data analyst about the important variables and the functional form for continuous variables. Table III gives some recommendations.

Based on practical experience and our own research, we believe that BE is a useful method under the assumptions of Table I. For continuous variables, a linear function is often an acceptable model in many applications. However, sometime the data may contain (strong) evidence against linearity; hence, it is always necessary to check for possible non-linearity. This can best be done by fitting a 'global' function. We consider the systematic search for better-fitting FP functions to be a good approach. Sometimes subject matter knowledge may lead to modification of the function selection process [10]. Depending on the aim of a study, the nominal significance level can be chosen separately for the BE and FP components of MFP. In contrast to spline approaches, modelling with MFP is less dependent on careful 'tuning' by an expert.

Even with a smaller number of variables, say 10, of which five are continuous, several million MFP models are possible. Obviously, some principles of 'good analysis practice' are required. Before model building begins, data must be checked for extreme values or influential points. Functions of continuous predictors selected in an MFP model should be checked for important local features that cannot adequately be represented by FP models.

Data-dependent model building introduces several biases, including selection bias. Combining selection with shrinkage seems to be a useful approach. However, much more research is needed on the effect of shrinkage in this context.

Sensitivity analyses are required to check whether the model obtained depends (strongly) on questionable assumptions. Sometimes these types of analysis may reveal that the proposed model is dubious. The search for a 'good' model may have to start again, but with some assumptions modified. For example, truncating a few extreme observations may or may not affect the model that is selected [27].

Whenever data-dependent modelling is used, the analyst should be aware that the 'final' model selected is partly a result of 'chance' and can have many weaknesses. Investigation of its stability should become standard practice [18]. Often this will reveal that the selected model includes components with only a weak influence. This type of 'internal' validation may indicate that parts of the model will fail to show an influence on the outcome in new data. However, an 'external' validation is the ultimate check of a model.

DOI: 10.1002/sim

Table III. Towards recommendations for model building by selection of variables and functional forms for continuous predictors under the assumptions of Table I.

Issue	Recommendation	Explanation
Variable selection procedure	Backward elimination; significance level as key tuning parameter, choice depends on the aim of the study	BE seems to be the best of all stepwise procedures. Although BE does not optimize any specific criteria, it will select models that are at least very similar to AIC or BIC (provided the appropriate significance level is chosen). Programs for all types of regression models are available. Combining BE with other tasks, e.g. function selection procedure, is often straightforward
Functional form for continuous covariates	Linear function as the 'default', check improvement in model fit by fractional polynomials. Check derived function for undetected local features	FP functions are sufficiently flexible for most purposes. A well-defined procedure (MFP) for the multivariable situation is available. Spline models or smoothers with a local character lack stability and transportability. Disagreement about the most sensible approaches, useful only in the hands of experienced practitioners
Extreme values or influential points	Check at least univariately for outliers and influential points in continuous variables. A preliminary transformation may improve the model selected. For a proposal, see Reference [27]	Outliers or influential points often influence the functions selected, particularly at the left or right end of a covariate distribution. They may even influence the selection of a variable in a model. Many methods have been developed to identify outliers and influential points, but none seems to be generally accepted in multivariable model building
Sensitivity analysis	Important assumptions should be checked by a sensitivity analysis. Highly context dependent	In multivariable model building many assumptions are required. Sensitivity analyses are important for checking whether main results are consistent when assumptions are varied
Check of model stability	The bootstrap is a suitable approach to check for model stability	Data-dependent model building requires many decisions and models selected may depend on a small fraction of the data. Repeating the model selection process in many (say $\geqslant$1000) bootstrap replications allows an assessment of model stability. Most important variables and a reasonable functional form can be identified
Complexity of a predictor	A predictor should be 'as parsimonious as possible'	Regarding statistical criteria such as MSE, complex predictors based on many variables or including wiggly functions may be only slightly better than simple predictors including the strong variables and strong non-linearity for continuous variables. However, complex predictors are more prone to overfitting, are less stable and are less likely to be reproducible in new data (external validation). Predictors based on many variables are unlikely to be transportable to other settings (general usefulness)

DOI: 10.1002/sim

Strong agreement between 'simple' predictors based on some strong factors and more complex predictors including also weak factors, as well as experiences with checks of model stability and also with external validation, result in the recommendation that a predictor should be 'as parsimonious as possible'.

