

## Analysing covariates with spike at zero: A modified FP procedure and conceptual issues

Heiko Becher<sup>\*,1</sup>, Eva Lorenz<sup>1</sup>, Patrick Royston<sup>2</sup>, and Willi Sauerbrei<sup>3</sup>

<sup>1</sup> Institute of Public Health, Medical Faculty, University of Heidelberg, Im Neuenheimer Feld 324, 69120 Heidelberg, Germany

<sup>2</sup> Hub for Trials Methodology Research, MRC Clinical Trials Unit and University College London, Aviation House, 125 Kingsway, London WC2B 6NH, UK

<sup>3</sup> IMBI, Freiburg University Medical Center, Stefan-Meier-Str. 26, 79100, Freiburg, Germany

Received 23 December 2011; revised 14 May 2012; accepted 14 May 2012

In epidemiology and in clinical research, risk factors often have special distributions. A common situation is that a proportion of individuals have exposure zero, and among those exposed, we have some continuous distribution. We call this a ‘spike at zero’. Examples for this are smoking, duration of breastfeeding, or alcohol consumption. Furthermore, the empirical distribution of laboratory values and other measurements may have a semi-continuous distribution as a result of the lower detection limit of the measurement. To model the dose–response function, an extension of the fractional polynomial approach was recently proposed. In this paper, we suggest a modification of the previously suggested FP procedure. We first give the theoretical justification of this modified procedure by investigating relevant distribution classes. Here, we systematically derive the theoretical shapes of dose–response curves under given distributional assumptions (normal, log normal, gamma) in the framework of a logistic regression model. Further, we check the performance of the procedure in a simulation study and compare it to the previously suggested method, and finally we illustrate the procedures with data from a case–control study on breast cancer.

*Keywords:* Dose–response model; Fraction unexposed; Fractional polynomials; Regression modelling.

### 1 Introduction

Covariables in medical research often show a biphasic distribution where most of the data are continuously distributed over a wide range but a proportion of the distribution is concentrated (‘spiked’) at one value. This is called a semi-continuous variable (Olsen and Schafer, 2001), a spike at zero (Robertson et al., 1994; Schisterman et al., 2006), mass at zero (Lachenbruch, 2001) or clump of zeros (Hallstrom, 2010). This type of distribution arises also for covariables that are based on laboratory measurements, where the analytical method is restricted by a limit of detection or limit of quantification. Typical examples, for example, in cancer or cardiovascular disease epidemiology are occupational exposures, for example, asbestos exposure, or alcohol and tobacco consumption where a proportion of individuals may be completely unexposed, and the exposure of those who had been exposed follows a continuous (positive) distribution. When assessing a dose–response relationship for this type of covariable, a simple solution would be to restrict the analysis to those exposed however, this has two major

\*Corresponding author: e-mail: heiko.becher@urz.uni-heidelberg.de, Phone: +49-6221-565031, Fax: +49-6221-565948

drawbacks: The sample size may be considerably reduced, and the effect of those exposed cannot be compared with those who are unexposed.

For analysing dose–response relationships, fractional polynomials (FPs) have been shown to be a useful and powerful tool (Royston and Sauerbrei, 2008; Royston et al., 2010). Here, the continuous covariable  $X$  undergoes a set of transformations, and the model selection is based on goodness-of-fit statistics according to a well-defined procedure which maintains the nominal  $\alpha$ -level, as described in Royston and Sauerbrei (2008, pp. 82–84). If zero or negative values of  $X$  can occur, a preliminary transformation of  $X$  to ensure positivity is needed for this procedure. We do not consider the case of negative values here since in the majority of practical applications, zero is the smallest observed value. For zero values, the suggested solution in Royston and Sauerbrei (2008) and earlier papers is to choose a non-zero origin and work with  $\tilde{x} = x + \gamma$ , where  $\gamma$  is, for example, the minimum increment between successive ordered sample values of  $X$ . Other choices of gamma are possible (Royston and Altman, 1994; Royston et al., 2010).

In earlier research, the correct model has been investigated under some specific distributional assumptions (Becher, 1992; Robertson et al., 1994). For example, if  $X$  is normally distributed into diseased and non-diseased with means  $\mu_1$  and  $\mu_0$  and equal variance  $\sigma^2$ , then the correct model requires  $X$  to be included untransformed in the model. If the variances differ, the FP term  $x^2$  is required. To our knowledge, the spike at zero situation has not yet been investigated theoretically. The standard FP approach (Royston and Altman, 1994), as well as the spline techniques (de Boer, 2001), do not specifically consider the spike at zero situation. Recently, we have suggested a method based on FPs to deal with the spike at zero situation in model fitting (Royston and Sauerbrei, 2008; Royston et al., 2010). In this method, a binary variable  $Z$ ,  $Z = \begin{cases} 1 & \text{if } x=0 \\ 0 & \text{if } x>0 \end{cases}$  is added as an additional variable, and a procedure to select the model has been developed. The procedure maintains the need to transform  $X$  to ensure positivity since according to this procedure the FP transformations are applied to all observations, including those with  $X = 0$ . In this previous work, we developed a function selection procedure with spike, here denoted FP-spike, which has two stages to select a model. In the first stage, the most complex model comprising  $Z$  and  $FP2 + (\tilde{x}, p_1, p_2)$  is compared with the null model on 5 d.f. (4 d.f. from the FP2 model plus one from the binary  $Z$  term). If the corresponding  $\chi^2$ -test is significant, the steps of the usual FP for selecting an FP function are followed, but with  $Z$  always included in the model. In the second stage (performed separately),  $Z$  and the selected FP function (may be linear) are each tested for removal from the model. If both parts are significant, the final model includes both; if one or both parts are non-significant adjusted for the other, the one with the larger  $p$ -value is removed. In the latter case, the final model comprises either the binary dummy variable or the selected FP function. If the binary variable is removed in the second stage, the spike at zero plays no specific role, a usual FP function is selected. However, since the selection of the FP terms in the first stage may have been affected by the presence of the binary dummy variable, the resulting model may differ from that from a standard FP analysis, because different power terms may have been chosen.

In this paper, we derive correct dose–response curves under several specific assumptions and suggest a modification of FP-spike. The modified version avoids the bias caused by the pre-transformation (adding  $\gamma$ ) of FP-spike.

The structure of the paper is as follows: First, we investigate the correct dose–response curve for a spike at zero situation under some specific assumptions for the distribution of the positive part of the covariable  $X$ . As a result, we propose a modification of FP-spike. Second, we choose a specific case, the log normal distribution, with two different parameter combinations. The effect of  $\gamma$  on selecting FP functions will be investigated in a simulation study. We will consider the standard FP-procedure, FP-spike and the modification of the latter. In addition, we use data from a case–control study to illustrate and compare the approaches to derive a dose–response model for covariates with a spike at zero.

