Multivariable model-building with continuous covariates: 1. Performance measures and simulation design

Harald Binder, Willi Sauerbrei & Patrick Royston

Universität Freiburg i. Br.

Nr. 105 July 2011

Zentrum für Datenanalyse und Modellbildung

Universität Freiburg

Eckerstraße 1

D–79104 Freiburg im Breisgau

und

Institut für Medizinische Biometrie und Medizinische Informatik

Universitätsklinikum Freiburg

Stefan-Meier-Straße 26

D–79104 Freiburg im Breisgau

und

MRC Clinical Trials Unit

222 Euston Road

London NW1 2DA, UK

binderh@imbi.uni-freiburg.de

Abstract

Many approaches exist for evaluating techniques for developing multivariable regression models with potentially non-linear effects of continuous covariates. In simulation studies to evaluate new proposals and to compare model-building techniques, researchers tend to consider oversimplified settings or unrealistically complex functional forms. The true shape of functions used for data generation is rarely representative of what would be expected in biomedical applications. In addition, the mean square error of prediction is often used as the main criterion for evaluation. This is insufficient when the effect of individual variables is to be assessed, e.g., in exploratory studies in clinical epidemiology. After reviewing some of the proposals for simulation designs, we suggest a new design that avoids oversimplification and tries to capture structure often found in biomedical settings. Specifically, we posit a non-trivial correlation structure between covariates, a challenge for many techniques of model selection. In addition to continuous covariates, binary covariates are included. As well as strong non-linear effects, some near-linear effects of covariates are considered. This allows one to judge whether a technique can distinguish between important non-linear effects and effects that might reasonably be represented by linear terms. We suggest several performance measures for capturing the potential impact of the various components of the simulation design on model structure. Some challenges of the simulation design are illustrated by diagnostic plots obtained after fitting a linear model, indicating the extent to which use of an under-specified model points towards potential non-linear effects.

Keywords: continuous covariates; model selection; non-linear effects; regression models; simulation.

1 Introduction

In many studies, many variables are collected and it is not clear which of them should be included in a model. Selection is an important task. Even if only fifteen variables were investigated and inclusion or exclusion was the only issue in building a regression model, $2^{15} = 32768$ possible models arise. Several methods of variable selection have been known for a long time and are often used in practice. Many of these use a stepwise approach for adding or removing covariates, guided by a cutoff for the significance level of a covariate or by information criteria such as the AIC. For a comprehensive overview, see for example References (1) or (2). Even when different models are selected, predictions from these models are often similar. Little is known about properties of the selection strategies (3). 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 covariates and gain insight into the relationship between the covariates and the outcome through the model structure. The distinction between prediction and explanation was emphasized by Copas (4), who noted that a good model "may include variables which are not significant, exclude others which are, and may involve coefficients which are systematically biased". Such a model would clearly fail to satisfy the explanatory aim of many studies ((5), pp. 26-29).

Multivariable model-building is even more difficult and controversial when continuous covariates such as age, systolic blood pressure, or (in cancer) tumour size are candidates for inclusion in a model. What functional form should such covariates assume in a multivariable model? Usually linearity is assumed, but it may describe the relationship with the outcome badly. Employing some technique that potentially allows for non-linear effects may substantially improve the fit (6). Generalized additive models (7), using a spline component for each continuous covariate (see e.g. Reference (8) for a comprehensive overview), and the multivariable fractional polynomial procedure (9; 10;

5), provide two strategies to address the common problem of model-building by selection of variables and functional forms for continuous covariates.

Despite obvious practical relevance, no systematic investigations of properties of such approaches and direct comparisons between them have been published for settings with a larger number of covariates (e.g. > 6). For such situations, theoretical results provide only limited insight (see, e.g., (11) for the impossibility of obtaining the distribution of estimators after model selection). A large simulation study is therefore currently the only way to assess the performance of different approaches. Naturally, the design and the evaluation criteria applied to summarize the results must be carefully chosen, depending on the aim of the simulation study.

Many simulation designs have serious problems, such as bias in favour of a specific approach. Often, an over-simplified design is used, having, for example, an implausibly small number of covariates, simple or even no correlation structure, and unrealistic assumptions. Only a small fraction of possible evaluation criteria may be considered. Such simulation studies do not promote deeper insight into the complex issue of selecting variables and functional forms for continuous covariates.

Here, we propose a simulation design and relevant performance measures to compare strategies for multivariable model building requiring selection of variables and of functional forms. To obtain results relevant to the analysis of real biomedical studies, we based key components of the underlying structure on a prognostic factors study. In such studies, interest often centres on the effects of individual variables, implying that explanatory models are more important than models for prediction (12). The predictive mean square error, which disregards the structure of the model, is not sufficient as a performance measure. We suggest in addition using Type I and Type II errors and other kinds of departure from the correct model.

Since the combination of variable and function selection is already challenging, we do

not consider interactions between covariates. The focus is on identifying strong main effects. In real data, interactions with respect to these main effects might be considered in a second step. We also do not consider high-dimensional data, e.g. arising from gene expression measurements in microarrays.

Naturally, no simulation design can cover all possible cases of interest. Other researchers wishing to evaluate techniques for multivariable model-building may modify our design as needed. However, retaining some of our proposed structure will enhance comparability between different studies.

In Section 2, we give an overview of simulation designs that have been proposed for evaluating multivariable model-building procedures allowing for non-linear effects of continuous covariates. We consider cases with more than three candidate covariates. Alternative criteria for model comparison are introduced in Section 3. In Section 4, the new simulation design for comparing multivariable model-building strategies is discussed in detail. A setting with 15 candidate covariates is considered, some covariates being continuous and some categorical. Aspects of the simulation design are illustrated by analyzing one data set in Section 5 assuming a linear function for each continuous variable. Specifically, we assess the extent to which the residuals point towards non-linear effects. Althoug such a poor approach as a blanket assumption of linearity may result in acceptable prediction performance, it becomes clear that the structure of the selected model does not adequately represent the real effects. Section 6 comprises final remarks. The proposed design is used in a companion paper (13) for comparing two spline approaches for multivariable modeling to the multivariable fractional polynomial (MFP) approach.

2 Some published simulation designs

In the following, some simulation designs are discussed that have been proposed for evaluating multivariable model-building approaches incorporating non-linear effects of continuous covariates. We consider only designs where all covariates are treated on an equal footing. Settings with special roles for some covariates are not discussed here, e.g., classical epidemiological settings where one covariate represents the effect of interest and the others are taken to be confounders (see Reference (14), for example). Also, we focus on designs with more than three covariates, i.e. settings where selection of covariates starts to become important. We look only at designs with additive covariate effects, i.e., no interactions are taken into account.

An early proposal for a design with a larger number of covariates is given in (15). There are $p = 10$ uncorrelated, uniformly distributed continuous covariates with range $[0, 1]$ for $n = 100$ or $n = 200$ observations. Two covariates have a non-linear effect and three have a linear effect. While the shape of the non-linear effects (shown in the left two panels of the first row of Figure 1) could be considered adequate for a biomedical setting, neither the lack of any correlation structure nor the uniform distribution are realistic.

