Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 4054–4063
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Detecting an interaction between treatment and a continuous covariate: A comparison of two approaches

Willi Sauerbrei[a,*], Patrick Royston[b], Karina Zapien[a]

[a] *Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Stefan-Meierstr. 26, 79104 Freiburg, Deutschland, Germany*

[b] *Cancer Group, MRC, Clinical Trials Unit, 222 Euston Road, London NW1 2DA, UK*

Received 30 October 2006; received in revised form 21 December 2006; accepted 21 December 2006
Available online 2 January 2007

Abstract

In clinical trials, there is considerable interest in investigating whether a treatment effect is similar in all patients, or that some prognostic variable indicates a differential response to treatment. To examine this, a continuous predictor is usually categorized into groups according to one or more cutpoints. The treatment/covariate interaction is then analyzed in factorial fashion using multiplicative terms. The use of cutpoints raises several difficult issues for the analyst. It is preferable to keep continuous variables continuous in such a model. To achieve this, the MFP algorithm for multivariable model-building with fractional polynomials was recently extended to a new algorithm called multivariable fractional polynomial interaction (MFPI). With the latter, covariates may be binary, categorical or continuous, and cutpoints are avoided. MFPI is compared with a graphical technique, the subpopulation treatment-effect pattern plot or subpopulation treatment effect pattern plot (STEPP). Differences between MFPI and STEPP are illustrated by re-analysis of a randomized trial in kidney cancer. The stability of the two procedures is investigated by using the bootstrap. The Type I error probability of MFPI to 'detect' spurious interactions is estimated by simulation. MFPI and STEPP are found to exhibit similar treatment/covariate interactions. The tail-oriented variant of STEPP is found to give more stable and interpretable results than the sliding window variant. The type 1 error probabilty of MFPI is found to be close to its nominal value.

Keywords: Clinical trials; Interaction; Continuous covariates; Fractional polynomials; Stability; Bootstrap; Type 1 error

1. Introduction

In most randomized clinical trials, detailed baseline data are collected on each patient at randomization. These data concern demographics, medical history, current signs and symptoms, and quantitative disease measures. One important use of such data is in subgroup analyses to assess whether differences in outcome (or lack thereof) between treatment groups depend on patient characteristics. The term predictive factor is often used for such characteristics in the clinical literature, whereas in biostatistics the term treatment–covariate interaction is generally used. We use both terms interchangeably.

* Corresponding author. Tel.: +49 761 203 6669; fax: 49 761 203 5002.
E-mail address: wfs@imbi.uni-freiburg.de (W. Sauerbrei).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.12.041

In a review of the use of baseline data in clinical trials, Assmann et al. (2000) found that most trial reports included subgroup analyses, but in less than half was a statistical test for interaction performed. They also concluded that most of the trials lacked power to detect any but very large subgroup effects. Nevertheless many subgroup analyses are often done. The main reason seems to be that researchers do not want to miss real treatment differences that depend on baseline characteristics. Often a search for subgroups is done in a simple way, although more elaborate algorithms searching for subgroups in a multi-dimensional space have also been proposed (Kehl and Ulm, 2006). The potential instability of such methods is a major obstacle to their general adoption.

The question of how to model interaction between a continuous covariate Z and a categorical covariate T in a regression (i.e. analysis of covariance) model has received a fair amount of attention during the last 20 years. Usually, Z is categorized into a number of groups according to one or more cutpoints and to analyse the interaction in a model with main effects and multiplicative interaction terms. A trend test of the effect of T over the ordered categories from Z may be performed and is likely to have more power than the more general unordered test. All of this raises several issues for the analyst, including dependence of the statistical significance of the interaction on the number and position of the cutpoints, the interpretation of the results when an unstable model with too many cutpoints is fitted, and in the case of a trend test, possible loss of power and faulty interpretation if a non-linear relationship is incorrectly assumed to be linear. Another approach is to avoid categorization but to assume linearity in Z at all levels of T—an assumption which may be incorrect.

To use all information from a continuous covariate, but to allow possible non-linearity in Z at all levels of T, Royston and Sauerbrei (2004) proposed the multivariable fractional polynomial interaction (MFPI) algorithm for investigating interactions between a continuous covariate and a binary or categorical (treatment) variable. The algorithm is an extension of the Multivariable fractional polynomial (MFP) procedure for the simultaneous selection of influential prognostic variables and the selection of the functional form for a continuous covariate (Sauerbrei and Royston, 1999). The best transformation of Z within the class of FP2 functions is selected, with the constraint of having the same powers for all the levels of the treatment variable. In the following, we will assume two levels of T, although the extension to more levels is straightforward. MFPI was shown in several examples to be able to identify treatment/covariate interactions which may be missed by methods which do not use the full information from a continuous covariate (Royston and Sauerbrei, 2004; Royston et al., 2004).