**Table 1** Distribution of alcohol consumption and univariate odds ratios for breast cancer among participants in a population-based case–control study, Germany, 1992–1995. In the original paper, adjusted odds ratios were presented.

Average ethanol intake (g/day)	Controls		Cases		Odds ratio	95% CI
	<i>n</i>	Percent	<i>n</i>	Percent		
0 (non-drinker)	239	17.3	158	21.7	1.00*	–
1–5	577	41.8	257	36.4	0.7	0.54, 0.90
6–11	295	21.4	124	17.6	0.66	0.49, 0.89
12–18	150	10.9	69	9.8	0.72	0.50, 1.03
19–30	84	6.1	59	8.4	1.1	0.73, 1.65
31+	36	2.6	44	6.2	1.91	1.14, 3.20
	1381	100.0	706	100.0		

\*Reference category.

### 1.1 Data example

We illustrate the approaches using data from a case–control study on pre-menopausal breast cancer which was performed in south-west Germany in 1992 to 1995 (Kropp *et al.*, 2001). The aim was to investigate the relation between alcohol consumption and breast cancer. The study had 706 cases and 1381 controls frequency matched by age and study region. Table 1 gives the average daily alcohol consumption by dose category, with adjusted odds ratios by category indicating a non-monotone dose–response relation. A large number of cases (21.7%) and controls (17.3%) reported zero exposure. Table 1 shows the data.

## 2 Deriving dose–response curves

### 2.1 General

We consider a binary response variable and the logistic regression model. Extensions to other models are discussed in Section 5. Let  $Y$  be a binary disease indicator which may take values 1 (diseased) or 0 (not diseased) and let  $X$  be the covariable of interest which has a continuous distribution with a spike at zero. We denote the corresponding probabilities with  $p_i = P(X = 0 | Y = i)$ . For the non-zero part of  $X$ , we denote the distribution of  $X$  as  $f_{X|X>0}$ , so that the density function of  $X$  can be given as  $f_{X|Y=i} = p_i + (1 - p_i)f_{X|X>0, Y=i}$  which is in the following simply denoted by  $f_i$ ,  $i = 0, 1$ . If we model the odds ratio  $OR_{X=x^* \text{ vs } X=x_0}$ ,  $x_0, x^* \neq 0$  as given in Becher (1992), we get

$$\begin{aligned}
 OR_{X=x^* \text{ vs } X=x_0} &= \frac{f_1(x^*)f_0(x_0)}{f_1(x_0)f_0(x^*)} = \frac{(1 - p_1)f_{X|X>0, Y=1}(x^*)(1 - p_0)f_{X|X>0, Y=0}(x_0)}{(1 - p_1)f_{X|X>0, Y=1}(x_0)(1 - p_0)f_{X|X>0, Y=0}(x^*)} \\
 &= \frac{f_{X|X>0, Y=1}(x^*)f_{X|X>0, Y=0}(x_0)}{f_{X|X>0, Y=1}(x_0)f_{X|X>0, Y=0}(x^*)}
 \end{aligned} \tag{1}$$

and if  $x_0 = 0$ ,  $x^* \neq 0$ , we get

$$OR_{X=x^* \text{ vs } X=0} = \frac{f_1(x^*)f_0(0)}{f_1(0)f_0(x^*)} = \frac{p_0(1 - p_1)f_{X|X>0, Y=1}(x^*)}{p_1(1 - p_0)f_{X|X>0, Y=0}(x^*)}. \tag{2}$$

Thus, the odds ratio for two non-zero values,  $OR_{X=x^* \text{ vs } X=x_0}$ , is independent of the spike. The odds ratio for a non-zero value  $x^*$  compared to zero,  $OR_{X=x^* \text{ vs } X=0}$ , is a function of the spike probabilities multiplied by the ratio of the density functions of diseased and non-diseased at  $X = x^*$ .

In practical applications, the non-zero part of  $X$  is usually positive (e.g., dose of an exposure variable); therefore, we chose the notation  $f_{X|X>0}$ . In theory, however, more general consideration is possible. See also the normal distribution as the first example below.

**2.2 Some analytical results**

We will now derive the OR when the distribution of the covariable  $X$  follows some standard distributions. These include the normal distribution, the log normal distribution, and the gamma distribution.

**2.2.1 Normal distribution**

Let

$$f_1(x) = \begin{cases} p_1 & x = 0 \\ (1 - p_1)\varphi_{\mu_1, \sigma}(x) & \text{if } x \neq 0 \end{cases}$$

and

$$f_0(x) = \begin{cases} p_0 & x = 0 \\ (1 - p_0)\varphi_{\mu_0, \sigma}(x) & \text{if } x \neq 0, \end{cases}$$

where  $\varphi_{\mu, \sigma}$  is the probability density function of a normal distributed variable with mean  $\mu_0$  in non-diseased and  $\mu_1$  in diseased and equal variance  $\sigma^2$ .

If  $x_0, x^* \neq 0$ , we get

$$OR_{X=x^* \text{ vs } X=x_0} = \exp\left(\frac{\mu_1 - \mu_0}{\sigma^2}(x^* - x_0)\right) \tag{3}$$

and if  $x_0 = 0, x^* \neq 0$ , we get

$$\begin{aligned} OR_{X=x^* \text{ vs } X=0} &= \frac{f_1(x^*)f_0(0)}{f_1(0)f_0(x^*)} = \frac{p_0(1 - p_1)}{p_1(1 - p_0)} \exp\left(\frac{\mu_1 - \mu_0}{\sigma^2}x^* + \frac{\mu_0^2 - \mu_1^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{\mu_1 - \mu_0}{\sigma^2}x^* + \frac{\mu_0^2 - \mu_1^2}{2\sigma^2} + \ln p_0 - \ln p_1 + \ln(1 - p_1) - \ln(1 - p_0)\right). \end{aligned} \tag{4}$$

Thus, the correct model requires  $X$  untransformed and a binary variable as indicator for  $X > 0$  and gives

$$OR_{X=x^* \text{ vs } X=0} = \exp(\beta_0 + \beta_1 x^*). \tag{5}$$

The coefficients are  $\beta_1 = \frac{\mu_1 - \mu_0}{\sigma^2}$  and  $\beta_0 = \frac{\mu_0^2 - \mu_1^2}{2\sigma^2} + \ln\left(\frac{p_0(1 - p_1)}{p_1(1 - p_0)}\right)$ . If the variances are unequal, some algebra shows that the correct model requires  $X$  untransformed,  $X$  squared, and a binary variable as indicator for  $X > 0$  as

$$OR_{X=x^* \text{ vs } X=0} = \exp(\beta_0 + \beta_1 x^* + \beta_2 x^{*2}). \tag{6}$$

The coefficients are  $\beta_1 = \frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}$ ,  $\beta_2 = \frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}$ ,  $\beta_0 = \frac{\mu_0^2}{2\sigma_0^2} - \frac{\mu_1^2}{2\sigma_1^2} + \ln\left(\frac{p_0(1-p_1)}{p_1(1-p_0)}\right)$ . For  $x_0, x^* \neq 0$ , we get

$$OR_{X=x^* \text{ vs } X=x_0} = \exp(\beta_1(x^* - x_0) + \beta_2(x^{*2} - x_0^2)) \quad (7)$$

with the same value for  $\beta_1$  and  $\beta_2$ .