6. DISCUSSION

In observational studies with many variables and an explanatory aim, the task of a practitioner is often to build a reliable regression model that fits the data well, is plausible, simple enough to be interpretable and useful in practice. Interest centers on the effects of single variables whose parameter estimates should reflect their relative importance in an unbiased way. Decisions about which variables to include in the model are required. Procedures most often used are the 'full' model, selection procedures based on information criteria or stepwise procedures [3, 12, 18]. Advocates of the full model criticize selection procedures mainly because the parameter estimates are biased. Other criticisms are that selection procedures are unstable and that stepwise procedures do not maximize any specific statistical criterion [12]. Unbiasedness of the estimates in a prespecified model is certainly a significant advantage, but it is obtained at the high cost of a difficult interpretation of the effects if a 'large' model is chosen, and limited general usefulness if it requires many variables that are often not measured outside the research setting. Furthermore, the fit may be sub-optimal if an assumed functional form for a continuous variable is far from the actual relationship in the data.

In spite of the acknowledged problems, we consider the correct use of BE to be advantageous. It starts with the model including all candidate variables and eliminates variables that have little influence on the model fit. A predefined nominal P-value, chosen according to the aims of the study [18], is used as the stopping criterion. Because only a subset of the variables remains in the model, it is much easier to interpret and to use in a more general setting. 'Strong' predictors are selected with high probability, resulting in a fit that is not substantially inferior to that of the model with all variables. Selection bias is not a serious problem for such predictors, and predictions derived from simpler or more complex models are similar [18]. In studies with a sufficient sample size and without extremely high correlations between predictors, StS usually selects the same variables as BE, although some differences may appear for weak effects.

In view of the similarities of the results for the same nominal significance level, differences between procedures have been overemphasized in the literature. This may be a type of publication bias, reflecting apparently alarming differences between methods in special cases. Stepwise procedures have limited value as a tool for prognostic modelling in small studies. With the additional aim of determining functional forms for continuous variables, even larger sample sizes are required and the assumed minimal sample size in our paper (see Table I) sounds reasonable. In search of an effective strategy in small data sets, Steyerberg *et al.* [28] conclude, 'apply shrinkage methods in full models that include well-coded predictors that are selected based on external information'. This statement does not contradict the proposals in our paper. Steyerberg and colleagues implicitly acknowledge that in small studies data-dependent variable selection will not result in a good model. Too much attention has been paid to issues such as properties and improvement of criteria in small studies and variants to search for a better computational algorithm.

Combining BE with a systematic search for an appropriate functional form within the FP class has produced good models in many examples [10, 11]. Splines are more flexible than FPs for

DOI: 10.1002/sim

modelling continuous covariates. However, their use brings with it the danger of instability. In the examples there were some very wiggly fitted functions, certainly difficult to reproduce in new data and therefore less useful for practical applications. Multiple minima for model selection criteria can pose an additional problem [23]. These weaknesses and missing guidance on how to select variables and determine functional forms with spline approaches are important reasons that the MFP procedure, which can be used for many types of regression models, is our preferred approach. Promising results from simulation studies of FP models have emerged [20, 29]. Nevertheless, larger simulation studies are needed to gain more insight into properties of the MFP procedure and to compare them with spline-based strategies.

Issues such as model stability, transportability and practical usefulness need more attention in model development. The latter are all connected with the often neglected criterion of external validation. Increasing their importance will result in models that are built with the aim to get the big picture right instead of optimizing specific aspects and ignoring others. With a good model building procedure, the analyst should be able to detect strong factors, strong non-linearity for continuous variables, strong interactions between variables and strong non-proportionality in survival models. With such a model one is less concerned about failing to include variables with a weak effect, failing to detect weak interactions or failing to find some minor curvature in a functional form of a continuous covariate. Such a model should be interpretable, generalizable and transportable to other settings. In contrast to results from spline techniques, which are often presented as a function plot, an FP function is a simple formula allowing general usage. Our aims agree closely with the philosophy of MFP and its extensions for interactions [30] and time-varying effects [31]. Modifications that may improve the usefulness of MFP are the combination with shrinkage and a more systematic check for overlooked local curvature.

ACKNOWLEDGEMENTS

We thank Philip S. Rosenberg and colleagues for sharing the data from the National Cancer Institute's Oral Cancer Study. We thank Christoph Minder, the associate editor and five anonymous reviewers for critical comments that helped us to improve the manuscript. We are grateful to Lena Barth and Karina Gitina for their assistance in preparing the manuscript and to the RiP Program at the Mathematisches Forschungsinstitut, Oberwolfach, Germany. A substantial portion of the work was carried out during a visit to Oberwolfach in October 2006.