Another technique, called the subpopulation treatment effect pattern plot or subpopulation treatment effect pattern plot (STEPP), involves dividing the observations into subgroups defined with respect to the covariate Z of interest and estimating the treatment effect T separately within each subpopulation (Bonetti and Gelber, 2000). To increase the number of patients that contribute to each point estimate, subpopulations are allowed to overlap. To create subpopulations, sliding window (SW) and tail-oriented (TO) variants have been proposed. STEPP was further extended and illustrated in reference Bonetti and Gelber (2004). Although properties of the SW and TO variants and expression of preference between them, or recommendations on the number of groups, have not been published, STEPP has gained some popularity in recent years for the analysis of treatment/covariate interactions, at least in breast cancer (Crivellari et al., 2003; Fisher et al., 2004).

Here we will discuss differences between MFPI and STEPP, compare results in a randomized trial for metastatic renal carcinoma, investigate stability of the two methods and assess type I error probability of MFPI. First we will concentrate on a single variable (white cell count, WCC), identified in an earlier analysis (Royston and Sauerbrei, 2004) with MFPI as interacting with treatment. For STEPP we will vary the number of subpopulations and compare the SW and TO variants. Then we will investigate stability of MFPI and STEPP by using the bootstrap. In a small simulation study we will investigate type I error of MFPI. Finally, for variables not identified as interacting with treatment according to MFPI, we will use the more flexible STEPP method to check whether the data gives any evidence of an interaction missed by MFPI. This may at least suggest the type II error of MFPI. Finally, we will discuss implications for investigations of continuous, potentially predictive factors in randomized controlled trials.

2. Data and methods

2.1. Data

The methods will be illustrated by re-analysing the MRC RE01 trial comparing interferon-α with medroxyprogesterone acetate (MPA) in patients with metastatic renal carcinoma. The study is a randomized trial recruiting 350 patients

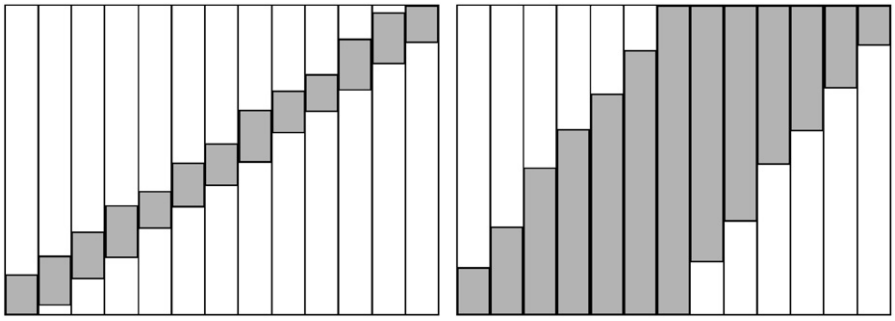

Fig. 1. Schematic depiction of the two sets of subgroups used in STEPP: sliding window (left) and tail-oriented (right). The horizontal axis indexes the various subpopulations for which treatment effects are estimated, and shows the range of covariate values (vertical axis) used to define the cohort of patients included in each subpopulation. The tail-oriented variant has the overall population as the centre group.

recruited between 1992 and 1997. In the first paper, based on 335 patients and 236 deaths, a 28% reduction in the risk of death in the interferon-α group was reported (Ritchie et al., 1999). No interaction was found in the trial report (Ritchie et al., 1999), but continuous covariates were not considered. Royston et al. (2004) used updated data with 322 deaths in 347 patients and found a significant interaction between treatment and WCC by using the MFPI procedure. Altogether they considered 10 potential predictive factors, of which six were continuous. Because of the high proportion of missing data, one continuous covariate (serum calcium) will not be considered, but all other data will be used here for further investigation.

2.2. Subpopulation treatment effect pattern plot (STEPP)

STEPP is based on dividing the observation into subgroups defined with respect to the covariate Z of interest, and estimating the treatment effect of T separately within each subpopulation. To increase the number of patients that contribute to each point estimate, subpopulations are allowed to overlap. This increases the precision of the individual estimates.

Two ways of defining subpopulations are proposed, as indicated in Fig. 1 by (a) and (b), respectively. The horizontal axis in Fig. 1 indexes the various subpopulations for which treatment effects are estimated, and shows the range of covariate values used to define the cohort of patients included in each subpopulation.

The TO variant has the overall population as the centre group. With increasing distance from the centre, more and more patients with high covariate values (to the left side) or low covariate values (to the right side) are deleted. Subpopulations in the SW variant have an overlapping part and a part that differs between neighbouring subpopulations. The number of subpopulations and the percentage of overlapping patients are important parameters of this variant. To define the size of the subpopulations the SW variant has two parameters, n_1 and n_2. A subpopulation must have at least n_2 patients, of which at least $(n_2 - n_1)$ patients are required to be different between neighbouring subpopulations. The amount of discreteness of the continuous variable determines the size of each subpopulation. The TO variant has a parameter g giving $(g - 1)$ subpopulations, where patients with larger values are eliminated and $(g - 1)$ subpopulations excluding patients with smaller values. For further details on how the subpopulations are created, see Bonetti and Gelber (2000).

The estimated treatment effects in the subpopulations defined by Z should be similar to the treatment effect in the overall population if Z does not modify the treatment effect, i.e. there is no interaction between Z and T. Plots showing the estimated treatment effect with confidence intervals (CIs) in the subpopulations, and tests based on the deviation of treatment effects in the subpopulations from those in the overall population, are suggested for the investigation of an interaction between Z and T. Each Z-based subgroup is represented by the mean Z. For more details see Bonetti and Gelber (2004).