### 2.2.2 Log normal distribution

When a log normal distribution is assumed the OR can be calculated in a similar way. The log normal distribution has two parameters,  $\mu$  and  $\sigma$ . If  $\sigma$  is equal in non-diseased and in diseased, the correct model requires the log-transformed positive part of the variable and a binary variable as indicator for  $x > 0$ . If  $x_0, x^* \neq 0$ , we get

$$OR_{X=x^* \text{ vs } X=x_0} = \exp\left(\frac{\mu_1 - \mu_0}{\sigma^2}(\ln x^* - \ln x_0)\right), \quad (8)$$

and if  $x_0 = 0$ , we get

$$OR_{X=x^* \text{ vs } X=0} = \exp(\beta_0 + \beta_1 \ln(x^*)) \quad (9)$$

with  $\beta_1 = \frac{\mu_1 - \mu_0}{\sigma^2}$  and  $\beta_0 = \frac{\mu_0^2 - \mu_1^2}{2\sigma^2} + \ln\left(\frac{p_0(1-p_1)}{p_1(1-p_0)}\right)$ . If  $\sigma^2$  is different in cases and controls, we get for  $x_0, x^* \neq 0$

$$OR_{X=x^* \text{ vs } X=x_0} = \exp\left(\left[\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}\right](\ln x^* - \ln x_0) + \left[\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right]((\ln x^*)^2 - (\ln x_0)^2)\right), \quad (10)$$

and if  $x_0 = 0$ , we get

$$OR_{X=x^* \text{ vs } X=0} = \exp(\beta_0 + \beta_1 \ln x^* + \beta_2 (\ln x^*)^2) \quad (11)$$

with  $\beta_0 = \ln\left(\frac{p_0(1-p_1)}{p_1(1-p_0)}\right) + \ln\left(\frac{\sigma_0^2}{\sigma_1^2}\right) + \frac{\mu_0^2}{2\sigma_0^2} - \frac{\mu_1^2}{2\sigma_1^2}$ ,  $\beta_1 = \frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}$  and  $\beta_2 = \frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}$ . Thus, the correct model requires terms in  $\ln x$  and  $(\ln x)^2$  and the binary indicator. This distribution may be appropriate for some common situations in epidemiology, like dose of alcohol intake (see example). Note that the log transformation is applied only to the non-zero part of the covariable  $X$ .

### 2.2.3 Gamma distribution

If the continuous part is gamma distributed, we get the density

$$f_i(x) = \begin{cases} p_i & x = 0 \\ (1 - p_i) \frac{\lambda_i^{q_i}}{\Gamma(q_i)} e^{-\lambda_i x} x^{q_i-1} & \text{if } x \neq 0 \end{cases}$$

with parameters  $\lambda_i$  and  $q_i$ . The same sort of algebra shows that the model requires  $x$  if the parameter  $\lambda$  is different in diseased and non-diseased and  $\ln x$  if the parameter  $q$  is different in diseased and non-diseased, and the binary indicator. In the special case of  $q_i = 1$ , we have an exponential distribution and get

$$OR_{X=x^* \text{ vs } X=x_0} = \exp((x^* - x_0)(\lambda_0 - \lambda_1)) \quad (12)$$

and

$$OR_{X=x^* \text{ vs } X=0} = \exp(\beta_0 + \beta_1 x^*) \quad (13)$$

with coefficients  $\beta_0 = \ln\left(\frac{p_0(1-p_1)}{p_1(1-p_0)}\right) + \ln(\lambda_1) - \ln(\lambda_0)$  and  $\beta_1 = \lambda_0 - \lambda_1$ . Thus, the correct model requires the binary indicator plus  $X$  untransformed.

### 2.3 Some properties and computational aspects

Some useful results follow directly from the above sections

- From the general result in Subsection 2.1, we get

$$\ln(OR_{X=x^* \text{ vs } X=0}) = \ln(OR_{X>0 \text{ vs } X=0}) + \ln \frac{f_{X|X>0|Y=1}(x^*)}{f_{X|X>0|Y=0}(x^*)}.$$

Therefore, if the distribution of  $X$  among the exposed is identical in cases and controls, the OR in comparison to  $X = 0$  is just the OR for exposed versus non-exposed.

- If the proportion of non-exposed is equal in cases and controls,  $OR_{X>0 \text{ vs } X=0} = 1$ , it does not follow that the binary indicator cancels out. To see this, consider, for example, the log normal distribution with  $\sigma_0 = \sigma_1$ . Here, we get  $\beta_0 = \frac{\mu_0^2 - \mu_1^2}{2\sigma^2}$ .
- There are conditions for which the binary factor cancels out. In the above example with  $\sigma = 1$  this holds when  $\frac{\mu_0^2 - \mu_1^2}{2\sigma^2} + \ln\left(\frac{p_0(1-p_1)}{p_1(1-p_0)}\right)$ . Another example is parameter combination A in the simulation below. A numerical example is

$$\ln OR_{X>0 \text{ vs } X=0} = 0.5, \mu_0 = 0, \mu_1 = 1.$$

For this we get

$$OR_{X=x^* \text{ vs } X=0} = \ln(x^*)$$

and

$$OR_{X=x^* \text{ vs } X=x_0} = \ln(x^*) - \ln(x_0).$$

Here,  $x_0 = 1.648$  (the expected value for  $Y = 0$ ). See the simulation study for implications in model fitting.

- If there is a spike at zero with  $OR_{X>0 \text{ vs } X=0} > 1$  and  $OR_{X=x^* \text{ vs } X=x_0} > 1$  for  $x^* - x_0 > 0$  then it follows that it exists  $\tilde{x}$  such that  $OR_{X=\tilde{x} \text{ vs } X=0} < 1$ .

The results in Subsection 2.2 are not restricted to the case-control design. For example, in a cross-sectional study, the prevalence odds is modelled, and the same results apply.

### 2.4 The FP procedure for the spike at zero situation

The results in Subsection 2.2 show that there is no theoretical justification to include the parameter  $\gamma$ , and that the transformation should be applied to the positive values only. In the following, we consider a spike at zero situation where the non-zero part is strictly positive. The FP-spike procedure of Royston et al. (2010) is given by

- Generate binary  $z$  variable (exposure yes/no)
- *First Stage:* Select FP including  $z$  in all models.  
Add small constant  $\gamma$ : all values are positive
- *Second Stage:* Test both  $z$  and the selected FP for removal.