In (16), also uncorrelated covariates are used. For $n = 1000$ observations, a main effects model with $p = 10$ continuous covariates is considered (uniformly distributed in $[0, 1]$), with two additional binary covariates. Such a mixture of continuous and binary covariates is typical of biomedical settings. Of the three continuous covariates that have an effect, one is similar to the design in (15) (top left panel of Figure 1). The effects for the other two are shown in the right two panels of the first row of Figure 1. The effect in the third panel of the top row is generated by a sine wave, which does not seem appropriate in a biomedical setting without replicated measurements. However, as only a small part of the oscillation is used in this design, the effect might still be considered

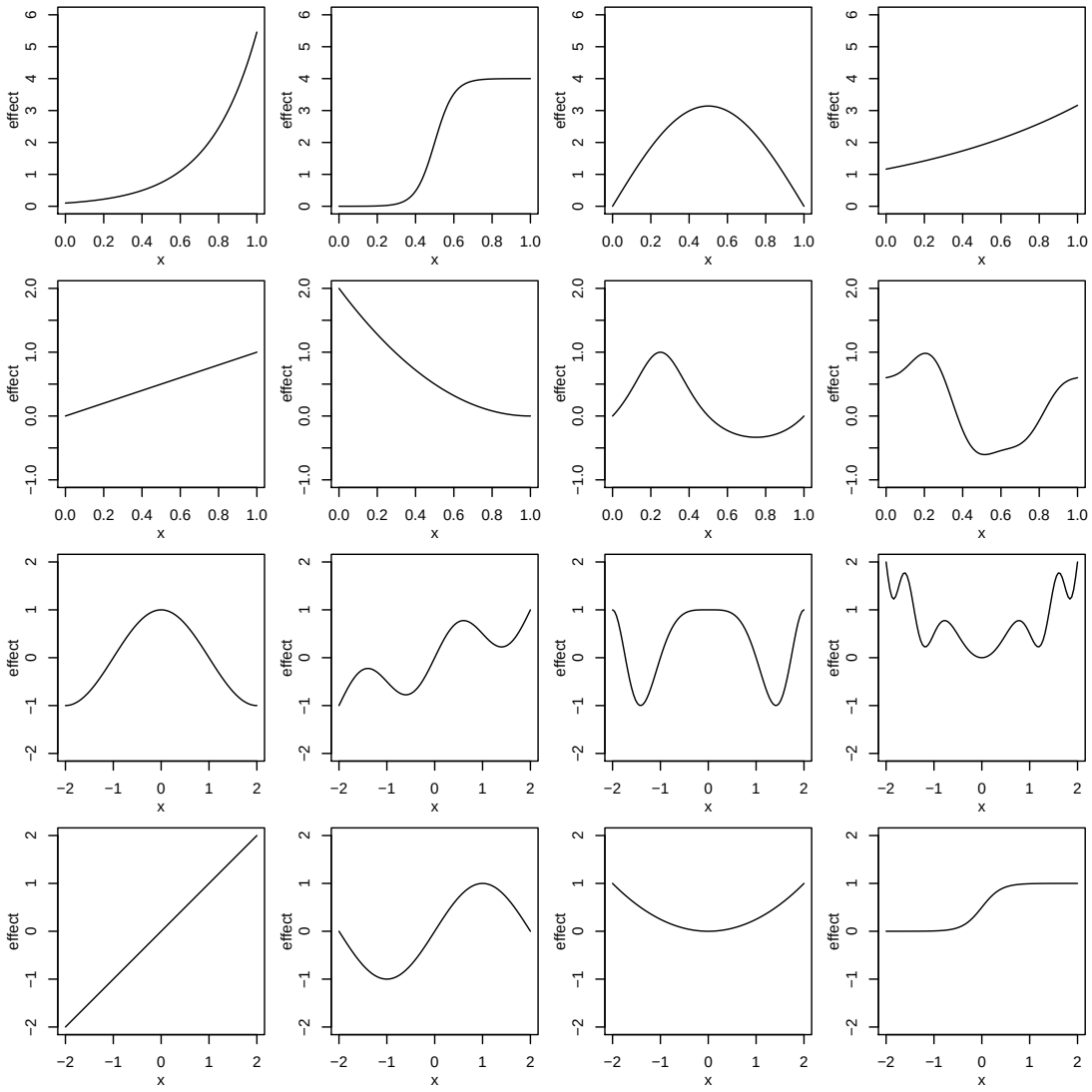

Figure 1: Non-linear functions considered for covariate effects in some simulation designs proposed in the literature (left two panels of first row: (15); right two panels of first row: (16); second row: (17); third row: (18), for $k = 1$ (left two panels) and $k = 2$ (right two panels); fourth row: (19)).

reasonable. This illustrates how sensitive some of the choices in simulation designs are to the range of the covariate values. For example, if a skewed distribution had been used in (16), some observations might have been assigned extreme covariate values that are affected by the (unrealistic) oscillations of the sine wave.

In (17), two kinds of correlation structure are proposed. In the *compound symmetry* design, each covariate is obtained via a linear combination of an independent component and a component that is common to all covariates. Both components are uniformly distributed. The weights of these components, which are the same for all covariates, allow one to vary this design from an independent uniform design, as used in (16), to one with highly correlated covariates. In the *trimmed* design, an AR(1) type of structure is used. The covariate with index 1 is taken from a standard normal distribution. For each subsequent covariate, values are obtained via a linear combination of the previous covariate and a variable drawn from a standard normal distribution. The covariates are then trimmed in $[-2.5, 2.5]$ and scaled to $[0, 1]$. This general rule supports such a correlation structure for a widely differing number of covariates. A disadvantage is that the resulting structure is not very realistic. The functions shown in the second row of Figure 1, are considered for the shape of the covariate effects. These functions are used in a setting where four of ten covariates are informative (where $n = 100$), and also in a setting where 12 covariates with adjacent indices of $p = 60$ covariates are informative (where $n = 500$). However, for evaluating model-fitting approaches in a biomedical context, the functions shown in the third and fourth panel of the second row might be especially problematic. While the overall shape could still be expected in a medical application, the local features introduced by the trigonometric functions, which generate these shapes, are difficult to interpret.

In (18), a different kind of correlation structure is considered. For a total of $p = 18$ covariates, six blocks of three correlated covariates are considered, sample sizes are $n =$

50 and $n = 200$. The covariates are from a standard normal distribution. The correlation of two covariates with indices j and k in the same block is given by $\rho^{|j-k|}$. Covariates in different blocks are independent. This corresponds to having several independent underlying influential factors, where the facets of each factor are represented by three covariates. This is in contrast to the compound symmetry design in (17) with only one common factor. Furthermore, different settings are considered in (18), where either the first two or five of the six blocks influence the response. The shape of the effect of three covariates in a block is given governed by a parameter k. One of the effects simply is x^k, the other two are shown in the third row of Figure 1, for $k = 1$ (left two panels) and $k = 2$ (right two panels). The strongly local structure with a large number of extrema, obtained for larger values of k, would not be expected in a biomedical application. This illustrates the danger of using a parameter, such as k, for generating functional forms automatically.