2.3. MFPI

To investigate possible interactions between treatment and continuous covariates, Royston and Sauerbrei (2004) proposed the MFPI algorithm as an extension of the MFP algorithm (Sauerbrei and Royston, 1999; Royston and

Sauerbrei, 2005). MFP was proposed for building regression models, by combining variable selection with determination of functional forms for continuous predictors. Variables are selected by backward elimination. The algorithm investigates in a systemic way whether the effect of a continuous covariate is better modelled by a non-linear function from the class of fractional polynomials (FP) or by a linear function.

A FP function with two power terms ($FP2$) is a double-power model $\beta_1 X^{p_1} + \beta_2 X^{p_2}$ with the powers p_1 and p_2 chosen from a set $S = \{-2, -1.-0.5, 0, 0.5, 1, 2, 3\}$ where X^0 denotes $\log X$ (Royston and Altman, 1994). For $p_1 = p_2$ ('repeated-powers function') FP2 is defined as $\beta_1 X^{p_1} + \beta_2 X^{p_2} \log X$. This gives eight FP1 (including linear) and 36 FP2 functions.

The MFPI algorithm models the prognostic effect of Z by FP2 transformations within treatment groups, but under the constraint of the same powers. This can be done in a univariate setting, or by adjusting the model for other covariates. Assume that the covariate T has two levels, coded 1, 2. The influence of the covariate Z on the estimated treatment effect is determined by $t(Z) = \hat{f}_2(Z) - \hat{f}_1(Z)$ where $\hat{f}_i(Z)$, $i = 1, 2$ are the estimated functions for the prognostic effect of Z in treatment group i. The plot of $t(Z)$ together with a pointwise confidence band is called a treatment-effect plot. Comparing the model with separate functions for Z in treatment groups with a 'main' effects model with the same function in both groups is a test of interaction. The difference in deviances is compared with χ^2 on 2 d.f. For the investigation of an interaction of treatment with binary or categorical variables, MFPI uses the usual method of testing for an interaction in a model with main effects and multiplicative interaction terms.

MFPI allows adjustment for other variables in a multivariable setting in the context of different types of regression models. Royston and Sauerbrei (2004) propose to determine an 'adjustment' model in a preliminary step, preferably by MFP, without considering the covariate Z. For more details see Royston and Sauerbrei (2004).

2.4. Investigation of stability

(In)stability is a critical issue when working with flexible models Breiman (1996). In the particular case of estimating a treatment effect function for a continuous covariate, a small number of influential points may drive the function, thus indicating an interaction which is mainly a result of overfitting the data. To explore stability of MFPI, we applied bootstrap resampling as in Sauerbrei and Schumacher (1992), with extensions to handle transformations of continuous variables and to estimate the treatment effect function. Each observation of outcome and covariates was used as the sampling unit in the bootstrap analysis. We included as potential confounders all variables other than that under investigation (WCC), and used MFP to select the confounder model with a nominal significance level of 0.05. WCC is the only variable identified by MFPI as a predictive factor and the stability of the function chosen will be investigated, adjusting for the other variables. As in Royston and Sauerbrei (2003), we estimated a mean function and empirical bootstrap CIs. Bootstrap resampling was also used to investigate the stability of STEPP, but without adjusting for covariates. For each STEPP group, we computed the bootstrap mean and empirical 95% CI for the treatment effect, using SW and TO variants.

2.5. Type I error of MFPI

To investigate the type I error of the MFPI procedure, we used the renal cancer data. The observed survival time and treatment variables were together permuted at random, and interaction with the continuous variable haemoglobin was assessed. In this way, the situation that the continuous variable and the treatment are independent was simulated. In this context the distribution of P-values from a test of interaction has a uniform distribution and the percentage of P-value < 0.05 is an estimate of the actual type I error probability for the 0.05 nominal significance level.

Investigation of the type II error would require a large simulation study, in which different functional forms for the treatment effect must be considered. This is beyond the scope of this paper. Here we will check whether STEPP, with its greater flexibility, points to possible interactions of continuous variables not identified by MFPI. The analysis in Royston et al. (2004) identified WCC as the only predictive variable, the other four continuous variables, three binary variables and WHO performance status (three categories) not seeming to modify the treatment effect. For the continuous variables we calculate P-values of the test for an interaction and compare the function graphically with the equivalent functions from a STEPP analysis using the TO variant.

3. Results

3.1. Estimation of treatment effect function

Inspection of the distribution of values of WCC showed several large and potentially influential outliers. To improve the robustness of FP analysis performed by the MFPI algorithm, we applied to WCC the preliminary transformation $g_\delta(.)$ proposed by Royston and Sauerbrei (2006). The latter function is linear in the central portion of the distribution, but pulls in the tails and greatly reduces the leverage of extreme observations. Fig. 2 shows the treatment effect function for WCC from the MFPI analysis using g_δ(WCC). A likelihood ratio test gives a significant test result ($P = 0.03$) for an interaction between treatment and an FP2 function of g_δ(WCC). The treatment effect function (see Fig. 2) indicates a substantially reduced risk in the interferon-α group for very small WCC values, a similar risk for WCC around 10 and a slightly increased risk in the interferon-α group for large values of WCC. However, the CI includes zero for all WCC values above 10.