FPI functions are defined as

$$f_{FPI}(x) = \begin{cases} 0 & \text{if } x = 0 \\ x^p & \text{if } x > 0 \end{cases} \text{ with } p \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$$

and  $x^0$  denotes  $\log(x)$ .

**Table 2** Parameter combinations of the log normal distributed variable  $X$  in cases (index 1) and controls (index 0).

	$\mu_0$	$\mu_1$	$p_0$	$p_1$	$\sigma$	$\beta_0$	$\beta_1$	$OR_{X=x \text{ vs } X=0}$	$OR_{X=x \text{ vs } X=x_0}$
A	0	1	0.3	0.206	1	0	1.0	$e^{(\log(x))} = x$	$e^{(\log(x)-\log(x_0))}$
B	0	0.5	0.2	0.100	1	0.686	0.5	$e^{(0.686+0.5\log(x))} = 1.986\sqrt{x}$	$e^{(0.5(\log(x)-\log(x_0)))}$
C	0	1	0.2	0.100	1	0.311	1.0	$e^{(0.311+\log(x))} = 1.365x$	$e^{((\log(x)-\log(x_0)))}$

FP2 functions are defined as

$$f_{FP2}(x) = \begin{cases} (0, 0) & \text{if } x = 0 \\ (x^{p_1}, x^{p_2}) & \text{if } x > 0 \end{cases} \quad \text{with } p_1, p_2 \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$$

and  $x^0$  denotes  $\log(x)$  and for  $p_1 = p_2$  we define  $x^{p_2} = x^{p_1} \log(x)$ .

We suggest a modified FP-spike procedure where FP-transformations are applied to positive observations only and the spike at zero remains unchanged, as motivated by the theoretical derivations given in the previous section. The model selection algorithm also remains unchanged.

### 3 Simulation study

#### 3.1 Design of the simulation

In a simulation, we investigate the performance of the procedures for some realistic scenarios. We consider a case-control scenario and one covariable where the continuous part of  $X$  is lognormal distributed and with a spike at zero. We choose the following parameter combinations A, B and C (Table 2):

The parameters  $\mu_0, \mu_1, p_0$  and  $p_1$  are chosen such that the coefficient  $\beta_0$  is either equal to (A) or different from (B, C) zero. In the intermediate combination C only one parameter in comparison to A and B is changed. The percentages of zero values are 20.6% (cases) and 30% (controls) (A), and 10% (cases) and 20% (controls) (B, C). The resulting correct dose-response functions are given in the last two columns of Table 2. In the combinations A and C, there is a linear dose-response curve ( $\beta_1 = 1$ ). The coefficient for the binary indicator variable is equal to zero in A and different from zero in C. In the second combination  $\beta_0$ , the coefficient for the binary indicator variable is different from zero, and the dose-response curve is concave.

A total of 500 cases and 500 controls were simulated with a simulation sample size of 1000. We then apply the following procedures:

- (i) fit the correct model (which includes for A the log-transformed variable only and for B and C the binary term and the log-transformed variable),
- (ii) fit the standard FP procedure using  $\gamma = 1$  (no specific role of the spike at zero),
- (iii) fit the FP-spike according to Royston *et al.* (2010) using different values for  $\gamma$  (0.001, 0.1, 1),
- (iv) fit the modified FP procedure (FP-spike-mod) as suggested in Subsection 2.4

To present the results of the simulation, we calculated the estimated dose-response curves from all methods (i)–(iv) applied to the 1000 generated datasets and present graphically the mean curve and the empirical 95% confidence limits which are obtained as the 2.5% and 97.5% fractile of all estimated dose-response curves. In addition, we present in a table key parts of FP models selected in methods (ii)–(iv). We used SAS version 9.2 for the simulation. We have developed a program for the procedures (iii) and (iv) which is based on the program developed for method (ii) and which is available on the website of one of us (W.S.).

**Table 3** Percentaged distribution of selected models.

Parameter combination A. Correct model is an FP1 (the binary term cancels out), $\exp(\log(x))$ .							
Selected model	Standard FP procedure			FP-spike			FP-spike-mod
	$\gamma$			$\gamma$			
	0.001	0.1	1	0.001	0.1	1	–
Null	0	0	0	0	0	0	0
Linear	0	0	0.8	0	0	0	0
FP1	0	0	0.7	0	0.2	0	79.5
FP2	100	100	98.5	0	0	0	1.5
Linear+spike	–	–	–	0	0	0	0
FP1+spike	–	–	–	95.6	94.3	91.7	15.5
FP2+spike	–	–	–	4.4	5.5	8.3	3.5
Spike	–	–	–	0	0	0	0
Parameter combination B. Correct model is an FP1 with spike, $\exp(0.686 + 0.5 \log(x))$ .							
Null	0	0	0	0	0	0	0
Linear	5.7	6.6	6.7	4.0	3.8	4.5	4.0
FP1	41.6	58.4	88.9	5.2	18.4	45.7	9.7
FP2	52.7	35.0	4.4	0.6	0.8	1.2	0.9
Linear+spike	–	–	–	6.2	6.7	7.6	6.6
FP1+spike	–	–	–	79.8	67.2	39.0	75.0
FP2+spike	–	–	–	4.1	2.9	2	3.8
Spike	–	–	–	0.1	0.2	0	0
Parameter combination C. Correct model is an FP1 with spike, $\exp(0.311 + \log(x))$ .							
Null	0	0	0	0	0	0	0
Linear	0	0	1.3	0	0	0	0
FP1	0	0.2	14.5	0	0	0	0.2
FP2	100	100	84.2	0	0.1	0	0
Linear+spike	–	–	–	0	0	0	0
FP1+spike	–	–	–	95.2	93.4	90.4	95.5
FP2+spike	–	–	–	4.8	6.5	9.6	4.3
Spike	–	–	–	0	0	0	0

### 3.2 Results

In the upper part of Table 3, we show which type of models have been selected for the different procedures, for parameter combination A. The correct model is an FP1 without a spike. In nearly all cases, the standard FP procedure selects an FP2 model, whereas FP-spike selects nearly always FP1+spike. The new procedure FP-spike-mod selected the correct FP1 without spike most often (79.5%). The problem of the standard FP approach is obvious from the left column of Figure 2. For  $\gamma$  values close to zero the selected FP2 functions falsely indicate a hook, whereas the selected function closely agrees with the true function if  $\gamma$  is not too large. These results illustrate that the standard FP approach is unsuitable for a variable with a spike at zero, even if the effect of the spike is small. The right column summarises the results for FP-spike and FP-spike-mod. Despite of differences concerning

the selection of the spike, all selected functions agree very well with the true function, a result of the absence of the spike in parameter combination A.

