Sauerbrei W and Royston P (2017): The Multivariable Fractional Polynomial Approach, with Thoughts about Opportunities and Challenges in Big Data. In: Hans-Joachim Mucha (Ed.)Big data clustering: Data preprocessing, variable selection, and dimension reduction. WIAS Report 29, Berlin: 36-54.

3 The Multivariable Fractional Polynomial Approach, with Thoughts about Opportunities and Challenges in Big Data

Willi Sauerbrei
Institute for Medical Biometry and Statistics
University of Freiburg, Germany
wfs@imbi.uni-freiburg.de
and
Patrick Royston
2MRC Clinical Trials Unit at UCL
London, UK

Abstract

Data analysts are often faced with many covariates and a suitable model for explanation requires the selection of a subset of variables with a relevant influence on the outcome. For continuous variables it is important to determine a suitable function which fits the data well. We will introduce the basic concept and philosophy of the multivariable fractional polynomial (MFP) approach, which tackles both issues simultaneously. In the context of comparing two treatments we will introduce MFPI as an extension to investigate for potential interactions with continuous covariates. The approach avoids well known problems introduced by categorization. We will also introduce various opportunities and challenges of fractional polynomial modelling in Big Data. Furthermore, we will argue that treatment comparisons need to be based on well-designed randomized trials. In general, observational data do not allow to derive an unbiased estimate of the treatment effect, even if the sample size is very large.

Introduction

The number of covariates potentially included in a regression model is often too large and a more parsimonious model may have advantages. Several variable selection strategies (e.g. all-subset selection with various penalties for model complexity, or stepwise procedures) have been proposed for a long time (Sauerbrei, 1999). As there are few analytical studies about their properties, their usefulness is controversial. With continuous covariates the usual assumption of linearity may be violated. The multivariable fractional polynomial (MFP) approach simultaneously determines a functional form for continuous covariates and deletes uninfluential covariates (Royston and Altman, 1994; Sauerbrei and Royston, 1999; Sauerbrei et al., 2007a; Royston and Sauerbrei, 2008).

Continuous covariates are measured in most of the studies in the health sciences and MFP has become a popular approach for multivariable model building. For variable selection it uses backward elimination and for continuous covariates it checks whether a suitable (non-linear) function from the class of fractional polynomials improves the fit significantly (Royston

DOI 10.20347/WIAS.REPORT.29

and Altman, 1994; Royston and Sauerbrei, 2008). The method also allows categorical and binary covariates. Extensions of MFP have been developed to look for interactions between continuous covariates and treatment (MFPI), between two continuous covariates (MFPIgen) and for interactions with time (non-proportional hazards, MFPT) in a Cox model (Royston and Sauerbrei, 2008; Sauerbrei et al., 2007b; Buchholz and Sauerbrei, 2011).

We have substantial experience in the analysis of real and simulated data with MFP, but restricted to 'larger' data sets. We will introduce key issues of MFP modelling and briefly discuss some opportunities and challenges when using fractional polynomial modelling in the context of 'big data', a highly relevant topic for the future. The phrase 'big data' is used for many different types of very large amounts of automatically collected data. Unfortunately, the concept of big data is not well-defined, since an essentially arbitrary dividing line seems to be imposed on the sample size, for no apparent reason. Nevertheless, the term seems to have stuck. In a recent Editorial, David Hand (2016) stresses the importance of distinguishing between two types of activity relating to big data. The first involves primarily data manipulation: sorting, searching, matching, and so on. Examples include online route finders and apps for updated status of bus and train traffic, with the associated issues addressed mostly by computer scientists and mathematicians. The second type of big data activity seeks to go beyond the data at hand, with the ultimate goals being either prediction of future data, or understanding of the mechanisms and processes that have generated the collected data. Achieving these goals will rely primarily on state-of-the-art statistical and machine learning methods. In addition, the method of data collection is relevant; briefly we may distinguish whether data come from a well-designed experiment (e.g. a randomized trial), a systematic collection (e.g. cancer registry) or whether they are 'found' data (e.g. internet poll). For a discussion see Keiding and Lewis (2016).

In this paper we have the second type of data in mind and as an application we will discuss key issues when comparing two treatments. Often, differences between effects of competing treatments are relatively small, but nevertheless relevant for patients. We will argue that data from a 'larger' randomized trial is required and that data from observational studies, even if the data set is 'Big' (very large), would not help to provide an unbiased estimate of treatment differences (Harford, 2014; Antes, 2015). We will also argue that the information from many RCTs is not fully exploited and discuss that MFPI should play a prominent role to investigate for potential interactions of a continuous covariate with treatment. Having 'Big Data' in mind and assuming that the selection of covariates and functional form for continuous covariates are an important part of the analysis, we will discuss opportunities and challenges of an analysis using MFP.

In this paper we have a 'larger' data set in mind, thoughts about big data are postponed to the specific subsection 6. In subsection 1 we discuss several key issues in variable selection. This is followed by subsection 2 on handling continuous covariates and the introduction of fractional polynomial modelling. The basic concept and philosophy of MFP modelling is introduced in subsection 3, followed by a short subsection on MFPI, the extension to investigate for interaction between a continuous and a binary covariate (subsection 4). MFPI can play an important role when comparing two treatments (subsection 5). Before giving concluding remarks, in subsection 6 we discuss issues of MFP modelling in the context of big data. We have extensively published on the methodology and therefore details will not be given. We refer to the original papers, our book and the MFP website.