Binder and Tutz (19) consider p = 6,10, 20, and 50 uncorrelated covariates from a standard normal distribution, truncated to the range $[-2, 2]$. They simulate sets of $n = 100$ observations with continuous or binary responses. The number of informative covariates is either three or six. The shape of the effect of a covariate is randomly sampled to be the centred and standardized version of one of the functions shown in the fourth row of Figure 1. Consequently, even within one simulation setting the shape of the true function varies over the repetitions.

In summary, the main problems of the simulation designs we have discussed arise in two areas: correlation structure and appropriateness of the non-linear functions for a biomedical setting. Some correlation structure is needed for a realistic evaluation. However, even automatically generated correlation structures, i.e., as specified by a general rule, such as in (17), do not match the complex structure seen in biomedical applications. Similarly, automatically generated shapes of non-linear functions, as in

(18), make it difficult to incorporate typical shapes seen in biomedical applications and might result in very unrealistic shapes (e.g., the rightmost function in the third row of Figure 1). In addition, the distribution of the covariate values should not be neglected. Uniform distributions are unrealistic, and settings with normally distributed covariates could be made more realistic by introducing some skewness.

An alternative way to specify correlation structure and functional form is to mimic the structure found in some real biomedical data. This is the idea behind the ART study (5), which is considered in what follows.

3 Criteria for evaluating selected models

Consider a true model of form

$$y_i = \beta_0 + \sum_{j \in J_{lin}} \beta_j x_{ij} + \sum_{j \in J_{nonlin}} f_j(x_{ij}) + \epsilon_i, \tag{1}$$

with response y_i, covariates $x_{ij}, j = 1, \ldots, p$, and error term $\epsilon_i \sim N(0, \sigma^2)$, i.e., there are the indices $J_{lin} \subset \{1, \ldots, p\}$ of covariates with truly linear effect (by definition also including ordinal and dummy-coded categorical covariates that have an effect), specified by the parameters β_j, the indices J_{nonlin} of covariates with truly non-linear effects, specified by the true functions f_j, and the indices $J_{noe} = \{1, \ldots, p\} \backslash (J_{lin} \cup J_{nonlin})$ of covariates with no effect.

A fitted model

$$\begin{aligned} \hat{y}_i \;=\; & \hat{\beta}_0 + \sum_{j \in \hat{J}_{lin} \cap J_{lin}} \hat{\beta}_{ij} x_{ij} + \sum_{j \in \hat{J}_{lin} \cap J_{nonlin}} \hat{\beta}_{ij} x_{ij} + \sum_{j \in \hat{J}_{lin} \cap J_{noe}} \hat{\beta}_{ij} x_{ij} \\ & + \sum_{j \in \hat{J}_{nonlin} \cap J_{nonlin}} \hat{f}_j(x_{ij}) + \sum_{j \in \hat{J}_{nonlin} \cap J_{lin}} \hat{f}_j(x_{ij}) + \sum_{j \in \hat{J}_{nonlin} \cap J_{noe}} \hat{f}_j(x_{ij}), \end{aligned} \tag{2}$$

with predicted response $\hat{y}_i, i = 1, \ldots, n$, is to be evaluated with respect to the true model (1). The fitted model is characterized by the indices $\hat{J}_{lin} \subset \{1, \ldots, p\}$ of the covariates that have been assigned a linear effect, the indices $\hat{J}_{nonlin} \subset \{1, \ldots, p\}$ of the covariates that have been assigned a non-linear effect, where $\hat{f}_j$ are the fitted functions, and the indices $\hat{J}_{noe}$ of covariates deemed to have no effect. That latter have to be determined by some kind of model selection approach. $\hat{J}_{lin}$ and $\hat{J}_{nonlin}$ typically will not be identical to J_{lin} and J_{nonlin}, respectively, i.e., the covariates $\hat{J}_{lin} \cap J_{noe}$ and $\hat{J}_{nonlin} \cap J_{noe}$ are erroneously included in the fitted model.

A straightforward criterion for comparing the true model (1) with the fitted model (2) utilizes the predictive mean square error (PMSE)

$$E[(y - \hat{y})^2], \tag{3}$$

which can easily be evaluated empirically, using a reasonably large number of newly generated observations, e.g., n_{new}=1000.

Prediction performance, quantified by the PMSE, is often taken as the sole measure for comparing fitted models. However, as a global measure, it fails to capture many aspects of model performance that may be of practical relevance, e.g., when performing exploratory data analysis, aimed at understanding the role of each covariate. For example, $\hat{J}_{lin}$ and $\hat{J}_{nonlin}$ often differ from J_{lin} and J_{nonlin}, which is only indirectly reflected in measures of square error. Generally, aspects such as the complexity of a model and of individual functions, Type I and Type II errors, and qualitative features of the fitted functions often are at least as important as prediction performance in a biomedical setting. We therefore focus on quantities aimed at capturing interpretability of the fitted models (12).

All measures discussed in the following are summarized in Table 1 and roughly catego-

Table 1: Summary of criteria for evaluating selected models and single fitted functions. Type I and type II error criteria can also be considered for the single functions (see text).

Selected model	Eqn.	Single functions	Eqn.
Predictive mean square error	(3)	variability	(4)
Type I error for covariate inclusion	(6)	PED_1	(5)
Type II error for covariate inclusion	(7)		
Type I error for fitting of non-linear effects	(8)		
Cross-tables of complexity levels			
Type II error for fitting of non-linear effects		Qualitative criteria	
Costs for erroneous inclusion/exclusion			

rized into criteria mainly used for judging selected models and those for judging individual fitted functions. However, many of the former can also be used to assess performance for a subset of the covariates or even a single covariate. Except for PMSE, all measures can be applied regardless of the response type, e.g., also for binary response models or models for time-to-event endpoints. Table 1 is by no means comprehensive, as there exist many quantities that could be considered as alternatives, e.g., bias, coverage, or the number of parameters used by a model. We subjectively chose those that we deemed the most important for assessing models in the context of selection of variables and of functional form. These measures cannot and should not replace graphical tools, such as plotting a random sample of fitted functions, but should complement them.