Fig. 3 presents the results from several STEPP analyses of WCC. It is clear that use of small subpopulations ($n_1 = 25$, $n_2 = 40$; upper left-hand plot) with the SW method results in considerable variation caused by overfitting the data. Increasing the sample size in each subpopulation reduces the variation and leads to treatment estimates which show a similar dependence on WCC. For example, for $n_1 = 50$, $n_2 = 80$, there is only one additional 'blip' for WCC around 7. The lower panel clearly indicates that results from the TO variant are less noisy and hence easier to interpret. The plot with $g = 4$ (lower right-hand plot) can be viewed as a rough approximation to the treatment effect function from MFPI (Fig. 2).

To check whether the interaction with WCC is an artefact of MFPI, Royston and Sauerbrei (2004) proposed an investigation in four subgroups defined by WCC. We use cutpoints of 6.5, 8 and $10 \times 10^9 1^{-1}$. The value of 10 was chosen because the beneficial effect of IFN disappears at about this point, whereas the other two cutpoints represent the first two quartiles of the distribution of WCC. Fig. 4 shows the estimated survival curves by treatment in the four subgroups of WCC. In accordance with the function shown in Fig. 2, a trend is seen towards a substantial survival advantage of interferon-α in group I with the lowest WCC, becoming weaker in subgroups II and III with larger counts, and no longer present in group IV. Estimated hazard ratios for the effect of interferon-α in comparison to MPA (with 95% CI) are 0.53 (0.34–0.83), 0.69 (0.44–1.07), 0.89 (0.57–1.37) and 1.32 (0.85–2.05) in WCC groups I–IV, respectively.

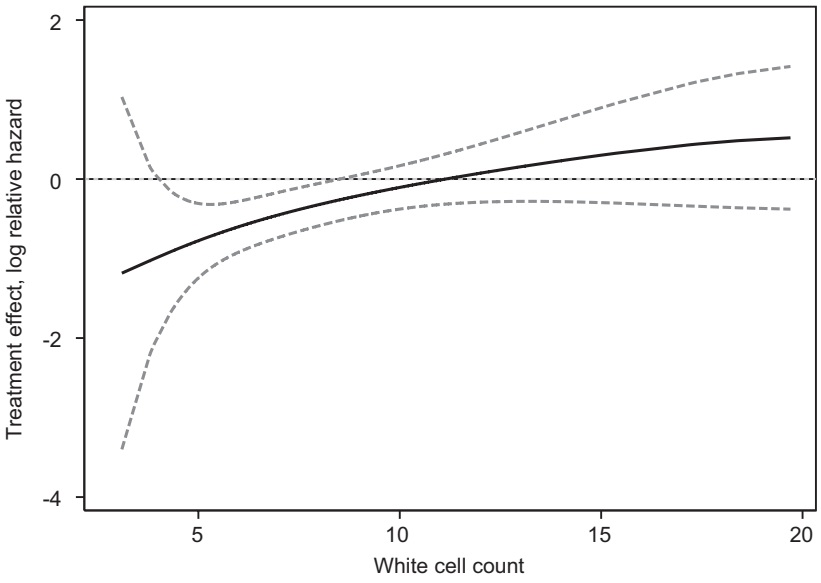

Fig. 2. Treatment effect plot for white cell count from the MFPI analysis of the renal cancer data. Estimated treatment effect with pointwise 95% confidence interval.