3.1 Model Building when Several Covariates are Available

In fitting regression models, data analysts are often faced with many covariates that may have an influence on an outcome variable. Consensus is that subject matter knowledge should generally guide model building, but it is often limited or at best fragile, making data-dependent model building necessary (Harrell, 2001). If the number of covariates is large, a parsimonious model involving a subset of the available covariates is often preferable (Sauerbrei, 1999). An aim of the analysis is the selection of covariates with more than a negligible influence on the outcome. In the health sciences the most popular methods for continuous, binary and censored survival data outcomes are normal-errors (linear) regression, logistic regression and Cox regression models. Issues and methods for variable selection are very similar among the three models mentioned. Usually, methods for variable selection and related issues have been developed and investigated for a normal-errors linear regression model and the methods, or at least their basic ideas, are commonly transferred to generalized linear models and to models for survival data. Sometimes additional problems, such as the definition of residuals or equivalents of R^2, exist. We refer to Andersen and Skovgaard (2010) for a text providing a useful unified treatment of regression models for different types of outcomes.

Relevant issues

In this part of subsection we assume that ’linearity’ is a suitable assumption for the effect of a continuous covariate and our main emphasis is on models for explanation (interpretation). We have more traditional methods for variable selection (e.g. backward elimination) in mind. There have been several recent developments in the literature on variable selection but we know of no strong argument favoring replacement of backward elimination with another procedure in the MFP algorithm (see subsection 3). Some of our arguments are hardly defensible for ’small’ sample sizes and high-dimensional data, such as -omics data. Such situations are implicitly excluded. Under our assumptions we consider the following issues as the most relevant to model selection: Aim (model for prediction or for explanation), model complexity, model stability, incorporating the model uncertainty concept, selection bias and shrinkage of regression coefficients as a potential way to correct for it. For more details see our website `http://mfp.imbi.uni-freiburg.de/`.

Aim of the model and model complexity

Many different aims are possible when developing a multivariable model and the specific aim has an influence on the suitability of a chosen approach. For a detailed discussion see section 2.4 in Royston and Sauerbrei (2008). In many analyses the most important distinction is between models aiming to derive a suitable covariate and models aiming to identify factors which seem to help explaining the value of an outcome. For a discussion see the paper entitled 'To explain or to predict' by Shmueli (2010). He illustrates that these phrases mean different things in different disciplines and mentions relevant distinctions and practical implications for explanatory and predictive modeling. For example, in the social sciences the term explanatory model is used nearly exclusively for testing causal theory. Although we agree with Shmueli that our approach to derive a model would be better called descriptive modeling, we will proceed with the better known (in the health sciences) term 'explanatory model'.

For stepwise variable selection procedures, the significance level (to be chosen by the analyst) is the key user-adjustable setting that influences model complexity. For details on stepwise procedures and a discussion of the close relationship between the significance level and the information criteria AIC and BIC see section 2.6 in Royston and Sauerbrei (2008). Deriving explanatory models is the main aim in this paper. There are several arguments that simpler models are preferable for such situations (Sauerbrei, 1999; Royston and Sauerbrei, 2008 section 2.9.4).

Model complexity, model stability and model uncertainty

Model complexity, model stability and model uncertainty are three different issues of data-dependent model building. However, they are closely related. A more complex model (in this context, a model including more covariates) is usually less stable as it almost invariably includes several covariates which have only a 'weak' effect on the outcome (Sauerbrei and Schumacher, 1992; Sauerbrei et al., 2015). When selecting a specific model, the uncertainty of the selection process is (usually) ignored. To improve models for prediction, the model uncertainty concept was introduced some 20 years ago (Chatfield, 1995, Draper, 1995). A predictor and its variance are estimated by averaging predictors from many (unstable) models. Usually the Bayesian framework is used for model selection and assessment of model uncertainty (Bayesian model averaging; Hoeting et al., 1999). Extending an approach by Buckland et al. (1997), Augustin et al. (2005) suggested using the bootstrap to handle model uncertainty. In contrast to the Bayesian approach which uses Occam's razor to reduce the number of models, Augustin et al. (2005) proposed using a screening step to eliminate covariates with at most a weak effect. Obviously, the number of models included in the second part for model averaging is severely reduced. For a detailed illustration see the example in Sauerbrei et al. (2015). In subsection 3 we will describe the MFP approach to select covariates and functional relationships for continuous covariates. We have also conducted some investigations in the context of function stability (Royston and Sauerbrei, 2003).

Variable selection and shrinkage

Concerning approaches for variable selection, the situation is very confusing. Triggered by the problem of identifying a small number of relevant covariates in a high-dimensional data, many procedures have been proposed recently. However, the number of helpful comparisons between strategies is limited. There is agreement that variable selection will cause biases in estimates of regression parameters and many of the more recent strategies combine variable selection with shrinkage in a regularized approach. For an overview of techniques see Hastie et al. (2009) and several issues are also discussed in Schumacher et al. (2012). In the context of low-dimensional data van Houwelingen and Sauerbrei (2013) assessed whether post-selection two-step approaches using global shrinkage proposed by van Houwelingen and Le Cessie (1990) or parameterwise shrinkage (PWSF, (Sauerbrei, 1999)) can improve selected models. They also compared results to models derived with the LASSO procedure (Tibshirani, 1996), probably the most popular approach to combine variable selection and shrinkage in a one-step approach. Concerning prediction ability the performance of backward elimination (BE) with a suitably chosen significance level was not worse compared to the LASSO and BE models selected were much sparser, an important advantage for interpretation and transportability. It could be shown that the PWSF approach compares favourably to global shrinkage. It was summarized that BE followed by PWSF is a suitable approach when variable selection is a key part of data analysis, provided that the amount of information in the data is not ’too small’.

In the context of using the MFP procedure to derive a multivariable model data-dependently, it was noted that regression parameter estimates of FP functions are biased and may need to be shrunken (Sauerbrei and Royston, 1999). The PWSF approach was considered as one potential way to handle this issue. However, for covariates which are either highly correlated or associated with regard to contents, such as several parameters describing a nonlinear FP2 function, the approach has weaknesses. For such cases the methodology was extended by so-called ’joint shrinkage factors’, a compromise between global and parameterwise shrinkage (Dunkler et al., 2016).