3.1 Continuous measures for fitted functions

The fitted functions $\hat{f}_j$ are evaluated according to continuous measures. In (20) and (21) it is suggested that functions that were fitted in bootstrap samples should be aggregated into an overall bagging (bootstrap aggregating) estimate for closer inspection. The variability of fitted functions can then be evaluated by

$$V_j = \frac{1}{n_{new}M} \sum_{i=1}^{n_{new}} \sum_{m=1}^{M} (\hat{f}_j^{(m)}(x_{ij}) - \hat{f}_j^{agg}(x_{ij}))^2, \tag{4}$$

using n_{new} observations from a test set with the same covariate distribution as the original data. The reference function $\hat{f}_j^{agg}$ is obtained from the estimated functions $\hat{f}_i^{(1)}, \ldots, \hat{f}_i^{(M)}$ in M repetitions of a simulation scenario as $\hat{f}_j^{agg}(x) = 1/M \sum_{m=1}^{M} \hat{f}_i^{(m)}(x)$, where the single fits $\hat{f}_i^{(m)}$ are centered. If a covariate has not been selected in a fitted model, the fitted function is chosen to be constant, i.e., $\hat{f}_i^{(m)} = 0$. For more sophisticated approaches to aligning different functions and obtaining average representatives, see Reference (22).

The closeness of a fitted function to the true function could be quantified by mean square difference. This measure is closely related to (3), but as indicated in (23), even very 'wiggly' fits that barely resemble a true smooth function can result in a small mean square difference.

Following (23), we suggest evaluating the mean square difference of the functions and of their first and second derivatives. Specifically, the measure

$$\text{PED}_1(\hat{f}_j, f_j) = E[(f_j'(x_{ij}) - \hat{f}_j'(x_{ij}))^2], \tag{5}$$

can be evaluated at n_{new} new observations, giving more weight to regions with more observations. A similar measure might be considered for the second derivative of the original functions. To avoid extreme values from fitted functions that tend towards $-\infty$ or ∞ in boundary areas with only few observations, these quantities are only evaluated for observations within the 5% and 95% quantiles (23).

3.2 Type I and Type II errors

When attempting to interpret a selected model, the components representing the effects of the individual covariates are important. The first question is whether a covariate is included or excluded. Depending on whether that covariate really has an effect, a Type

I or a Type II error may result. A Type I error occurs when a covariate has no effect and is nevertheless included in a model, i.e., the rate of Type I errors with respect to covariate inclusion is given by

$$\frac{|J_{noe} \cap (\hat{J}_{lin} \cup \hat{J}_{nonlin})|}{|J_{noe}|}. \tag{6}$$

Correspondingly, the simplest definition of the Type II error rate with respect to covariate inclusion is given by

$$\frac{|(J_{lin} \cup J_{nonlin}) \cap \hat{J}_{noe}|}{|J_{lin} \cup J_{nonlin}|}. \tag{7}$$

However, Type I and Type II errors with respect to the fitted shape are also important and should be considered. For a covariate with no effect or a linear effect, selecting a non-linear function can be considered a Type I error. Correspondingly, the rate of Type I errors with respect to shape is given by

$$\frac{|(J_{noe} \cup J_{lin}) \cap \hat{J}_{nonlin}|}{|J_{noe} \cup J_{lin}|}. \tag{8}$$

It follows that even if a non-linear fit looks nearly linear, a Type I error will be recorded, because a non-linear fit still results in a more complicated model equation.

We propose three ways to investigate Type II errors with respect to shape:

1. Quantify deviation via continuous measures, such as (5).

2. Cross-tabulate the true levels of complexity and fitted levels of complexity. For levels of complexity "not included", "linear", and "non-linear" this results in 3×3 tables, which could, e.g., be aggregated via ordinal variants of Cohen's kappa.

3. Define qualitative features for each of the true functions that can be checked in the fitted functions, e.g. monotonicity, local extrema in a certain region, larger slope in a certain region. These depend on the specific functions. If a fitted function $\hat{f}_j$

Table 2: Scheme for assigning costs for erroneous exclusion or inclusion of covariates. "$\text{Cor}(y, x_j)$" indicates the marginal correlation between the covariate with index j and the response.

true effect	selected model	cost
$j \in J_{lin} \cup J_{nonlin}$	$j \in \hat{J}_{lin} \cup \hat{J}_{nonlin}$	0
	$j \in \hat{J}_{noe}$	$\text{Cor}(y, x_j)$
$j \in J_{noe}$	$j \in \hat{J}_{lin} \cup \hat{J}_{nonlin}$	$\min_{l \notin J_{noe}} \text{Cor}(y, x_l) \cdot \frac{\max_{l \in J_{noe}} \text{Cor}(y,x_l) - \text{Cor}(y,\eta_j)}{\max_{l \in J_{noe}} \text{Cor}(y,x_l) - \min_{l \in J_{noe}} \text{Cor}(y,x_l)}$
	$j \in \hat{J}_{noe}$	0

does not exhibit all the characteristics of the underlying true function, a Type II error is recorded.

More detailed analyses can be performed by looking at individual criteria and the conditions under which they are satisfied.

The Type I and Type I error criteria (6), (7), and (8), and complexity cross-tabulation aggregate the errors across all covariates, while with the qualitative criteria, each covariate is conidered separately. Also, the former criteria could be considered for each covariate, but this might be too demanding for a larger number of covariates. When a simulation design incorporates a mixture of many different types of covariate effects, aggregating Type I and Type II errors provides an indication of the average performance that can be expected. In a situation where a data analyst does not know much about the true effects, this might be more useful compared with performance conditional on some specific covariate effect.

3.3 Costs for erroneous exclusion or inclusion

To combine Type I and Type II errors into a single measure, costs for erroneous inclusion/exclusion of covariates can be assigned and summed for each selected model. Different costs should be assigned depending on the role and importance of a covariate.

Table 2 illustrates the scheme proposed here.

The costs for a covariate are based on the absolute value of the marginal correlation of the corresponding true model component with the response, e.g., as determined from a test set. For an influential covariate, the costs are equal to this value. Therefore, excluding a covariate with a larger effect results in larger costs.

The definition of costs for erroneously including a covariate that has no effect is based on two ideas. First, erroneous inclusion is a less severe mistake than omitting a covariate that has an effect. Therefore, the maximal cost that can be incurred by erroneous inclusion is set to the smallest cost that can be incurred by erroneously excluding a covariate that has an effect. Second, the error of erroneously excluding a covariate and instead erroneously including a highly correlated covariate should not be punished twice in terms of costs. Therefore, a cost of zero is assigned to the covariate that has no effect on the response, but has the largest (absolute) marginal correlation. For the covariate with the smallest correlation, the cost is taken to be the minimal cost that can be incurred by erroneous exclusion, as explained above. For covariates in between, costs are allocated by the corresponding linear transformation of the correlations.

For example, consider a setting with three influential covariates with marginal correlations of 0.5, 0.4, and 0.2, respectively, and two non-influential covariates with marginal correlations of 0.3 and 0.1, respectively. A model that includes only the first of the influential covariates, and the first of the non-influential covariates, incurs a cost of 0.4+0.3=0.7. The erroneously included covariate does not increase costs, as it potentially has picked up some of the information of the erroneously excluded influential covariates, which already resulted in a cost increase. The relatively large value of the marginal correlation indicates that the non-influential covariate might be a reasonable surrogate. In contrast, if the second non-influential covariate had been included, this would have increased the costs by 0.2.