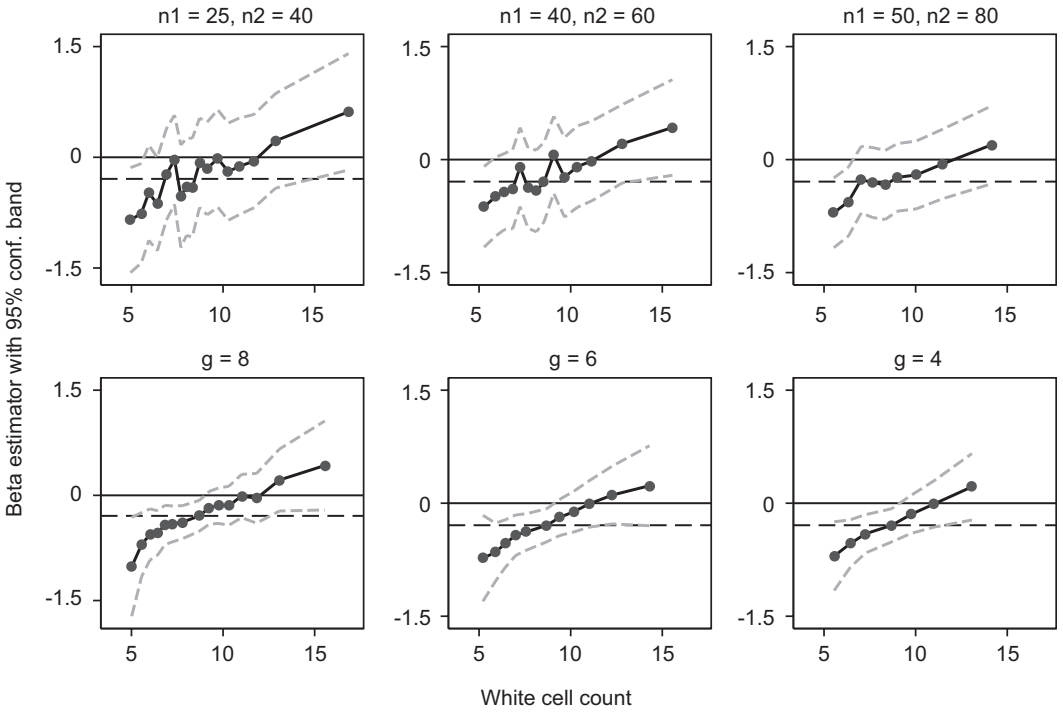

Fig. 3. STEPP plots for WCC in the renal data, constructed with several choices of parameter values. Upper panel: sliding window; lower panel, tail-oriented. The plotted points denote the estimated treatment effects in each subgroup, with corresponding 95% confidence intervals (faint dashed lines). The solid horizontal lines at zero represent no treatment effect and the dashed lines the estimated overall treatment effect.

3.2. Stability analysis

We checked the stability of the MFPI-derived treatment effect function and of the STEPP functions in 1000 bootstrap samples. For the MFPI analyses, we first selected the multivariable adjustment model and then the FP2 treatment effect function for WCC (with preliminary transformation). Selecting a new adjustment model in each bootstrap replication will increase instability. The STEPP analyses were restricted to unadjusted models. We present the result of 20 random sample curves from MFPI in the left panel of Fig. 5. Most of the individual curves are similar to the curve from the original analysis presented in Fig. 2. The mean of the 1000 bootstrap replications gives a nearly identical curve, with small differences appearing for more extreme values (WCC < 5 or > 15). The estimated effects from 11 subpopulations using the TO variant of STEPP with $g=6$ agree very closely with the functions from MFPI. For the bulk of the distribution of WCC values the 95% pointwise CI derived from the 1000 bootstrap replications is a little wider than the interval from the original analysis. For larger values (say, > 12), the data becomes sparse and the bootstrap intervals become much wider, reflecting greater uncertainty in the FP2 functions selected by MFPI.

Fig. 6 presents a random sample of 20 curves from 1000 bootstrap replications by using STEPP with $n_1 = 40$, $n_2 = 60$ (SW) or $g = 6$ (TO). A large amount of instability is apparent for the SW variant. Functions for the TO variant are more variable than the functions from MFPI, but the trend corresponding to the result from the original analysis stands out clearly. For TO, the bootstrap interval is smaller than the equivalent interval from MFPI, partly a consequence of the additional variation introduced in the MFPI analyses by selecting the adjustment model. Note that with STEPP the range of the covariate is restricted in the tails by the grouping procedure. Considering the treatment effect function outside the range of the mean covariate values in the extreme groups would be extrapolation.

3.3. Type I error of MFPI

The situation that the continuous variable and the treatment are independent was simulated using random permutation, as described above. The distribution of P-values from 1000 simulations of a test of interaction is close to uniform