3.2 Continuous Covariates

Continuous covariates are often encountered in life. We measure age, weight, blood pressure and many other things. In medicine, such measurements are often used to assess risk or prognosis or to select a therapy. However, the question of how best to extract useful information from continuous covariates is an important challenge (Rosenberg et al., 2003), in the multivariable context interrelated with the selection of covariates for inclusion in a model. In a short summary, topic group 2 of the STRATOS (STRengthening Analytical Thinking for Observational Studies) initiative states (Sauerbrei et al., 2014):

“In practice, multivariable models are usually built through a combination of (i) a priori inclusion of well-established ’predictors’ of the outcome of interest and (ii) a posteriori selection of additional variables, based often on arbitrary, data-dependent procedures and criteria such as statistical significance or goodness-of-fit measures. There is a consensus that all of the

many suggested model building strategies have weaknesses (Miller, 2002) but opinions on the relative advantages and disadvantages of particular strategies differ considerably. The effects of continuous predictors are typically modeled by either categorizing them (which raises such issues as the number of categories, cutpoint values, implausibility of the resulting step-function relationships, local biases, power loss, or invalidity of inference in case of data-dependent cutpoints) (Greenland, 1995) or assuming linear relationships with the outcome, possibly after a simple transformation (e.g. logarithmic). Often, however, the reasons for choosing such conventional representation of continuous variables are not discussed and the validity of the underlying assumptions is not assessed.

To address these limitations, statisticians have developed flexible modeling techniques based on various types of smoothers, including fractional polynomials (Royston and Altman, 1994; Royston and Sauerbrei, 2008) and several 'flavors' of splines. The latter include restricted regression splines (Harrell, 2001; Boer, 2001), penalized regression splines (Wood, 2006) and smoothing splines (Hastie and Tibshirani, 1990). For multivariable analysis, these smoothers have been incorporated in generalized additive models."

To categorize or to model?

For continuous covariates, a simple and popular approach is to assume a linear effect, but the linearity assumption may be questionable. To avoid this strong assumption, researchers often apply cutpoints to categorize the covariate, implying regression models with step functions. This simplifies the analysis and may or may not simplify interpretation of results. It seems that the usual approach in clinical and psychological research is to dichotomize continuous covariates, whereas in epidemiological studies it is customary to create several categories, often four or five, allowing investigation of a crude dose-response relationship. However, categorization discards information and raises several critical issues such as how many cutpoints to use and where to place them (Altman et al., 1994; Royston et al., 2006). Sauerbrei and Royston (2010) illustrate several critical issues by investigating prognostic factors in patients with breast cancer. As a more suitable approach to analysis, they propose to model continuous covariates with fractional polynomials (FP). See Royston and Sauerbrei (2008) for a monograph on this topic and the related website `http://mfp.imbi.uni-freiburg.de/`.

Fractional polynomials

Class of FP functions

The class of fractional polynomial (FP) functions is an extension of power transformations of a covariate. For most applications FP1 and FP2 functions are sufficient.

FP1: $\beta_1 x^{p1}$

FP2: $\beta_1 x^{p1} + \beta_2 x^{p2}$

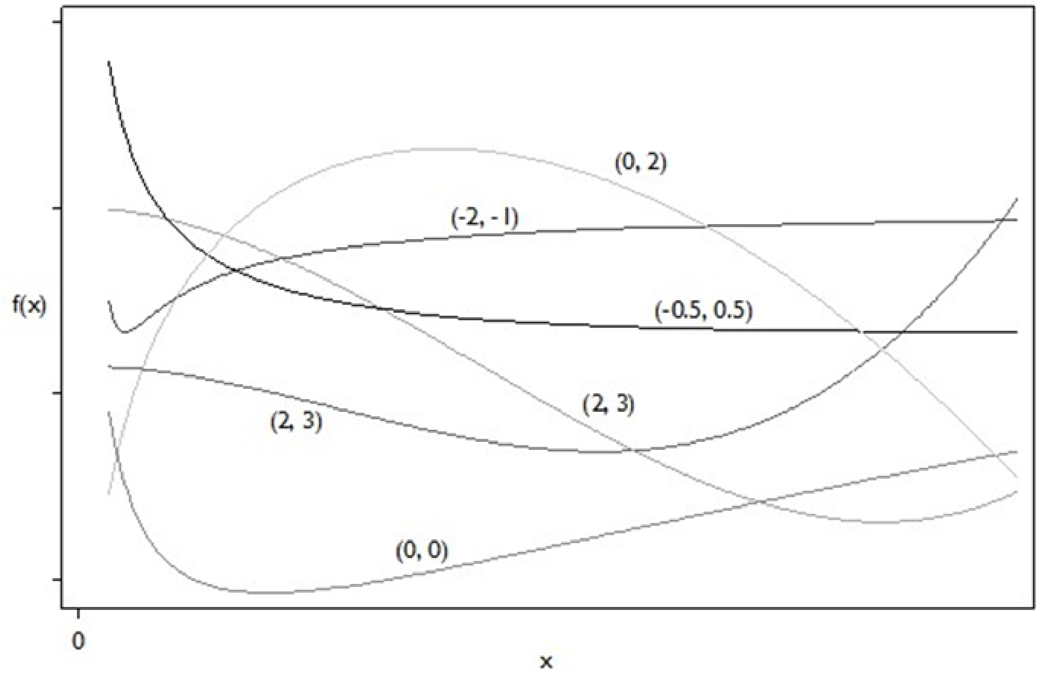

Figure 18: Various shapes of FP2 functions with different power terms $p1$ and $p2$.