For parameter combination B (Table 3), the correct model is again an FP1 with a spike, but this time with a value highly different from zero (0.686). Furthermore, the true function is non-linear. As a result of this more complex situation, we observe more variability in the functions selected (middle part of Table 3). For the standard FP procedure and FP-spike, the model selected depends strongly on  $\gamma$ . For a small  $\gamma$ , the latter selects an FP1+spike model in about 80%, similar to the modified spike procedure. This is observed in the modified procedure with 9.7% selecting FP1 without spike, 75% selecting an FP1 with spike and few other models. Applying the spike procedure ( $\gamma = 0.001, 0.1, 1$ ) leads in 39–80% to an FP1 spike model and by applying the standard procedure ( $\gamma = 0.001, 0.1, 1$ ) in 41–89% to an FP1 model.

Some more simulation results are given in Lorenz (2010).

For parameter combination C, we have an FP1 with a spike, which is intermediate between A and B ( $\beta_0 = 0.311$ ), and a linear dose–response. Again, the new FP-spike-mod procedure shows the best properties in selecting the correct model (lowest part of Table 3).

Figures 1–3 show the mean (short dashed line) of the estimated dose–response curves with corresponding empirical pointwise 95% confidence band (shaded area) for all analysed parameter combinations and for the different FP-procedures. The correct curve is marked as the bold line.

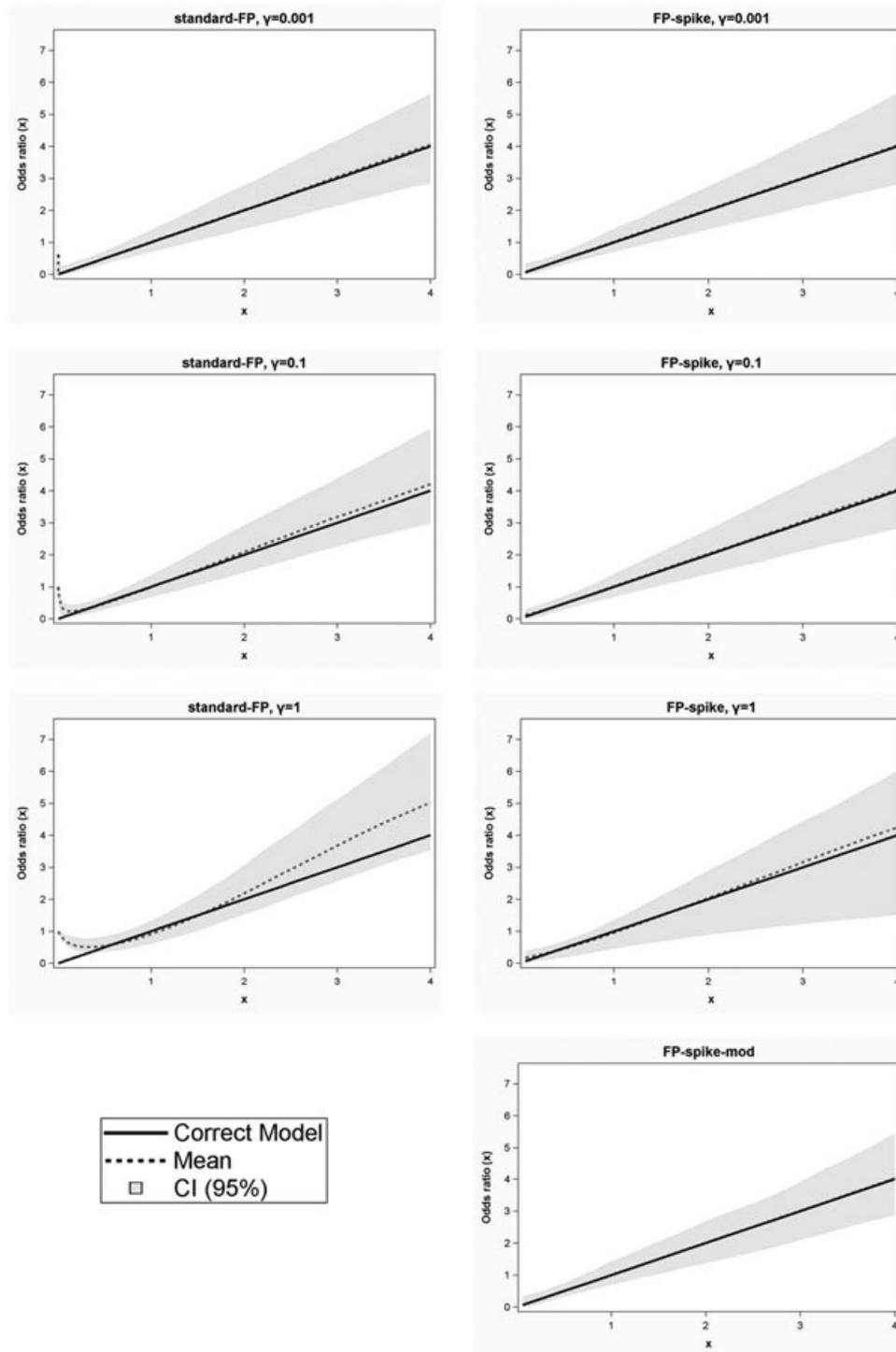
For the standard FP-procedure ignoring the spike at zero for both parameter combinations, the discrepancy between the correct curve and the mean curve increases with increasing value of  $\gamma$ . In the bottom plot of the left column in Figures 1–3, the means of the estimated odds ratios lie above those derived from the theoretical investigations. It is due to the selection of 0 as a baseline value, which constrains the odds ratio curve to pass through the origin of the graph. The shape of the curve is still correct but the origin is shifted downwards compared with the correct odds ratio curve.

For the old FP-spike procedure with  $\gamma = 0.001, 0.1, 1$  in the right column of Figures 1–3, the width of the confidence bands and the discrepancy between the correct function curve and the mean curve increase with increasing value of  $\gamma$ . For example, the 95% confidence limits of the dose–response function (A) at 3 are [2.10; 3.88] for the modified procedure, [2.16; 4.17] for the standard procedure with  $\gamma = 0.001$ , [2.27; 4.37] for  $\gamma = 0.1$  and [2.57; 5.10] for  $\gamma = 1$ .