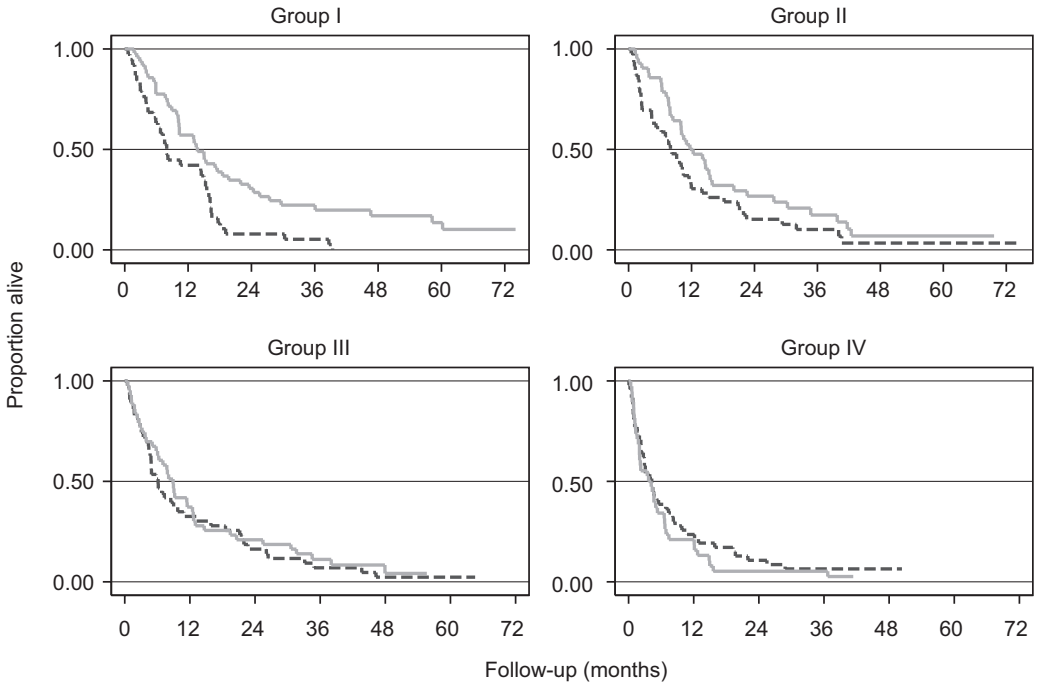

Fig. 4. Kaplan–Meier plots by treatment group in four subgroups of the renal cancer data, defined by WCC. Solid grey lines denote interferon, dashed black lines MPA. Cutpoints on WCC used: 6.5, 8, 10.

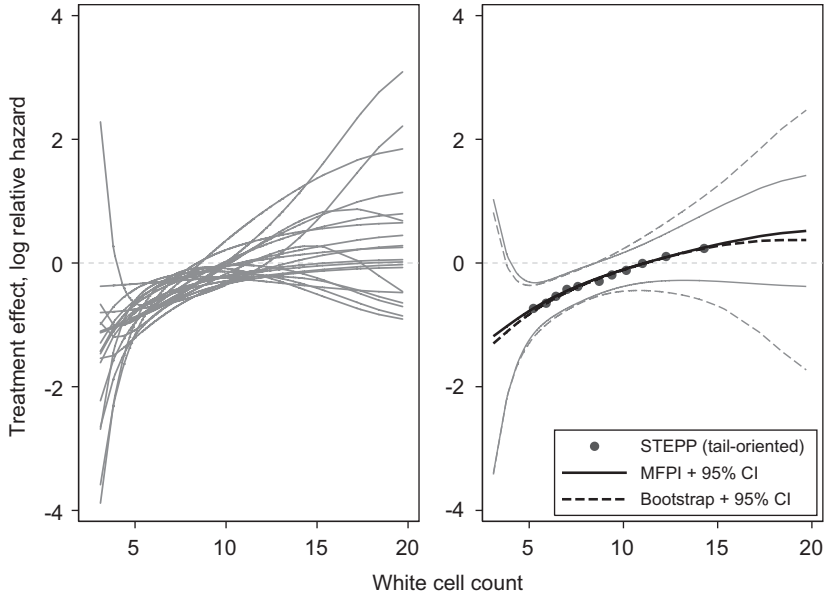

Fig. 5. Bootstrap analysis. Left panel: treatment effects plots in 20 bootstrap replications, using MFPI with data-driven adjustment model. Right panel: mean and 95% CI of treatment effect function from 1000 bootstrap replications, also showing treatment function on original data with 95% CI, and tail-oriented STEPP function.

(data not shown). The P-value of an Anderson–Darling test of a standard uniform distribution is 0.51. In 54 of the 1000 permutations the P-value is < 0.05, showing that the type I error of the MFPI procedure is close to its nominal level.

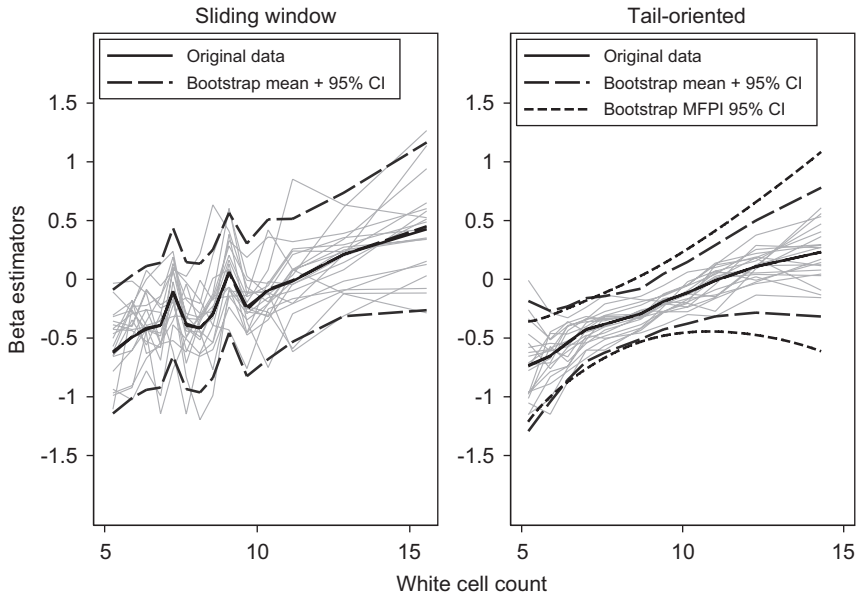

Fig. 6. Results from 1000 bootstrap replications of STEPP analysis of WCC. Left panel: sliding window ($m = 40$, $n = 60$); right panel, tail-oriented ($g = 6$). Thick lines represent the original data, bootstrap mean and 95% CI. Thin lines are results from 20 bootstrap replications selected at random.