For the exponents $p1$ and $p2$ a set **S={-2, -1, -0.5, 0, 0.5, 1, 2, 3}, with $\mathbf{0} = \log \mathbf{x}$** was proposed. For $p1 = p2 = p$ ('repeated powers') an FP2 function is defined as $\beta_1 x^p + \beta_2 x^p \log x$. This defines 8 FP1 and 36 FP2 models. The values $p1 = 1,\ p2 = 2$ define the quadratic function. The class of FP functions seems to be small, but it includes very different types of shapes (Fig. 18). General FPm functions are well-defined and straightforward, but will not be discussed here as they are rarely used. FP3 or more complex FP functions may improve the fit in some cases (particularly in a univariate analysis), but in the multivariable context, which is the main issue here, we are not aware of any relevant example. Occasionally they also find a use as effective approximations to intractable mathematical functions (Royston and Altman, 1997).

Selecting an FP function

A suitable function should fit the data well, and also be simple, interpretable and generally usable. To assess whether a covariate has a significant effect, the FP function selection procedure (FSP) starts by comparing the best fitting allowed FP (often FP2) function of a continuous covariate x with the null model (Royston and Sauerbrei, 2008 section 4.10). If significant, the procedure proceeds by comparing FP functions with a 'simple' (usually linear) default function. Using FSP the default function is often selected. More complex FP functions are chosen only if they fit the data much better (based on a significance criterion), which implies that sample size (effective sample size in survival data) plays an important role. Modifications are required in 'big data', see subsection 6.

Before starting to select a suitable function, the analyst must decide on a nominal p-value (α) and on the degree (m) of the most complex FP model allowed. Typical choices in medicine are $\alpha = 0.05$ and FP2 ($m = 2$). In the following we describe FSP when FP2 is chosen. It is straightforward to adapt the procedure for use with other FP degrees. Based on minimizing the deviance (minus twice the maximized log likelihood), the best FP1 and best FP2 function are determined. The following test procedure assumes that the null distribution of the difference in deviances between an FPm and an FP $(m-1)$ model is approximately central χ^2 on two degrees of freedom. For details see section 4.9.1 of Royston and Sauerbrei (2008). The FP function is determined for the variable x using the following closed test procedure:

1 Test the best FP2 model for x at the α significance level against the null model using four d.f. If the test is not significant, stop, concluding that the effect of x is 'not significant' at the α level. Otherwise continue.

2 Test the best FP2 for x against the default (usually a linear function) at the α level using three d.f. If the test is not significant, stop, the final model being the default. Otherwise continue.

3 Test the best FP2 for x against the best FP1 at the α level using two d.f. FP2 selects two power terms and estimates two corresponding parameters, therefore 4 d.f.; correspondingly FP1 has 2 d.f., giving a difference of two d.f. If the test is not significant, the final model is the best FP2, otherwise the final model is the best FP1. End of procedure.

Note that the α level for the selection of the FP function can be different to the significance level of backward elimination. If α = 1 in the latter then x is always selected and step 1 is redundant. Using the flavor of a closed test procedure ensures that the overall type 1 error is close to the nominal significance level. For some results concerning type 1 error and power we refer to simulation studies described in section 4.10.5 of our book.

3.3 MFP: an Approach to Multivariable Model-building with Several Continuous Covariates

MFP is an approach to multivariable model-building which retains continuous covariates as continuous, finds non-linear functions if sufficiently supported by the data, and removes weakly influential covariates by backward elimination (BE). The main issues of the approach arise from the two key components: variable selection with backward elimination and selection of an FP function to model non-linearity.

The MFP algorithm - basic concept

Like backward elimination, the MFP algorithm starts with all candidate covariates entered as linear terms (the 'full model') and investigates whether any covariates can be eliminated. However, for each of the continuous covariates the FSP is used to check whether a non-linear function fits the data significantly better than a linear function. After a first cycle some covariates will often be eliminated and for some continuous covariates a better fitting non-linear function may have been determined. The algorithm starts a second cycle, but the new starting model now has fewer covariates (as some were eliminated) and perhaps non-linear functions for some of the continuous covariates. In the second cycle all covariates are reconsidered (even if they were not significant at the end of the first cycle) and the FSP is used again to determine the 'best' fitting FP function (it may be different because other 'adjustment' covariates are in the model). This yields the result of cycle 2 which is the starting point for cycle 3. In most cases the model does not change anymore in cycle 3 or 4 and the algorithm stops with the final MFP model.

Important is the order of 'searching' for model improvement by better fitting non-linear functions. Obviously, mismodelling the functional form of a covariate with a strong effect is more critical than mismodelling the functional form of a covariate with a weak effect. The order is determined by ascending p-values from likelihood ratio tests for elimination from the full model. Covariates with a small p-value are considered first. Boxes 6.1 and 6.2 in Royston and Sauerbrei (2008) illustrate the algorithm in an example. Most often 0.05 is used as the significance criteria for both variable elimination and function selection, however, these two important parameters for variable and function selection can (and should) be flexibly chosen by the analyst. Depending on the aim of an analysis more or less stringent significance criteria may be preferable.

MFP modelling - philosophy and related matters

For a detailed description of the algorithm and some relevant issues see Chapter 6 in Royston and Sauerbrei (2008). In the discussion of it, we consider in detail four relevant issues (1 - Philosophy of MFP; 2- Function Complexity, Sample Size and Subject-Matter Knowledge; 3- Improving Robustness by Preliminary Covariate Transformation; 4- Conclusion and Future). Our thoughts about these issues are summarized in a table entitled 'Towards recommendations for model building by selection of variables and functional forms for continuous predictors in observational studies, under the assumption of Tab 1.3.' This table is adapted from Sauerbrei et al. (2007a), where we expressed thoughts about our philosophy of MFP modelling:

"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 (Royston and Sauerbrei, 2003) and time-varying effects (Sauerbrei et al., 2007b). Modifications that may improve the usefulness of MFP are combination with shrinkage and a more systematic check for overlooked local curvature."