4 Simulation design

The covariate structure of the simulation design proposed in the following is based on the ART study, described in ((5), Ch. 10). The latter design is based in turn on the data from the GBSG study (24), a well-known biomedical data set that has been analyzed in several places (see (10), for example). However, we made several changes to Royston and Sauerbrei's design (5). There are five additional covariates (z_{i11} to z_{i15}), providing one additional covariate with a strong linear effect and more covariates without effect. As a further major change, we suggest using different functional forms. Specifically, we have introduced one true function with a local effect. We have also simplified the correlation structure, using only four levels for the absolute values of the underlying correlations. To adapt the design for use in large-scale simulation studies, several other minor changes have been made.

Naturally, the simulation design could be modified for future applications, depending on the particular objectives when evaluating multivariable modelling techniques. For example, different correlation structures or effect sizes might be considered. Some of the covariates might even be 'forced' into the models, reflecting scenarios where the mandatory covariate(s) are of primary interest and others are included only for the purpose of adjustment.

4.1 True model

4.1.1 Covariate structure

Fifteen underlying variables form the basis for deriving 17 covariates with different kinds of distributions (e.g., skewed continuous covariates or categorical covariates). The underlying variables are generated from a standard normal distribution $z_{ij} \sim N(0, 1), j = 1 \ldots, 15$. The partial correlations for them are indicated in Figure 2. All correlations

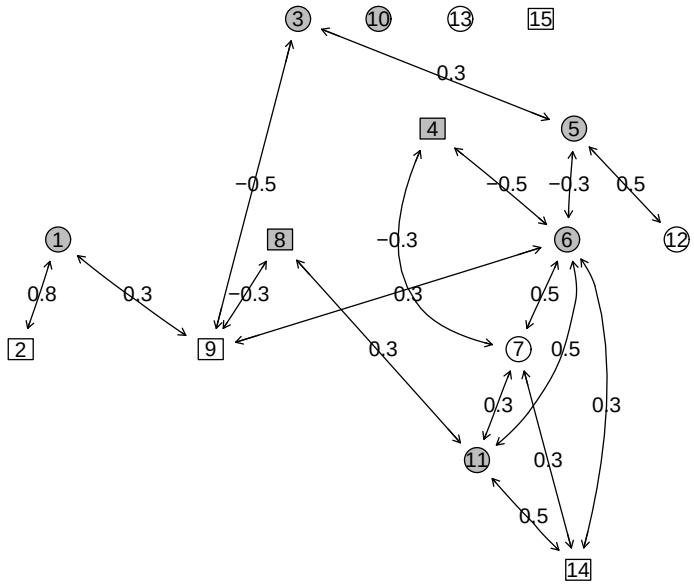

Figure 2: Partial correlations of the variables $z_{ij}, j = 1, \ldots, 15$, underlying the covariates. Variables that form the basis for continuous covariates are indicated by circles, variables that correspond to categorical covariates are indicated by rectangles. Variables corresponding to covariates that have an effect on the response are indicated by gray shading.

not explicitly indicated are equal to zero. For obtaining covariates of the type given in Column 2 of Table 3 and for introducing a skewed distribution for all continuous covariates except x_{i1}, x_{i11}, x_{i12}, and x_{i13}, the covariates are determined via the transformations given in Column 3. For avoiding extreme values for the continuous covariates, $x_{i3}, x_{i5}, x_{i6}, x_{i7}$, x_{i10}, x_{i11}, x_{i12}, and x_{i13} are truncated at the third quartile plus 5 time the interquartile range for every generated data set. The type of a covariate, i.e., continuous or categorical, is indicated in Figure 2. The continuous covariate x_{i6} is related to many other continuous and categorical covariates. This makes separation of effects on the response difficult. The strongest correlation in the proposed simulation design is between the continuous covariate x_{i1} and the binary covariate x_{i2}. This structure requires careful, systematic handling of both categorical and continuous covariates by a multivariable model-building procedure.

4.1.2 Effect on the response

The contribution of each covariate with respect to a continuous response is given in Column 5 of Table 3. Plots of the functions for the continuous covariates are shown in Figure 3. While some of the functions already have a rather complicated shape, fitting of the functions is further complicated by the distribution of the covariate values. For illustration, the rugs in Figure 3 show the empirical distribution in an example data set of size $n = 200$. This shows, for example, that the strong decrease for values of x_{i5} larger than 10 is difficult to identify because of limited support in the data. Similarly, the flattening of the effect for larger values of x_{i6} can only be identified from a small number of observations.

The fraction by which explained variation decreases, compared with the true model, when removing the respective components is given in Column 6, evaluated on a test set of size $n = 10000$. The covariates x_{i3}, x_{i5}, x_{i6}, and x_{i11} are seen to have the strongest

Table 3: Covariate structure for the multivariable simulation study. [] indicates that the non-integer part of the argument is removed, and $I()$ is the indicator function, which takes value 1 if its argument is true, and value 0 otherwise. For pairs of continuous covariates and combinations of continuous and binary variables, the marginal correlation is given (only the covariate name is indicated if absolute value < 0.3). For pairs of binary variables, the odds ratio is provided (indicated by italics). "ΔR^2" indicates the fraction by which the R^2 decreases, compared to the true model, when the respective component is removed in a test set of size n=10000 (total R^2 of the true model: 0.49).