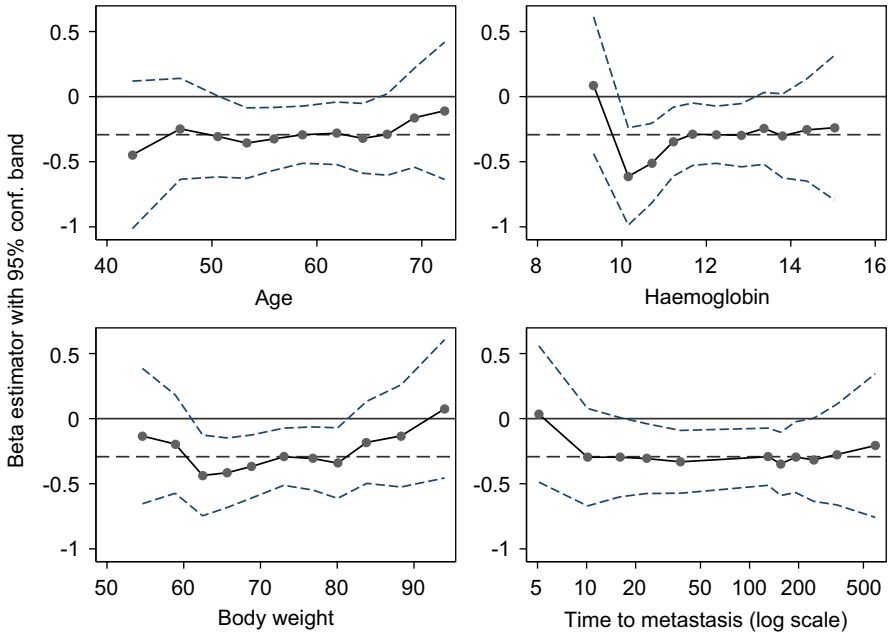

Fig. 7. STEPP analyses for the four continuous variables in the renal data which do not apparently interact with treatment ($P > 0.5$) according to MFPI analysis. Tail-oriented with $g = 6$. For other details, see legend for Fig. 3.

For the four continuous variables not identified by MFPI as interacting with treatment, we plot the treatment effect functions and results from the TO variant of STEPP with $g = 6$ in Fig. 7. The P-values for the four tests of interaction from MFPI are all > 0.5, and also for none of these variables is there any evidence of a trend towards an interaction. In principle these results are confirmed by the functions from a STEPP analysis using the TO variant, which are more

difficult to interpret. They point to some possible interactions in smaller subpopulations. Other apparent subpopulations may be identified by changing g or by considering results from the SW variant.

4. Discussion

We have shown that both MFPI and STEPP facilitate an informative investigation of interaction between treatment and a continuous covariate. The MFPI algorithm estimates a continuous treatment effect function which is nonlinear. It tests for an interaction by comparing models using separate functions in treatment groups with a 'main' effects model with the same function in all groups (Royston and Sauerbrei, 2004), essentially an analysis of covariance. In contrast, STEPP estimates a treatment effect in a series of subpopulations with overlapping groups of patients. Two variants of STEPP, SW and TO, have been suggested. Each variant requires an important decision concerning the number(s) of subpopulations to be considered. Naturally, that is a very crucial parameter determining the complexity and (in)stability of the resulting plot. Related tests of the null hypothesis of no differential treatment effects have been proposed (Bonetti and Gelber, 2000, 2004), also depending crucially on the number of groups. As no guidance for the preferred test and number of groups is available, we did not consider this aspect here.

We have compared MFPI and STEPP in data from a randomized trial with several potential continuous predictive factors. We concentrated on the only predictive factor (WCC) identified by MFPI in an earlier analysis (Royston et al., 2004). If the TO variant of STEPP is used with a smaller number of subgroups, the results of MFPI and STEPP agreed remarkably closely. Using the bootstrap we showed that the estimated treatment effect functions are relatively stable, in agreement with a more detailed analysis of determining multivariable models with continuous predictors (Royston and Sauerbrei, 2004). The mean of 1000 bootstrap replications was nearly identical to the function from the original analysis, confirming that the original function was not an artefact arising from complex model building. This was confirmed through checks in four ordered subgroups based on the distribution of WCC. In contrast to the subgroups used in STEPP, these subgroups did not overlap. The marked similarity between the treatment effect functions from MFPI and the TO variant of STEPP (see Fig. 5) provides another check of MFPI. Furthermore, the bootstrap analysis confirmed that CIs from models determined data-dependently by MFPI are too narrow. According to a small simulation study, the type I error of the MFPI test of interaction is very close to its nominal level.

It is encouraging to see the close agreement between MFPI and a carefully 'tuned' flavour of STEPP. According to our investigations in a relatively small study ($n = 347$), the SW variant seriously overfits the data and is too unstable for practical use. Of course the severity of these problems depends on the amount of overlap between the subpopulations. The TO variant displays the treatment effect of the overall population in the centre of the covariate distribution and shows how much it deviates as more and more patients are eliminated from a subgroup. Using $g = 6$ for the tuning parameter of the TO variant, we obtained results in close agreement with the treatment effect function from MFPI. The bootstrap replications also show substantial stability. Without an MFPI analysis, we would have come to the same conclusion about predictive factors in the renal cancer study if we had used STEPP in the first place, particularly if the TO variant with $g = 6$ had been chosen. MFPI gives a smooth function for this effect and the type I error for the test for interaction is near to its nominal value. Similar investigations are still needed for STEPP.