In a large simulation study comparing MFP with various spline approaches, we provided some evidence for the conclusions given in the table of recommendations', but further simulation studies are needed (Binder et al, 2013). It is planned to conduct them in topic group 2 of the STRATOS initiative 'Selection of variables and functional forms in multivariable analysis' (Sauerbrei et al, 2014).

3.4 Extension of MFP to Investigate for Interactions

Given the enormous amount of resources spent on conducting a large clinical trial, it is surprising that greater efforts are not made to try to extract more information from clinical trials data. In the context of potential interactions between continuous covariates and treatment, we have argued for the use of MFPI (multivariable fractional polynomials - interaction) for such investigations (Royston and Sauerbrei, 2004; Sauerbrei and Royston, 2007). Unfortunately, dichotomization is still the 'standard', even though most of the well-known problems of categorization mentioned above transfer to analyses for interactions. The key ideas of MFPI are: first, MFPI estimates for each treatment group a fractional polynomial function representing the prognostic effect of the continuous covariate of interest, optionally adjusting for other covariates. Second, the difference between the functions for the treatment groups is calculated and tested for significance. The testing is done through an analysis of interaction between treatment and the FP function. A plot of the difference (e.g., log hazard ratio) against the covariate, together with a 95% CI, is termed a 'treatment-effect plot'. A treatment-effect plot for a continuous covariate not interacting with treatment would be a straight line parallel to the x-axis, whereas a treatment-covariate interaction would be indicated by a non-constant line, often increasing or decreasing monotonically. For more details see subsection 6 in our book (Royston and Sauerbrei, 2008). In a recent simulation study we were able to illustrate striking advantages of MFPI over methods based on dichotomization or categorization (Royston and Sauerbrei, 2013; Royston and Sauerbrei, 2014). Based on these results, we slightly changed our recommendation for the most suitable approach (our new default). For details see the website or the latter paper.

3.5 Opportunities of MFPI when Comparing Treatments

By re-analyzing data from an MRC randomized trial in patients with renal cancer, we illustrated additional opportunities to investigate for interactions of a continuous covariate with treatment (Royston et al., 2004). In Fig. 19 we show Kaplan-Meier estimates in all patients and in patients defined by 4 subgroups based on white cell count (WCC) values. These subgroups are motivated by the treatment effect function for WCC (top right). Using MFPI we investigate ten continuous covariates as potential modifiers of the treatment effect. Nine covariates did not exhibit any important interaction, but for WCC the test for interaction was significant at the 1% level. We use Kaplan-Meier plots in subpopulations as check of the derived treatment effect function.

The five plots of Kaplan-Meier estimates show that the proportional hazard assumption of the Cox model is acceptable in all populations and we estimated treatment effects in each of the groups. The estimated hazard ratio (HR: Interferon to MPA; 95% confidence interval) in all patients is 0.75 (0.60 - 0.93), which clearly shows the benefit of interferon. However, in subgroups defined by increasing values of WCC we observe increasing estimates agreeing with the treatment effect function and the impression from the plots for subgroups (I: 0.53 (0.34 - 0.83), II: 0.69 (0.44 - 1.07), III: 0.89 (0.57 - 1.37), IV: 1.32 (0.85 - 2.05)).

There is a large effect favoring interferon in group I (very low WCC values). The advantage disappears for patients with higher WCC values. Analyses in subgroups support the estimated treatment effect function.

Concerning the interpretation of results from MFPI analyses, we need to distinguish between prospectively planned analyses and a retrospectively conducted search for markers which may have an influence on the effect of treatment. Results from a retrospective search need to be seen as hypothesis generation, requiring validation in new data. For hypothesis generation we recommend using small p-values (e.g. 0.01), otherwise larger p-values may be acceptable. In any case, we strongly recommend checking estimated treatment effect functions by conducting analyses in subgroups.

3.6 Analyzing Big Data with MFP - on Opportunities and Challenges

So far we have no experience analyzing ’Big Data’ with MFP or more generally with FP methodology. In the following we will consider two very different ’big data’ situations and point to potential opportunities and challenges when using FPs for the analysis.

Large(r) sample size

Having a large sample size offers many opportunities for MFP analyses but also raises several issues of our test-based FP function selection procedure. Obviously FSP needs to be adapted because a very large sample size would (nearly) always result in selecting the most