REFERENCES

1. Wyatt JC, Altman DG. Prognostic models: clinically useful or quickly forgotten? *British Medical Journal* 1995; **311**:1539–1541.
2. Copas JB. Regression, prediction and shrinkage (with Discussion). *Journal of the Royal Statistical Society*, *Series B* 1983; **45**:311–354.
3. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference* (2nd edn). Springer: New York, 2002.
4. Pocock SJ, Wang D, Pfeffer MA, Yusuf S, McMurray JJV, Swedberg KB, Östergren J, Michelson EL, Pieper KS, Granger CB on behalf of the CHARM investigators. Predictors of mortality and morbidity in patients with chronic heart failure. *European Heart Journal* 2006; **27**:65–75.
5. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* 2006; **25**:127–141.
6. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman & Hall: New York, 1990.
7. Wood SN. *Generalized Additive Models. An Introduction with R*. Chapman & Hall/CRC: Boca Raton, FL, 2006.
8. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Statistical Science* 1996; **11**:89–121.

DOI: 10.1002/sim

9. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with Discussion). *Applied Statistics* 1994; **43**(3):429–467.
10. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society*, *Series A* 1999; **162**:71–94. Corrigendum: 2002; **165**:399–400.
11. Royston P, Sauerbrei W. Building multivariable regression models with continuous covariates in clinical epidemiology, with an emphasis on fractional polynomials. *Methods of Information in Medicine* 2005; **44**:561–571.
12. Harrell FE. *Regression Modeling Strategies*, *with Applications to Linear Models*, *Logistic Regression*, *and Survival Analysis*. Springer: New York, 2001.
13. Chatfield C. Confessions of a pragmatic statistician. *The Statistician* 2002; **51**:1–20.
14. Sauerbrei W, Royston P, Binder H. Variable selection for multivariable model building, with an emphasis on functional form for continuous covariates. *FDM-Preprint 98*, University of Freiburg, 2007.
15. Royston P, Sauerbrei W. *Multivariable Model-Building*: *A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Continuous Variables*. Wiley: New York, 2008.
16. Mantel N. Why stepdown procedures in variable selection? *Technometrics* 1970; **12**:621–625.
17. Sauerbrei W. Comparison of variable selection procedures in regression models—a simulation study and practical examples. In *Europäische Perspectiven der Medizinischen Informatik*, *Biometrie und Epidemiologie*, Michaelis J, Hommel G, Wellek S (eds). MMV, Medizin-Verlag: München, 1993; 108–113.
18. Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *Applied Statistics* 1999; **48**:313–329.
19. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press: Cambridge, 2003.
20. Ambler G, Royston P. Fractional polynomial model selection procedures: investigation of Type I error rate. *Journal of Statistical Computation and Simulation* 2001; **69**:89–108.
21. Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs. *Computational Statistics and Data Analysis* 2006; **50**:3464–3485.
22. Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine* 1992; **11**:2093–2109.
23. Rosenberg PS, Katki H, Swanson CA, Brown LM, Wacholder S, Hoover RN. Quantifying epidemiologic risk factors using nonparametric regression: model selection remains the greatest challenge. *Statistics in Medicine* 2003; **22**:3369–3381.
24. Copas JB, Long T. Estimating the residual variance in orthogonal regression with variable selection. *The Statistician* 1991; **40**:51–59.
25. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *Series B* 1996; **58**:267–288.
26. Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Statistics in Medicine* 2003; **22**:639–659.
27. Royston P, Sauerbrei W. Improving the robustness of fractional polynomial models by preliminary covariate transformation. *Computational Statistics and Data Analysis* 2007; **51**:4240–4253.
28. Steyerberg E, Eijkemans M, Harrell F, Habbema J. Prognostic modelling with logistic regression analysis: in search of a sensible strategy in small data sets. *Medical Decision Making* 2001; **21**:45–56.
29. Holländer N, Schumacher M. Estimating the functional form of a continuous covariate's effect on survival time. *Computational Statistics and Data Analysis* 2006; **50**:1131–1151.
30. Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine* 2004; **23**:2509–2525.
31. Sauerbrei W, Royston P, Look M. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal* 2007; **49**:453–473.

DOI: 10.1002/sim