We used the MFPI algorithm exactly as proposed in our original paper (Royston and Sauerbrei, 2004). However, further refinements of the procedure are possible. The least flexible option would be use of linear functions, which is what is often done. Alternatively, FP1 functions could be used. Greater flexibility might involve allowing the powers to be different in the treatment groups. However, except in very large samples, the advantage of increased flexibility would probably be outweighed by the increased instability. More experience is needed, both in analyses of real data sets and in simulation studies, to examine these issues further. Note that STEPP, when used with many cutpoints, is extremely flexible and may appear to show complex treatment effect patterns (see for example the top left panel of Fig. 3) that could be artefactual. Even with a small number of subpopulations, the TO variant may suggest evidence for interaction not detected by MFPI (see Fig. 7). However, such fluctuations in the estimated function may simply be the result of overfitting.

STEPP has gained some popularity in recent years (at least in breast cancer research). Besides the methodological papers (Bonetti and Gelber, 2000, 2004) where breast cancer was used as an example, it has also been used by some of the leading breast cancer study groups, e.g. the IBCSG (Crivellari et al., 2003) and the NSABP (Fisher et al., 2004). With the current trend towards much larger clinical trials, investigations of interactions will become more important and will have more power. The serious problems of categorizing continuous data in the investigation of prognostic factors

are well-known (Altman et al., 1994; Royston et al., 2006). Investigations using the full information from continuous data are becoming more popular. We have illustrated similar issues (Royston and Sauerbrei, 2004) for predictive factors, and proposed MFPI as a way to improve the investigation of predictive factors. We hope that the promising results from this paper will increase the use of MFPI, helping to identify factors relevant to determining rational, evidence-based treatment policies.

References

Altman, D.G., Lausen, B., Sauerbrei, W., Schumacher, M., 1994. The dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. Journal of the National Cancer Institute 86, 829–835.

Assmann, S.F., Pocock, S.J., Enos, L.E., Kasten, L.E., 2000. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 355, 1064–1069.

Breiman, L., 1996. The heuristics of instability in model selection. Ann. Statist. 24, 2350–2381.

Bonetti, M., Gelber, R.D., 2000. A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. Statist. Med. 19, 2595–2609.

Bonetti, M., Gelber, R.D., 2004. Patterns of treatment effects in subsets of patients in clinical trials. Biostatistics 5, 465–481.

Crivellari, D., Price, K., Gelber, R.D., Castiglione-Gertsch, M., Rudenstam, C.-M., Lindtner, J., Fey, M.F., Senn, H.-J., Coates, A.S., Collins, J., Goldhirsch, A., 2003. Adjuvant endocrine therapy compared with no systemic therapy for elderly women with early breast cancer: 21-year results of International Breast Cancer Study Group trial IV. J. Clinical Oncol. 21, 4517–4523.

Fisher, B., Jeong, J.-H., Bryant, J., Anderson, S., Dignam, J., Fisher, E.E.R., Wolmark, N., 2004. Treatment of lymph-node-negative, oestrogen-receptor positive breast cancer: long-term findings from National Surgical Adjuvant Breast and Bowel Project randomised clinical trials. Lancet 364, 858–868.

Kehl, V., Ulm, K., 2006. Responder identification in clinical trials with sensored data. Comput. Stat. Data Anal. 50, 1338–1355.

Ritchie, A., Griffiths, G., Parmar, M., for the MRC Renal Cancer Collaborators, 1999. Interferon-α and survival in metastatic renal carcinoma: early results of a randomised controlled trial. Lancet, 353 14–17.

Royston, P., Altman, D.G., 1994. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). Appl. Statist. 43 (3), 429–467.

Royston, P., Sauerbrei, W., 2003. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. Statist. Med. 22, 639–659.

Royston, P., Sauerbrei, W., 2004. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Statist. Med. 23, 2509–2525.

Royston, P., Sauerbrei, W., 2005. Building multivariable regression models with continuous covariates, with a practical emphasis on fractional polynomials and applications in clinical epidemiology. Methods of Information in Medicine 44, 561–571.

Royston, P., Sauerbrei, W., 2006. Improving the robustness of fractional polynomial models by preliminary covariate transformation: a pragmatic approach. Comput. Statist. Data Anal., in press, doi:10.1016/j.csda.2006.05.006.

Royston, P., Sauerbrei, W., Ritchie, A.W.S., 2004. Is treatment with interferon-α effective in all patients with metastatic renal carcinoma? A new approach to the investigation of interactions. British J. Cancer 23, 794–799.

Royston, P., Altman, D.G., Sauerbrei, W., 2006. Dichotomizing continuous predictors in multiple regression: a bad idea. Statist. Med. 25, 127–141.

Sauerbrei, W., Royston, P., 1999. Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. J. Roy. Statist. Soc. Ser., A 162, 71–94 (Corrigendum: J. Roy. Statist. Soc. (Ser. A) 165: 399–400, 2002).

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