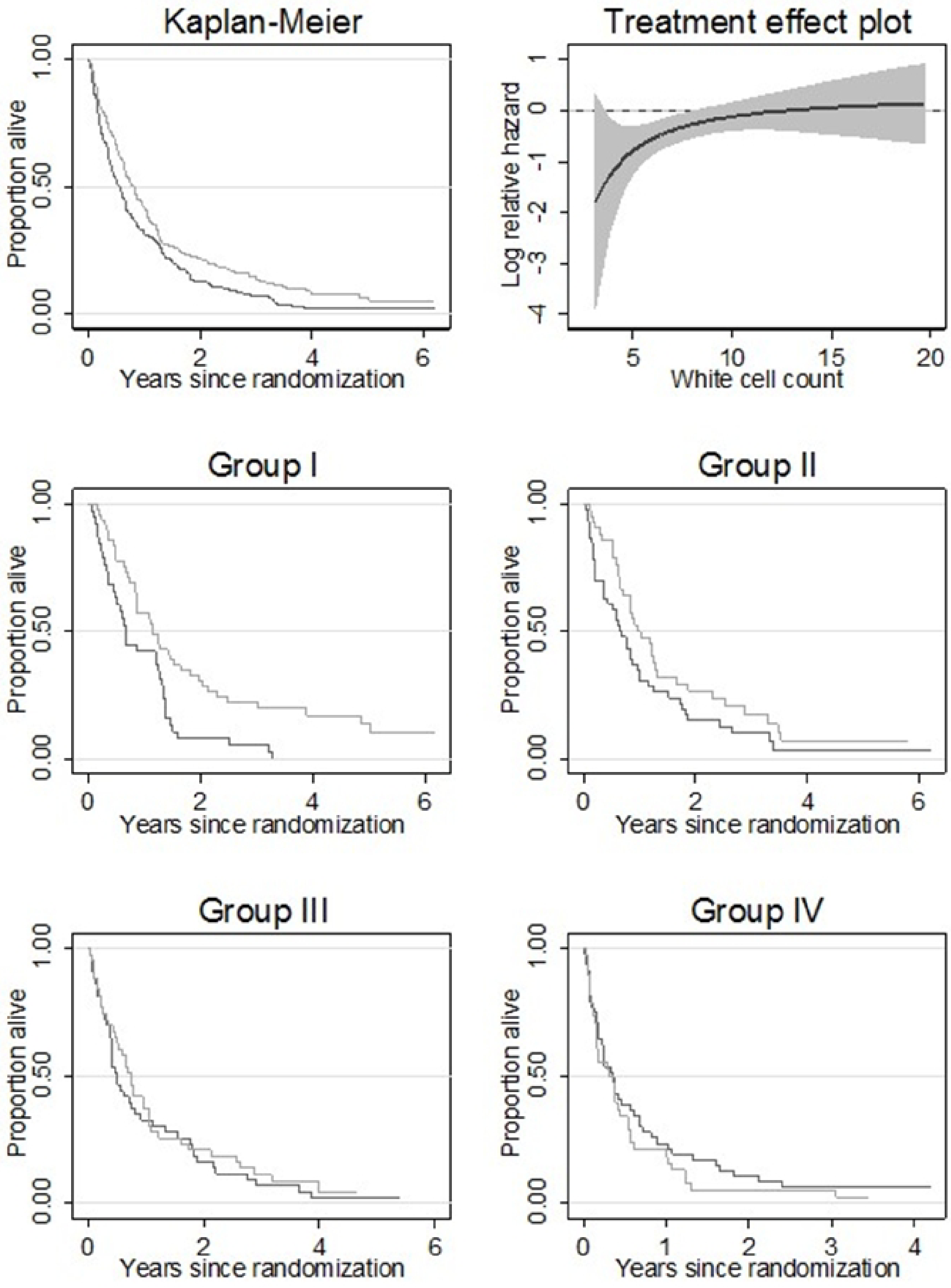

Figure 19: Plots of Kaplan-Meier estimates of overall survival probability for patients treated with interferon (pale) or MPA (dark) and estimated treatment effect function (with 95% CI) comparing the treatment effect dependent on white cell count (WCC), the only significant covariate interacting with treatment (top right). The four Kaplan-Meier plots (middle and bottom) show survival estimates in subgroups determined by WCC values.

complex allowed FPm (typically FP2) function.

Careful consideration of whether FP2 functions should be allowed is one simple way to handle this issue. For example, restriction to the FP1 class could be suitable if subject matter knowledge provides a strong argument that a function should be monotonic. In that case the best fitting transformation of the eight functions would be selected. In a similar way FSP could be modified to select the best of the thirty-six FP2 functions if a non-monotonic function would be suitable.

Another approach would be to use the area between curves (ABC) criteria to replace significance tests in the FSP. ABC was proposed by Govindarajulu et al. (2007) to quantify the distance between smoothed curves and later adapted to quantify the distance between two curves estimating time-varying effects in the Cox model for survival data (ABCtime, Buchholz et al., 2014). In a procedure similar to FSP, distances between best FP2, best FP1 and the linear function could be considered. However, further work on a suitable metric to compare two curves is needed. What is a relevant ABC value to conclude that the best FP2 fits 'substantially better' than the linear function or the FP1 function? This issue needs experience in real studies and in simulations.

In the context of MFP modelling it is also important to adapt the variable selection part of MFP. One simple possibility is to choose the BIC (Bayes Information Criterion; Schwarz, 1978) as the criterion for model selection. The penalty constant of BIC is log(n), which may help to restrain the selection of many significant covariates with a very small effect. BIC or extremely small p-values such as 0.000001 may also be used in FSP for the selection of an FP function. A different line would to try adapting the variable selection (backward elimination) part by using ideas from the change-in-estimates approach (Greenland et al., 1989). That may also help for categorical covariates, also needing adaption for the case of very large sample sizes.

However, practical experience is needed to see whether these ideas are sufficient to adapt the current MFP procedure to handle the problem of variable and function selection in very large data sets. Very large sample sizes offer also many new possibilities for MFP methodology. For potential interactions between two continuous covariates we have proposed MFPIgen as an extension of MFPI (see section 7.11 in Royston and Sauerbrei, 2008). However, to conduct such an analysis a 'large' sample size is needed. For 'very' large sample sizes an adaption as discussed above may be required.

Having very large datasets allows the analyst to partition the data and give the often neglected model validation aspect much more weight. To get some ideas about external validation of a 'derivation' model, data partitioning is often done, even with 'medium sized' data sets. In very large data sets natural partitions may be available (e.g. three hospitals with large datasets each) and a partition could be possible without the severe disadvantage of losing power, which is often low anyway. See van Houwelingen (2000) for related discussions. Related is the possibility of dividing the data into several (well-defined) subpopulations, conduct an (MFP) analysis in each of them and summarize results in a meta-analysis. Using 'big data' from nine SEER registries, we proposed a new approach for the meta-analysis of functions (Sauerbrei and Royston, 2011).