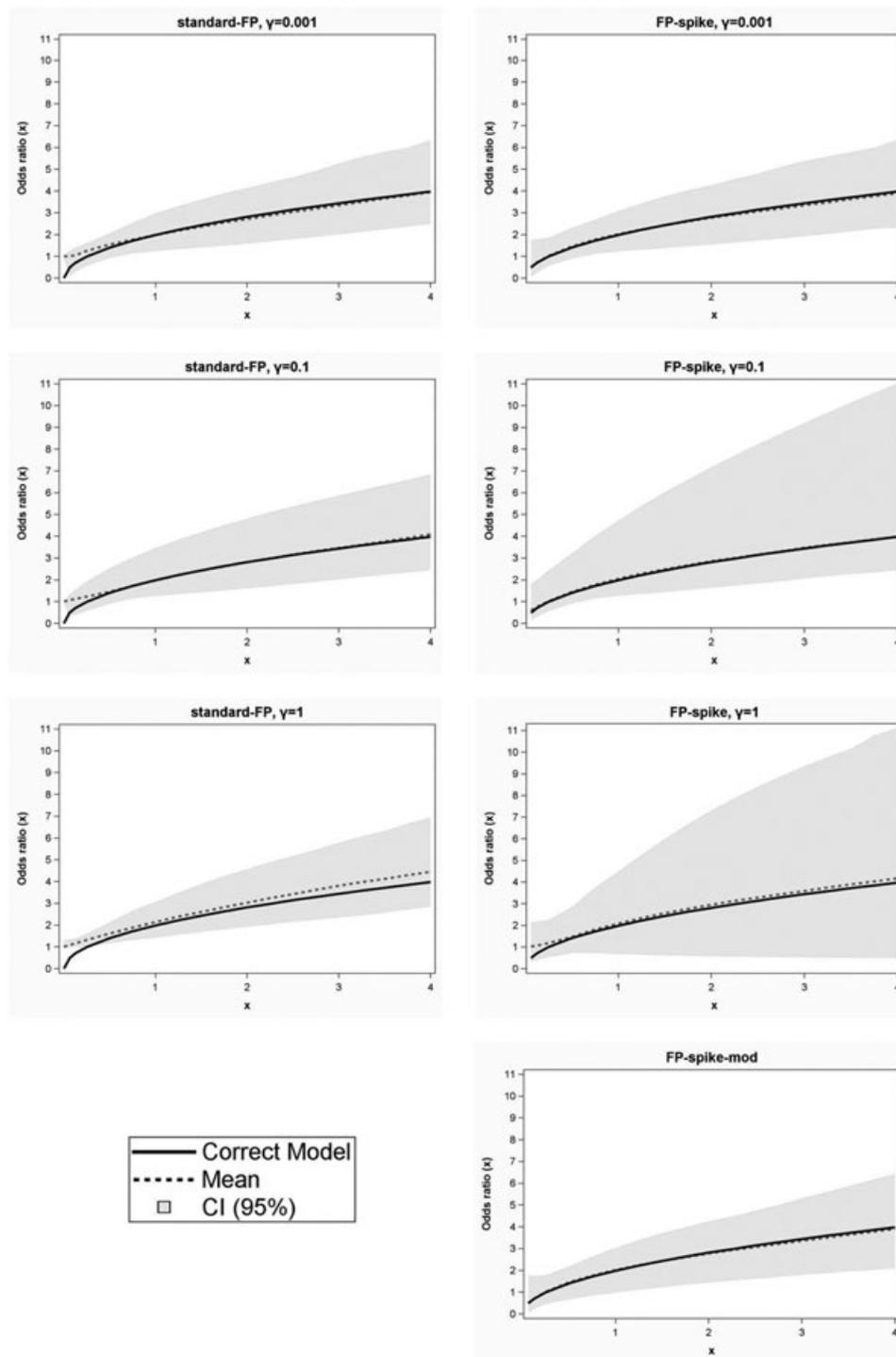
For the modified procedure without a constant in the bottom plot of Figures 1–3, the correct function curve is closest to the mean of the estimated dose–response curve. It also has the narrowest confidence band and therefore is the best of all procedures.

## 4 Analysis of data example

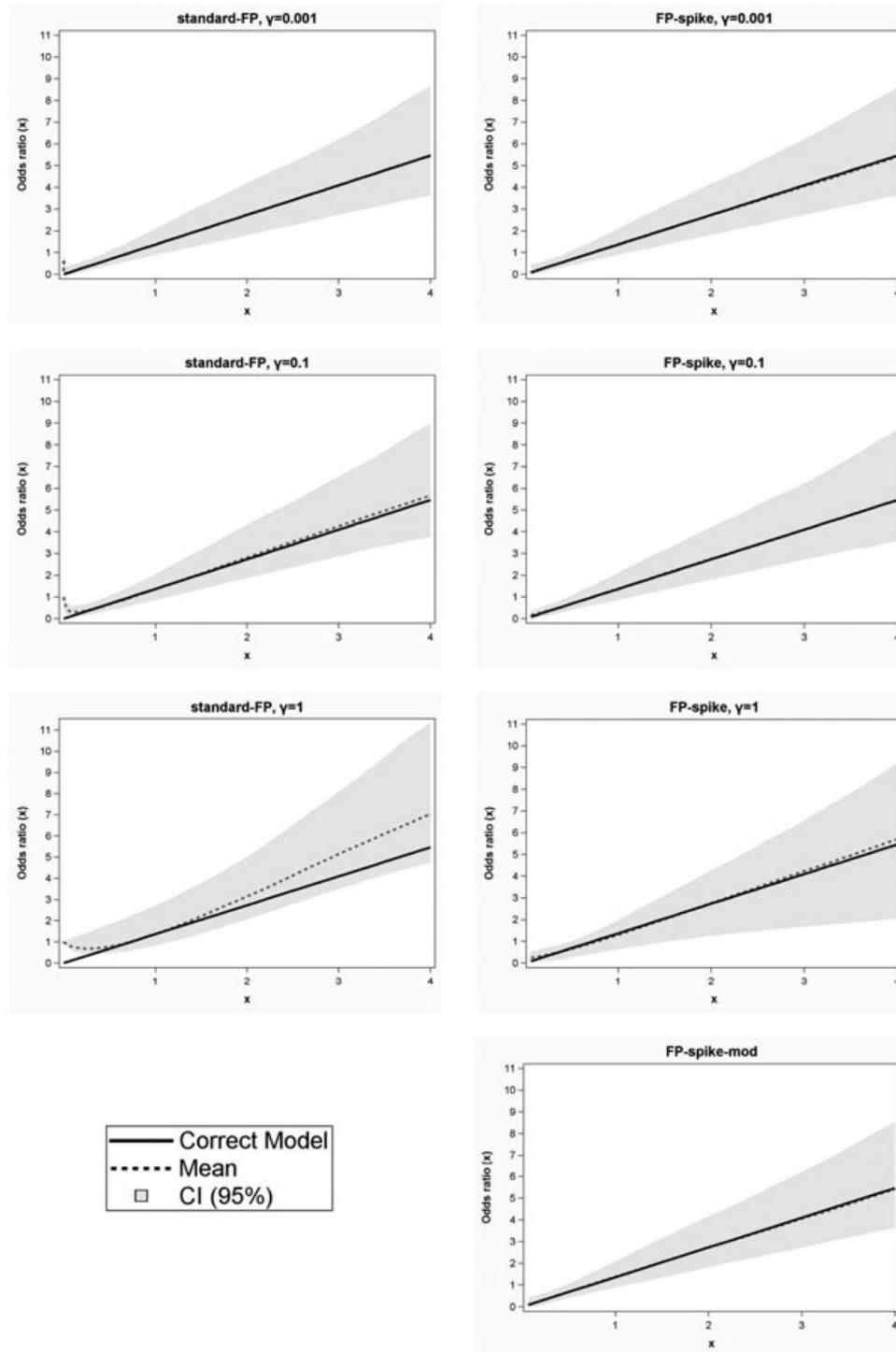
In practice, the true distribution of covariables is unknown. In this chapter, we re-analyse the data from a case–control study presented in the introduction. In the analysis of Royston *et al.* (2010), this dataset was also used as an example, and the constant  $\gamma$  was chosen as  $\gamma = 1$ . Here, we analysed the data with three different  $\gamma$  values:  $\gamma = 0.001, 0.1, 1$ . Figure 4 shows the fitted odds ratio functions. According to all models, a small intake of alcohol (up to 20 g/day) is associated with a slightly reduced risk of breast cancer and a high dose yields an increased risk. With a standard FP analysis ( $\gamma = 1$ ), an FP2 model with the power values [0; 0.5] is selected. According to the deviance criterion, its fit is only slightly worse (2636.2 vs. 2635.4) than the FP2+spike model. The fitted curves are very similar, the odds ratio from the spike FP2 model being somewhat larger at high alcohol consumption levels. The odds ratios, derived by the FP-spike procedure at  $x = 1$  g/day (OR = 0.9) are slightly higher than the ones derived by the standard procedure (OR = 0.8). In comparison to the odds ratio curve, computed by the standard FP procedure, we also display the odds ratio for categorical analysis (step function) and with a kernel density approach in Figure 4. Here, we calculated OR ( $X = x$  vs.