variable	type	covariate	correlation/OR	predictor	ΔR^2
z_{i1}	continuous	$x_{i1} = [10z_{i1} + 55]$	x_{i2} (-0.6), x_{i9a}, x_{i9b}	$3.5x_{i1}^{0.5} - 0.25x_{i1}$	0.05
z_{i2}	binary	$x_{i2} = I(z_{i2} < 0.6)$	x_1 (-0.6), x_{i15} (*1.1*)	-	0
z_{i3}	continuous	$x_{i3} = \exp(0.4z_{i3} + 3)$	x_{i5} (0.3), x_{i9a}, x_{i9b}	$2 \cdot (\log(\frac{x_{i3}+10}{25}))^2$	0.13
z_{i4}	ordinal	$x_{i4_a} = I(z_{i4} \geq -1.2)$,	x_{i6} (-0.3), x_{i8} (*0.9*), x_{i9a} (*1.1*), x_{i9b} (*1.2*), x_{i7}	$-0.4x_{i4_a}$	0.02
		$x_{i4_b} = I(z_{i4} \geq 0.75)$	x_{i6} (-0.3), x_{i9b} (*1.1*), x_{i7}	-	0
z_{i5}	continuous	$x_{i5} = \exp(0.5z_{i5} + 1.5)$	x_{i3} (0.3), x_{i12} (0.5), x_{i6}	$-(0.15x_{i5} + 0.75\exp(-\frac{(log(x_{i5})-1.5)^2}{0.4}))$	0.18
z_{i6}	continuous	$x_{i6} = [\max(0, 100\exp(z_{i6}) - 20)]$	$x_{i4a}(-0.3)$, x_{i4b} (-0.3), x_{i7} (0.4), x_{i11} (0.4), x_{i5}, x_{i9a}, x_{i9b}, x_{i14}	$0.25\log(x_{i6} + 1)$	0.12
z_{i7}	continuous	$x_{i7} = [\max(0, 80\exp(z_{i7}) - 20)]$	x_{i6} (0.4), x_{i11} (0.3), x_{i4a}, x_{i4b}, x_{i14}	-	0
z_{i8}	binary	$x_{i8} = I(z_{i8} < -0.35)$	x_{i4a} (*0.9*), x_{i9a} (*2.0*), x_{i9b} (*2.4*), x_{i11}	$0.4x_{i8}$	0.03
z_{i9}	categorical	$x_{i9a} = I(0.5 \leq z_{i9} < 1.5)$,	x_{i4a} (*1.1*), x_{i8} (*2.0*), x_{i1}, x_{i3}, x_{i6}	-	0
		$x_{i9b} = I(z_{i9} \geq 1.5)$	x_{i4a} (*1.2*), x_{i4b} (*1.1*), x_{i8} (*2.4*), x_{i15} (*0.9*), x_{i1}, x_{i3}, x_{i6}	-	0
z_{i10}	continuous	$x_{i10} = 0.01[100(z_{i10} + 4)^2]$	-	$0.021x_{i10}$	0.04
z_{i11}	continuous	$x_{i11} = [10z_{i11} + 55]$	x_{i6} (0.4), x_{i7} (0.3), x_{i14} (-0.4), x_{i8}	$0.04x_{i11}$	0.19
z_{i12}	continuous	$x_{i12} = [10z_{i12} + 55]$	x_{i5} (0.5)	-	0
z_{i13}	continuous	$x_{i13} = [10z_{i13} + 55]$	-	-	0
z_{i14}	binary	$x_{i14} = I(z_{i14} < 0)$	x_{i11} (-0.4), x_{i15} (*1.1*), x_{i6}, x_{i7}	-	0
z_{i15}	binary	$x_{i15} = I(z_{i15} < 0)$	x_{i2} (*1.1*), x_{i9b} (*0.9*), x_{i14} (*1.1*)	-	0

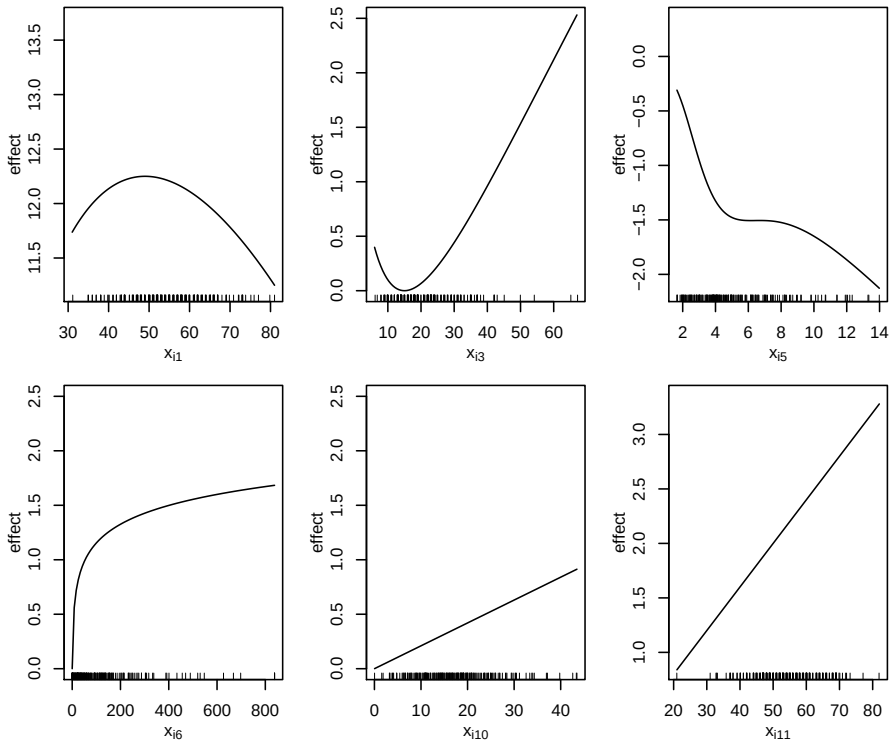

Figure 3: True functions used for the comparison for univariate and multivariable fitting problems. The empirical distribution in a data set of size $n = 200$ is indicated by rugs.

influence, while the influence of all other covariates with a non-zero effect is similar. The true functions for these covariates with a strong effect exhibit a linear effect (x_{i11}), a non-linear effect with local features (x_{i5}) or a global non-linear effect (x_{i3} and x_{i6}).

Certain pairs of covariates are of particular interest when the performance of a multivariable model-building technique is evaluated:

- x_{i1} and x_{i2} are strongly correlated, but only the former has an effect on the response. However, the shape of the effect is quadratic, i.e., it is difficult to represent it via a linear component (also compared with the other covariates with non-linear effects). A model-selection approach that avoids the complexity of non-linear effects could easily misattribute part of the effect to the simpler binary covariate x_{i2}.
- There is considerable correlation between the continuous covariates x_{i6} and x_{i11}, which both have a strong effect on the response. The former has a non-linear influence, the latter a linear effect. Due to the correlation, part of the non-linear effect might be misattributed to x_{i11}.
- For the pair x_{i5} and x_{i12}, as well as for x_{i6} and x_{i7}, the covariate with the smaller index has a strong non-linear effect on the response, while the covariate with the larger index has no effect. Therefore, part of the non-linear effect might be misattributed.

4.1.3 Sample size and effect size

There are virtually no empirically supported recommendations for the number of observations needed to provide a reasonable starting-point for selecting models with potentially non-linear effects and a given number of covariates. As a difficult setting, we suggest simulation scenarios with $n = 200$ observations. For the 17 covariates in our pro-

posed simulation design, this corresponds to roughly 12 observations per covariate. For a setting in which better identification of the true functional form could be expected, $n = 500$ should be considered. These recommendations are intended for continuous response data and need to be modified for other response types. For example, for a binary response, the number of observations in the smaller outcome group determines the amount of information. In a time-to-event setting, it is determined by the number of events.

The difficulty of a simulation scenario depends not only on the number of observations but also on the signal-to-noise ratio. For scenarios with intermediate difficulty, we suggest a signal-to-noise ratio of 1, obtained from an error variance of $\sigma^2 = 0.868$ and corresponding to an $R^2 = 0.5$ for the true model, as seen from the relation

$$R^2 = 1 - \frac{1}{\text{signal-to-noise ratio} + 1}.$$

For a more difficult setting, a signal-to-noise ratio of 0.25 ($\sigma^2 = 3.47$, $R^2 = 0.2$) could be considered, and for more informative settings, signal-to-noise ratios of 4 ($\sigma^2 = 0.217$, $R^2 = 0.80$) and 9 ($\sigma^2 = 0.096$, $R^2 = 0.90$) could be used.