Very large number of covariates and small sample size

This situation is becoming more and more relevant in the health sciences, often called ’omics’ research. This term encompasses multiple molecular disciplines that involve characterization of global sets of biological molecules such as DNAs, RNAs, proteins and metabolites (IoM (Institute of Medicine), 2012). Typical sample sizes are between 100 and 500 (for survival data the effective sample size, the number of events, is often much smaller) and the number of covariates range from several hundreds to several hundred thousand. Obviously, deriving a ’suitable’ model is a challenge. ’Traditional’ statistical modelling approaches cannot be used and many strategies have been adapted and developed during the last years. Considering a preliminary covariate screening step, various methods of regularization and the combination of variable selection and shrinkage play a key role. However, usually it is assumed that the effect of a continuous covariate is linear. We could imagine that consideration of the eight functions from the FP1 class could improve some of the models. After a pre-selection of covariates (say selection of the top 500) it would be easy to consider (in univariate analyses) whether any of the seven non-linear FP1 functions provides a much better fit compared to a linear function. The p-value or the ABC criterion may be used for the comparison.

To identify extreme values in omics data, Boulesteix et al. (2011) used a simple pre-transformation, originally proposed to improve robustness of MFP models (Royston and Sauerbrei, 2007), and compared gene rankings derived from the original and the transformed values. For some datasets they could identify striking differences in the gene rankings, caused by altering single observations. The approach could be extended to consider the best FP1 function as the pre-transformation and an extension to multivariable models should be possible.

Concluding Remarks

We have provided a brief overview of multivariable model building based on fractional polynomials for modelling continuous covariates. We have concentrated on the MFP approach which combines backward elimination as a strategy for variable selection with the selection of a suitable function from a well-defined class of fractional polynomials. The aim of a multivariable model has a substantial influence on the suitability of a model building procedure (Shmueli, 2010). Different strategies can produce very different models, but predictors from different models are often (very) similar (Sauerbrei et al., 2015). In the health sciences models for explanation play a more important role and we have such models in mind in our discussion. For the variable selection part we have discussed model complexity as the key issue and shrinkage as a potential way to correct for bias introduced by data dependent modelling. The complexity of a BE model can be easily controlled by the significance level and we use it as the key parameter for both parts of MFP.

Comparing two (or more) treatment strategies is one of the most important investigations in the health sciences. From a statistical point of view the popular phrase 'individualized treatment' implies interactions of treatment with 'several' patient characteristics. So far interactions with a continuous covariate are usually investigated by categorizing (dichotomizing) the continuous covariate and investigate for treatment differences in subgroups. In the context of prognostic factors, risk factors and many others, the severe disadvantages introduced by categorization have been well known for many years (Altman et al., 1994, Royston et al., 2006). Obviously, most of the problems transfer to models investigating for interactions. For a more detailed discussion see (Hingorani et al., 2013). One of their recommendations reads *"Standards in statistical analysis of prognosis research should be developed which address the multiple current limitations. In particular, continuous variables should be analysed on their continuous scale and non-linear relationships evaluated as appropriate."*

The MFP procedure was extended to MFPI as a more suitable way to investigate for interactions between a binary treatment (extension for categorical covariates are straightforward) and a continuous patients characteristic. In an example we have demonstrated that MFPI analyses can help to identify prognostic factors which interact with treatment, in some areas in medicine they are called predictive factors. To report relevant details of MFP and MFPI analyses is straightforward, another important factor of our approaches. The importance of transparent reporting and reproducible research has become a key issue in medical research. As software for MFP and MFPI is generally available (originally all routines have been programmed by Patrick Royston in Stata, for details see the website) it should be possible to reproduce an analysis, provided the data are publicly available.

So far we have no experiences using MFP and MFPI in the context of 'Big Data'. We have outlined some potential chances and problems in using our approaches. However, the key issues depend on the specific problem and the way data were collected. In the health sciences there are many promises related to 'Big Data' but it is obvious that more data will not solve every problem (Antes, 2015). Very large sample sizes can be helpful to reduce the random error and increase power, but potential biases are the more relevant in many analyses. Consequently, for investigations to compare treatments and to search for treatment modifying factors, we have used the data from a randomized trial. Concerning data quality, Hand (2016) points out that 'large does not necessarily mean good, useful, valuable or interesting. Big does not necessarily mean accurate or comprehensive'. With this short overview we aim to illustrate that fractional polynomial methodology can be used sensibly for many analyses requiring modelling of continuous covariates.

References

Altman, D.G., Lausen, B., Sauerbrei, W. and Schumacher, M (1994): Dangers of using 'Optimal' cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer*

Institute, 86: 829–835

Andersen, P.K. and Skovgaard, L.T. (2010): *Regression with Linear Predictors*. Springer, New York

Antes, G. (2015): *A new Science(ability)?* Lab Times Online

Augustin, N., Sauerbrei, W. and Schumacher, M. (2005): The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling 5: 95–118*

Binder, H., Sauerbrei, W. and Royston, P. (2013): Comparison between splines and fractional polynominals for multivariable model-building with continous covariates: a simulation study with continous response. *Statistics in Medicine 32: 2262–2277*

Boer, C. de (2001): *A Practical Guide to Splines*. revised edn. Springer, New York

Boulesteix A.-L., Guillemot V. and Sauerbrei W. (2011): Use of pretransformation to cope with extreme values in important candidate features. *Biometrical Journal 53(4): 673–688*

Buchholz, A. and Sauerbrei, W. (2011): Comparison of procedures to assess non-linear and time- varying effects in multivariable models for survival data. *Biometrical Journal 53(2): 308–331*

Buchholz, A., Sauerbrei, W. and Royston, P. (2014): A measure for assessing functions of time-varying effects in survival analysis. *Open Journal of Statistics 4: 977–998*

Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997): Model selection: an integral part of inference. *Biometrics 53: 603–618*

Chatfield, C. (1995): Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society, Series B 158: 419–466*

Draper, D. (1995): Assessment and propagation of model selection uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B 57: 45–97*

Dunkler, D., Sauerbrei, W. and Heinze, G. (2016): Global, Parameterwise and Joint Post-Estimation Shrinkage. *Journal of Statistical Software 69: 8*