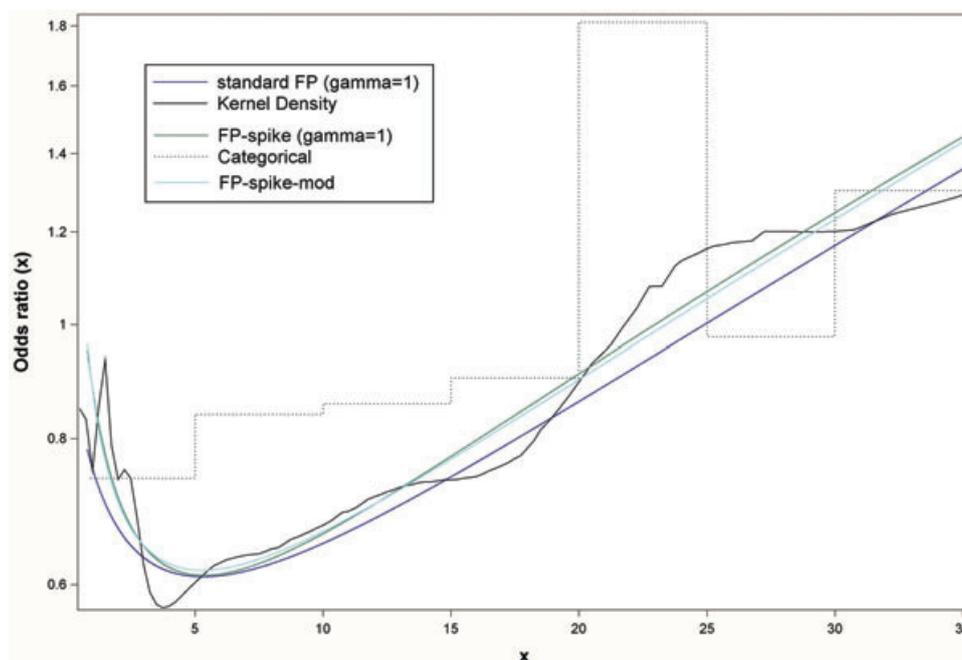
**Figure 1** Estimated dose–response curves of parameter combination A. Left column: FP procedure with  $\gamma = 0.001, 0.1, 1$ . Right column: FP-spike procedure with  $\gamma = 0.001, 0.1, 1$ . Bottom: MFP-spike-mod (no  $\gamma$ ).



**Figure 2** Estimated dose–response curves of parameter combination B. Left column: FP procedure with  $\gamma = 0.001, 0.1, 1$ . Right column: FP-spike procedure with  $\gamma = 0.001, 0.1, 1$ . Bottom: MFP-spike-mod (no  $\gamma$ ).



**Figure 3** Estimated dose–response curves of parameter combination C. Left column: FP procedure with  $\gamma = 0.001, 0.1, 1$ . Right column: FP-spike procedure with  $\gamma = 0.001, 0.1, 1$ . Bottom: MFP-spike-mod (no  $\gamma$ ).



**Figure 4** Breast cancer data. Fitted odds ratios from the risk of breast cancer plotted against alcohol consumption.

**Table 4** Breast cancer data. Analysis of alcohol consumption ( $x$ ) with a spike at zero.  $z$  is a dummy variable indicating non-drinkers. See text for details.

Method	Deviance	Dev. diff.	d.f.	Model selected	Power(s)	Estimated coefficients		
						$\beta_1$	$\beta_2$	$\beta_{\text{spike}}$
Null model	2670.9	—	5	—	—	—	—	—
Categorical analysis		31.6	5	—	—	—	—	—
FP procedure ( $\gamma = 0.001$ )		33.0	4	FP2	0.5, 1	-0.32	0.06	—
FP procedure ( $\gamma = 0.1$ )		33.8	4	FP2	0.5, 0.5	-0.66	0.18	—
FP procedure ( $\gamma = 1$ )		34.7	4	FP2	0, 0.5	-0.77	0.62	—
FP-spike ( $\gamma = 0.001$ )		35.4	5	FP2+spike	0, 0.5	-0.67	0.58	3.93
FP-spike ( $\gamma = 0.1$ )		35.4	5	FP2+spike	0, 0.5	-0.70	0.60	1.11
FP-spike ( $\gamma = 1$ )		35.4	5	FP2+spike	-0.5, 0.5	2.71	0.43	0.49
FP-spike-mod		35.4	5	FP2+spike	0, 0.5	-0.67	0.58	-0.71

$X = 0$ ) as  $\frac{p_0 \tilde{f}_{X|Y=1}(x)}{p_1 \tilde{f}_{X|Y=0}(x)}$ , when  $p_i$  is the proportion of non-drinkers in cases ( $i = 1$ ) and controls ( $i = 0$ ), and  $\tilde{f}_{X|Y=i}(x)$  is a kernel density estimate at  $X = x$  for cases ( $i = 1$ ) and controls ( $i = 0$ ) using the SAS procedure *proc kde* with variable *alc10*, as the logarithm to the base 10 of the daily alcohol intake in gram and bandwidth parameter  $BWM = 0.785$ .

In the following Table 4, we give the results of the analysis for different methods. A comparison between the null model and the selected FP2+spike model shows a highly significant association between alcohol consumption and case-control status. Simplifying the first stage (FP2+spike) model

in the second stage is not possible here, since dropping either the spike or the FP2 terms results in a significant worsening of the fit [Royston et al. (2010); Table 4].