In their basic form, R^2 and the signal-to-noise ratio are limited to continuous response models. However, the signal-to-noise ratio can be transferred to generalized linear and generalized additive models, i.e. for binary or counting responses (see (25), for example). For time-to-event endpoints, several adaptations of R^2 are available (see, e.g., (26) for an overview).

4.2 Type II error criteria

Several of the covariates have no effect on the response. If they are nevertheless included in a fitted model (or in the case of a continuous covariate, have a non-constant fit), a

Table 4: Features of the true functions to be satisfied by a fitted function to avoid being considered a Type II error.

covariate	monotonicity	extrema	slope	inflection point
x_{i1}	-	exactly one maximum between 45 and 55	decreasing	-
x_{i3}	strictly increasing for > 20	exactly one minimum between 12 and 20	-	not more than one, for > 20
x_{i5}	strictly decreasing for < 3 and > 9	-	absolute value for 7 smaller than for 2 and 10	exactly one, between 3 and 9
x_{i6}	strictly increasing	-	decreasing	-
x_{i10}	strictly increasing	-	constant	-
x_{i11}	strictly increasing	-	constant	-

Type I error occurs. In contrast, the continuous covariates x_{i1}, x_{i3}, x_{i5}, x_{i6}, and x_{i10} do have an effect on the response. Even if these have non-constant fits, however, the estimated shape may not adequately represent the true structure.

The qualitative criteria for judging Type II errors cannot be given in general form because they depend on the shape of the true function. We must decide how narrowly these criteria should be set. We suggest using rather narrow definitions. For example, medical understanding is little increased if a fitted function indicates only that there is an extremum (maximum or minimum) somewhere within a wide range of covariate values. Use of a narrower definition will increase the numbers of Type II errors, but at the same time will also increase the chance of identifying model-building approaches that can quite accurately detect features of the true functions.

Table 4 shows the criteria to be satisfied by the fitted functions for them to be considered adequate in the context of the proposed simulation design. Besides forming the basis for calculating Type II error rates, each individual requirement is considered in a more detailed analysis.

For covariate x_{i1}, the maximum of the true function is located in a region with many observations. We must therefore locate this extremum rather precisely. Since there is

no local curvature in the true function, we require a decreasing slope at all values. For covariate x_{i3}, the true function has an inflection point for very large values which is hardly visible and supported by very few observations. Identification of this inflection point is therefore not essential. For excluding local curvature, we must identify no more than one inflection point. For covariate x_{i5}, the true function has a small local minimum between values 6 and 8, but this is not required for the fitted functions. The only condition for the latter is that there is some indication of a smaller slope (in absolute value) at value 7 compared with values 2 and 10, which can be met more easily. For covariates x_{i10} and x_{i11} not only is a constant slope (i.e. linearity) checked, which can also be seen from 3×3 tables of complexity, but also the sign of the slope is considered.

5 Example

In the following, we illustrate the proposed simulation design by analyzing data generated from the design but assuming linear effects of all continuous covariates. Specifically, $n = 200$ observations are generated, with the error term chosen such that the signal-to-noise ratio is approximately equal to 1, i.e. the R^2 of the true model is about 0.5.

While a multivariable linear model cannot adequately represent the non-linear shapes in the proposed simulation design, it can nevertheless potentially explain some of the effects of the covariates on the outcome. In addition, the residuals from such a model should indicate whether non-linear structure was missed. Often, such a linear model, and corresponding diagnostic plots, might be the starting-point for a subsequent analysis using techniques that allow for non-linear effects.

Figure 4 shows the residuals from a multivariable linear model, obtained by backward elimination, guided by AIC. These are smoothed by the popular *loess* technique, which fits local polynomials (here: of degree 2) (27). For the covariates with a non-linear true

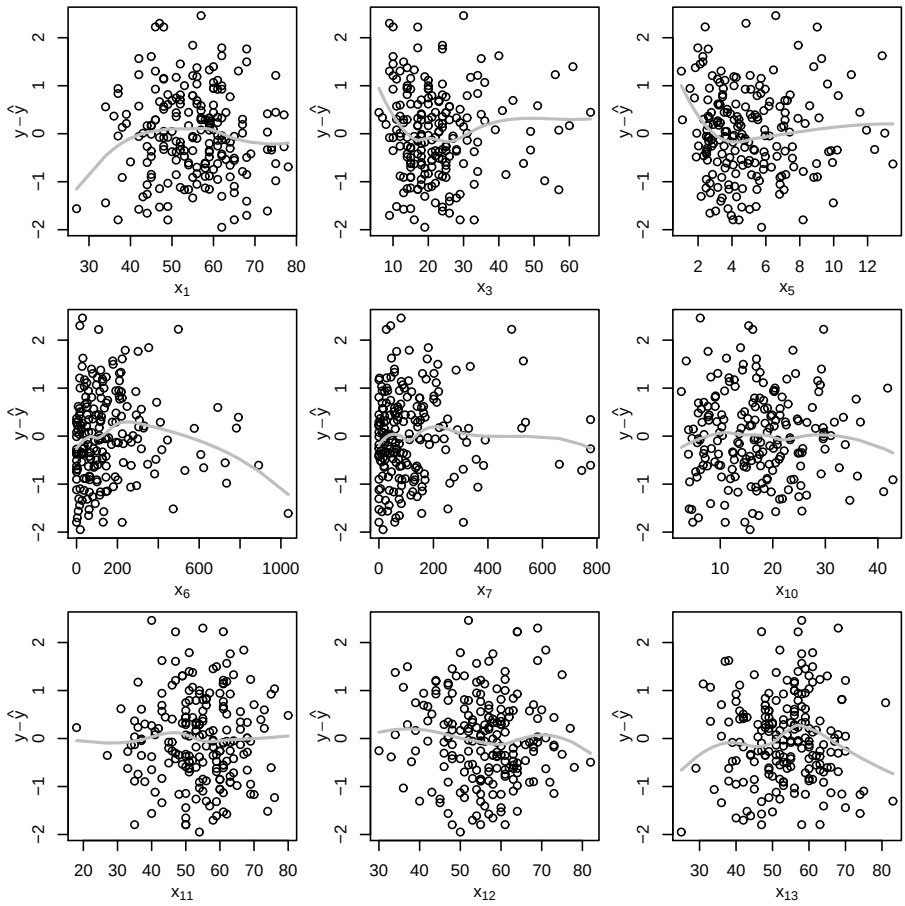

Figure 4: Residuals from the fit of a linear model (obtained via backward elimination) together with loess fits (grey curves) for $n = 200$ observations.

function, the smoothed residuals clearly indicate that structure has been missed. However, the smoothed residuals also indicate structure for covariates without effect. For example, the fitted smooth function is similar for x_{i6} and x_{i7}, whereas only the former really has non-linear structure and the latter has no effect at all. We may therefore anticipate that it will also be challenging for multivariable techniques that can incorporate non-linear effects to distinguish between truly non-linear structure and artifacts.