Govindarajulu, U.S., Spiegelman, D., Thurston, S.W., Ganguli, B. and Eisen, E.A. (2007): Comparing Smoothing Techniques in Cox Models for Exposure-Response Relationships. *Statistics in Medicine 26: 3735–3752*

Greenland, S. (1989): Modeling and variable selection in epidemiologic analysis. *American Journal for Public Health 79(3): 340–349*

Greenland, S. (1995): Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology (Cambridge, Mass.) 6 4: 450–454*

Hand, D.J. (2016): Editorial: 'Big data' and data sharing. *Journal of the Royal Statistical Society, Series A 179, 3: 629–631*

Harford, T. (2014): Big data: are we making a big mistake? *Financial Times*

Harrell, F.E. (2001): *Regression modeling strategies, with applications to linear models, logistic regression, and survival analysis*. Springer, NewYork

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

Hastie, T.J., Tibshirani, R. and Friedman, J. (2009): *The Elements of Statistical Learning*. 2nd edn. Springer, New York

Hingorani, A.D., van der Windt, D., Riley, R.D., Abrams, K., Moons, K.G.M., Steyerberg, E.W., Schroter, S., Sauerbrei, W., Altman, D.G., Hemingway, H. for the PROGRESS Group (2013): Prognosis research strategy (PROGRESS) 4: Stratified medicine research. *British Medical Journal 346*

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999): Bayesian model averaging: A tutorial. *Statistical Science 14: 382–417*

IOM (Institute of Medicine) (2012): *Evolution of Translational Omics: Lessons Learned and the Path Forward*. The National Academies Press, Washington, DC

Keiding, N. and Louis, T.A. (2016): Perils and potentials of self-selected entry to epidemiological studies and surveys. *J.R.Statistical Society, Series A 2: 319–376*

Miller, A. (2002): *Subset Selection in Regression*. Taylor & Francis: Boca Raton, Florida

Rosenberg, P.S., Katki, H., Swanson, C.A., Brown, L.M., Wacholder, S. and Hoover, R.N. (2003): Quantifying epidemiologic risk factors using nonparametric regression: model selection remains the greatest challenge. *Statistics in Medicine 22: 3369–3381*

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

Royston, P. and Altman, D.G. (1997): Approximating statistical functions by using fractional polynomial regression. *The Statistician 46: 411–422*

Royston, P., Altman, D.G. and Sauerbrei, W. (2006): Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine 25: 127–141*

Royston, P. and Sauerbrei, W. (2003): Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Statistics in Medicine 22: 639–659*

Royston, P. and Sauerbrei, W. (2004): A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine 23: 2509–2525*

Royston, P. and Sauerbrei, W. (2007): Improving the robustness of fractional polynomial models by preliminary covariate transformation: a pragmatic approach. *Computational Statistics and Data Analysis 51: 4240–4253*

Royston, P. and Sauerbrei, W. (2008): *Multivariable Model-Building - A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley.

Royston, P. and Sauerbrei, W. (2013): Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis. *Statistics in Medicine 32(22): 3788–3803*

Royston, P. and Sauerbrei, W. (2014): Interaction of treatment with a continuous variable: simulation study of power for several methods of analysis. *Statistics in Medicine 33: 4695–4708*

Royston, P., Sauerbrei, W. and Ritchie, A. (2004): Is treatment with interferon-alpha effective in all patients with metastatic renal carcinoma? A new approach to the investigations of interactions. *British Journal of Cancer 90: 794–799*

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

Sauerbrei, W., Abrahamowicz, M., Altman, D.G., le Cessie, S. and Carpenter, J. on behalf of the STRATOS initiative (2014): STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. *Statistics in Medicine 33: 5413–5432*

Sauerbrei, W., Buchholz, A., Boulesteix, A.-L. and Binder, H. (2015): On stability issues in deriving multivariable regression models. *Biometrical Journal 57: 531–555*

Sauerbrei, W. and Royston, P. (1999): Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society, Series A 162: 71–94*

Sauerbrei, W. and Royston, P. (2007): Modelling to extract more information from clinical trials data: on some roles for the bootstrap. *Statistics in Medicine 26: 4989–5001*

Sauerbrei, W. and Royston, P. (2010): Continuous Variables: To Categorize or to Model? In: Reading, C. (Ed.): *The 8th International Conference on Teching Statistics- Data and Context in statistics education: Towards an evidence based society*. International statistical Institute, Voorburg

Sauerbrei, W. and Royston, P. (2011): A new strategy for meta-analysis of continuous covariates in observational studies. *Statistics in Medicine 30(28): 3341–3360*

Sauerbrei, W., Royston, P. and Binder H. (2007a): Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine 26: 5512–5528*

Sauerbrei, W., Royston, P. and Look, M. (2007b): A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal, 49: 453–473*

Sauerbrei, W. and Schumacher, M. (1992): A Bootstrap Resampling Procedure for Model Building: Application to the Cox Regression Model. *Statistics in Medicine 11: 2093–2109*

Schumacher, M., Holländer, N., Schwarzer, G., Binder, H. and Sauerbrei, W. (2012): Prognostic Factor Studies. In: Crowley, J., Hoering, A. (Eds): *Handbook of Statistics in Clinical Oncology*. Third Edition, Chapman and Hall/CRC, 415–470

Shmueli, G. (2010): To explain or to predict? *Statistical Science 3: 289–310*

Tibshirani, R. (1996): Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B 58(1): 267–288*

Van Houwelingen, H.C. (2000): Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine 19: 3401–3415*

Van Houwelingen, H.C. and le Cessie, S. (1990): Predictive value of statistical models. *Statistics in Medicine 9: 1303–1325*

Van Houwelingen, H.C. and Sauerbrei, W. (2013): Cross-validation, shrinkage and variable selection in linear regression revisited. *Open Journal of Statistics 3: 79–102*

Wood, S. (2006): *Generalized Additive Models*. Chapman & Hall/CRC, New York