For this particular dataset, deviances and functional forms are very similar between the three approaches and the seven resulting models.

## 5 Discussion

Although spike at zero occurs in practice very often, few papers have particularly addressed this problem. Robertson et al. (1994) proposed to add a binary indicator for the spike, and later Royston and Sauerbrei (2008) and Royston et al. (2010) proposed to model the positive part of the covariate by using FPs. They proposed the two-stage procedure FP-spike. Our theoretical results under specific distributional assumptions show that a binary indicator is needed for suitable dose–response curves when a covariate has a spike at zero. This paper shows that the usual FP approach is insufficient for such a situation, and a modification of FP-spike has been developed (FP-spike-mod). We developed a theoretical justification for this modification and showed empirically the advantages of the method.

The aim of this simulation study was to (i) assess the performance of the modified spike at zero procedure without the need for a constant  $\gamma$  and (ii) of the procedure as a function of the constant  $\gamma$  (0.001, 0.1, 1). According to the previous theoretical investigations, the study was performed on the assumption of biased estimates caused by the constant term  $\gamma$ . Figures 2 and 3 showed that we derive the best fit of the correct OR curve with the modified procedure without adding a constant  $\gamma$ . The discrepancies between the correct OR curve and the simulation results as well as the width of the confidence bands increase with increasing value of  $\gamma$ .

The inclusion of a binary variable exposed yes/no may lead to a dose–response function for which  $\lim_{x \rightarrow 0} OR_{X=x} \text{ vs } X=0 \neq 1$ . This may not be satisfactory for several reasons: If the goal is to estimate the risk for a low dose, the result may not be meaningful. Formally, this is correct under specific distributional assumptions. If, on the other hand, an a priori assumption can be made that the dose–response curve is monotone, starting at 1 for dose zero, then the suggested procedure is not recommended for low dose risk estimation, and the standard FP procedure may be more appropriate. A possible outcome of the procedure with spike is that the final selected model contains the binary term only. In that case, one can conclude that the exposure is significantly related to the outcome and that specific exposure values have no further effect on the outcome. Provided that such a result is based on a study with sufficient power to detect a functional influence, it would be easy to interpret and often important in practice. Summing up the results from simulation, the modified procedure gives the expected results and thereby manifests our theoretical assumptions. The simulation has shown that it is preferable to leave the values  $X = 0$  untransformed and transform only the positive values without adding a constant in advance. Varying the constant value has little influence on the shape of the OR curves, but it does have an influence on the selected power terms. The procedure proposed by Royston and Sauerbrei (2008) has good statistical properties and brings about an improvement in modelling dose–response curves in the case of a ‘spike at zero’.

The extended procedure which has been developed in this paper and which is consistent with the theoretical solution has the best statistical properties of the procedures analysed.

To illustrate a potential application of FP-spike, we used the procedure to reanalyse a breast cancer study as described above. The binary indicator emphasizes the role of the non-drinkers. In the data example of the breast cancer study, the deviances from the second stage of the procedure [Table 2: published in Royston et al. (2010)] show that neither the FP component nor the binary indicator variable can be rejected from the model. Both procedures give approximately equivalent results, whereas the spike procedure and the modified procedure are somewhat closer to the odds ratio curve derived by the kernel density estimation. Deviances and functional forms are very similar between the two approaches and the seven resulting models. However, data from other studies may show larger differences between FP-spike and FP-spike-mod.

From Table 4 and Figure 4, it is apparent that the FP-spike procedure and the FP-spike-mod procedure had very similar results with respect to the dose–response curve and to the deviances. The model with the lowest deviance is not the FP-spike-model. We recommend, however, to use this procedure generally because (i) we have shown that this model is theoretically justified and (ii) a choice of  $\gamma$ , which is arbitrary and has been subject to some discussions, is avoided. The general method of FPs can be applied to regression models, as described in Royston and Sauerbrei (2008, p. 10.14). In survival time data, time-varying effects add a further component in the FP method (Abrahamowicz and MacKenzie, 2007). The extension described here with the spike at zero situation is generally applicable as well. Dose–response modelling is of interest within all types of regression models as well, such as, for example, Cox regression. The general principles as given here can similarly be applied.

In summary, we believe the FP-spike-mod procedure is a useful extension of the published method to analyse spike at zero variables. The method is supported by theoretical considerations. Further research to deal with confounding and interaction in this context both in terms of designing multivariable simulation studies and application to very large datasets from epidemiological studies is ongoing.

**Acknowledgements** E. Lorenz was supported by the German Research Foundation (DFG), GRK 793 and BE 2056/10-1. P. Royston was supported by U.K. Medical Research Council grant number MC\_US\_A737\_0002. P. Royston and W. Sauerbrei received some support from the RiP-program, Mathematisches Forschungsinstitut, Oberwolfach, Germany. We thank J. Chang-Claude for providing the data from the breast cancer study.

### Conflict of interest

*The authors have declared no conflict of interest.*

## References

- Abrahamowicz, M. and MacKenzie, T. A. (2007). Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine* **26**, 392–408.
- Becher, H. (1992). The concept of residual confounding in regression models and some applications. *Statistics in Medicine* **11**, 1747–1758.
- de Boer, C. (2001). *A Practical Guide to Splines*. (revised edn.). Springer, New York.
- Hallstrom, A. P. (2010). A modified Wilcoxon test for non-negative distributions with a clump of zeros. *Statistics in Medicine* **29**, 391–400.
- Kropp, S., Becher, H., Nieters, A. and Chang-Claude, J. (2001). Low and moderate alcohol consumption and breast cancer risk by age 50 among women in Germany. *American Journal of Epidemiology* **154**, 624–634.
- Lachenbruch, P. A. (2001). Power and sample size requirements for two-part models. *Statistics in Medicine* **20**, 1235–1238.
- Lorenz, E. (2010). Eine Simulationsstudie zur Untersuchung einer erweiterten Fractional Polynomial (FP) Prozedur für die Situation eines 'spike at zero'. *Diplomarbeit, Ruprecht-Karls-Universität, Heidelberg*.
- Olsen, M. K. and Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* **96**, 730–745.
- Robertson, C., Boyle, P., Hsieh, C. C., Macfarlane, G. J. and Maisonneuve, P. (1994). Some statistical considerations in the analysis of case-control studies when the exposure variables are continuous measurements. *Epidemiology* **5**, 164–170.
- Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics* **43**, 429–467.
- Royston, P. and Sauerbrei, W. (2008). *Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. Wiley, Chichester.
- Royston, P., Sauerbrei, W. and Becher, H. (2010). Modelling continuous exposures with a 'spike' at zero: a new procedure based on fractional polynomials. *Statistics in Medicine* **29**, 1219–1227.
- Schisterman, E. F., Reiser, B. and Faraggi, D. (2006). ROC analysis for markers with mass at zero. *Statistics in Medicine* **25**, 623–638.