6 Discussion

Building reliable multivariable regression models is a difficult task which is further complicated when selection of appropriate functional forms for continuous covariates is required. Many techniques exist for doing this, but guidance is largely absent. More detailed evaluation of regression modelling techniques that allow for a potentially non-linear influence of covariates is needed. The basic tool for such investigations is simulation, since real data examples are only of limited value for comparing techniques. However, the simulation design should reflect a real situation which is typically more complex than most of the simple designs considered in the literature. We therefore decided to base key components of our design on a real data example. This can provide deeper insight compared with the original example alone.

After considering existing simulation designs and discussing their weaknesses, we proposed a new simulation design for evaluating multivariable regression modeling techniques with continuous covariates. Building on the ART study (5), the proposal includes a realistically large number of continuous and categorical covariates with a complex correlation structure. For continuous covariates with non-linear influence, we consider global shapes as well as local effects in the true functions.

In an illustrative analysis assuming linearity, the residuals show that data generated

from the proposed design is challenging, if identification of only the truly influential covariates and functional forms is desired. Of course, the difficulty depends on the number of observations and the signal-to-noise ratio. Because we believe that real data poses similar challenges, we proposed settings that are manageable for linear regression techniques, but might prove problematic for approaches that can support non-linear effects. We do not know any similarly complex designs that truly challenge multivariable modelling techniques and might thereby indicate their limits.

Helpful criteria for evaluating the quality of models is another issue requiring closer attention. The mean square error of prediction does not tell us much about the properties of selected models. For more detailed evaluation of approaches to model selection, we have suggested several performance measures that focus not only on prediction performance but also on the interpretability of fitted models. These measures are useful for identifying Type I errors, reflecting overfitted models, but also for highlighting Type II errors, where true structure has been missed.

Although the current design has been developed for a continuous response variable, most of the measures also apply to models with other response types (e.g. binary or time-to-event outcomes). However, adapting the design will raise additional issues. For example, in Cox proportional hazards models for time-to-event data, misspecifying the function for one covariate may adversely affect the estimates for all other covariates, even if they are uncorrelated (28). Similarly, modelling time-dependent effects must be considered in a time-to-event setting (see (29), for example).

We hope that the proposed design (and others similarly complex), customized to the needs of a given simulation study, together with the performance measures we describe here, will be widely adopted for evaluating various model selection techniques. We are not aware of any other design that adequately reflects many of the challenges of multivariable modelling with continuous covariates in a biomedical setting. Our design may

therefore serve as a good starting-point for modifications suited to particular needs. To facilitate its use, we will make the design and the measures available through an easy-to-use R package. We hope that this will provide a building-block towards developing further guidance for multivariable model-building with continuous covariates in real world applications.

Acknowledgements

Harald Binder and Willi Sauerbrei gratefully acknowledge support from Deutsche Forschungsgemeinschaft (SA 580/4-2). Patrick Royston was supported by the UK Medical Research Council (Grant MC_US_A737_0002).

References

[1] Harrell FE. *Regression Modeling Strategies*. Springer: New York, 2001.

[2] Miller AJ. *Subset Selection in Regression*. Chapman & Hall: London, 2002.

[3] Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *Applied Statistics* 1999; **48**(3):313–329.

[4] Copas JB. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society B* 1983; **B 45**(3):311–354.

[5] Royston P, Sauerbrei W. *Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. Wiley: Chichester, UK, 2008.

[6] Abrahamowicz M, du Berger R, Grover SA. Flexible modeling of the effects of serum

cholesterol on coronary heart disease mortality. *American Journal of Epidemiology* 1997; **145**(8):714–729.

[7] Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman & Hall: London, 1990.

[8] Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press, 2003.

[9] Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics* 1994; **43**(3):429–467.

[10] Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: Transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society A* 1999; **162**(1):71–94.

[11] Leeb H, Pötscher BM. Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* 2006; **34**(5):2555–2591, doi: 10.1214/009053606000000821.

[12] Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine* 2007; **26**:5512–5528, doi:10.1002/sim.3148.

[13] Binder H, Sauerbrei W, Royston P. Multivariable model-building with continuous covariates: 2. Comparison between splines and fractional polynomials. FDM-Prepring, University of Freiburg, 2011.

[14] Benedetti A, Abrahamowicz M. Using generalized additive models to reduce residual confounding. *Statistics in Medicine* 2004; **23**(24):3781–3801, doi:10.1002/sim.2073.

[15] Friedman JH, Silverman BW. Flexible parsimonious smoothing and additive modeling. *Technometrics* 1989; **31**(1):3–21.

[16] Zhang HH, Wahba G, Voelker YLM, Ferris M, Klein R, Klein B. Variable selection and model building via likelihood basis pursuit. *Journal of the American Statistical Association* 2004; **99**(467):659–672.

[17] Lin Y, Zhang HH. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics* 2006; **34**(5):2272–2297.

[18] Avalos M, Grandvalet Y, Ambroise C. Parsimonious additive models. *Computational Statistics & Data Analysis* 2007; **51**(6):2851–2870.

[19] Binder H, Tutz G. A comparison of methods for the fitting of generalized additive models. *Statistics and Computing* 2008; **18**(1):87–99, doi:10.1007/s11222-007-9040-0.

[20] Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Statistics in Medicine* 2 2003; **22**(4):639–59, doi:10.1002/sim.1310.

[21] Binder H, Sauerbrei W. Stability analysis of an additive spline model for respiratory health data using knot removal. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2009; **58**(5):577–600.

[22] Ramsay JO, Silverman BW. *Functional Data Analysis*. 2nd edn., Springer: New York, 2005.

[23] Binder H, Sauerbrei W. A new measure for judging the shape of function estimates and penalizing wiggliness 2010.

[24] Schumacher M, Bastert G, Bojar H, Hubner K, Olschewski M, Sauerbrei W, Schmoor C, Beyerle C, Neumann RL, Rauschecker HF. Randomized 2 x 2 trial

evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. *Journal of Clinical Oncology* 1994; **12**:2086–2093.

[25] Tutz G, Binder H. Boosting ridge regression. *Computational Statistics & Data Analysis* 2007; **51**(12):6044–6059.

[26] Hielscher T, Zucknick M, Werft W, Benner A. On the prognostic value of survival models with application to gene expression signatures. *Statistics in Medicine* 2010; **29**(7-8):818–829, doi:10.1002/sim.3768.

[27] Cleveland WS, Grosse E, Shu WM. Local regression models. *Statistical Models in S*, Chambers JM, Hastie TJ (eds.). chap. 8, Pacific Grove, California, 1992.

[28] Gerds TA, Schumacher M. On functional misspecification of covariates in the Cox regression model. *Biometrika* 2001; **88**:572–580.

[29] Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and nonlinear effects of continuous covariates on survival. *Statistics in Medicine* 2007; **26**(2):392–408.