

Modeling variables with a spike at zero. Examples and practical recommendations

Journal:	<i>American Journal of Epidemiology</i>
Manuscript ID	AJE-00781-2015.R3
Manuscript Type:	Practice of Epidemiology
Key Words:	case-control study, dose-response model, fraction unexposed, fractional polynomials, regression modeling

SCHOLARONE™
Manuscripts

Review

Modeling variables with a spike at zero. Examples and practical recommendations

Eva Lorenz, Carolin Jenkner, Willi Sauerbrei and Heiko Becher

Author affiliations: Institute of Public Health, Medical Faculty, University of Heidelberg, Im Neuenheimer Feld 324, 69120 Heidelberg, Germany (Eva Lorenz, Heiko Becher); IMBI, Freiburg University Medical Center, Stefan-Meier-Str. 26, 79100 Freiburg, Germany (Carolin Jenkner, Willi Sauerbrei); and Institute of Medical Biometry and Epidemiology, University Hospital Hamburg-Eppendorf, Germany (Heiko Becher).

Running head: Modeling variables with a spike at zero.

Correspondence to Prof. Heiko Becher, Institute of Medical Biometry and Epidemiology, University Hospital Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany (phone: 040-7410-59550, fax: 040-7410-57790, e-mail: h.becher@uke.de)

Abbreviations: CI, confidence interval; FP, fractional polynomial; FP1, first degree fractional polynomial; FP2, second degree fractional polynomial; FSP, function selection procedure; HT, hormone therapy; OR, odds ratio; SAZ, spike at zero

Correspondence to Prof. Heiko Becher, Institute of Medical Biometry and Epidemiology, University Hospital Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany (phone: 040-7410-59550, fax: 040-7410-57790, e-mail: h.becher@uke.de)

1
2
3 Abstract
4

5 In most epidemiological studies and in clinical research generally, there are variables with a spike
6 at zero, namely variables where a proportion of individuals have zero exposure (e.g. never
7 smokers) and among those exposed the variable has a continuous distribution. For modeling such
8 variables different options exist, such as categorization where the non-exposed form the reference
9 group, or ignoring the spike by including the variable with or without some transformation or
10 modeling procedures in the regression model.
11
12
13
14
15
16
17
18

19
20 It has been shown that such situations can be analyzed by adding a binary indicator
21 (exposed/non-exposed) to the regression model, and a method based on fractional polynomials to
22 estimate a suitable functional form for the positive part has been developed.
23
24
25
26
27

28 In this paper we compare different approaches using data from three case-control studies
29 conducted in Germany. These are the MARIE study on breast cancer, conducted from 2002 to
30 2005; the RHEIN-NECKAR-LARYNX study on laryngeal cancer, conducted from 1998 and
31 2000 and a study on lung cancer, conducted from 1988 to 1993. Strengths and limitations of
32 different procedures are demonstrated, and some recommendations for practical use are given.
33
34
35
36
37
38
39

40 Keywords: case-control study; dose-response model; fraction unexposed; fractional polynomials;
41 regression modeling
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 A goal in the analysis of epidemiological data is often the estimation of a dose-response
4 relationship for risk factors which are composed of positive continuous values and zeros. Typical
5 examples are occupational exposures, e.g. asbestos exposure or tobacco consumption where a
6 proportion of individuals may be completely unexposed (spike at zero, SAZ), and the exposure of
7 those who have been exposed follows a continuous distribution.
8
9

10
11
12
13
14
15
16
17 There are both statistical problems, and problems with regard to interpretation arising from this
18 situation. The simplest method to deal with the situation is to categorize the variable, and use the
19 non-exposed group as reference. This is intuitively appealing, easy to interpret and popular.
20
21
22 However, there are major disadvantages which have been described in many papers, such as
23
24
25
26
27 Altman et al. (1), Vickers and Lilja (2), and Barendregt et al. (3).
28
29

30
31
32 To account for the SAZ variable, Jedrychowski et al. (4) included a binary variable smoker/non-
33 smoker into the model in addition to the dose variable. This was an ad hoc approach without
34 giving a formal rationale. The method was more formally described in Robertson et al. (5). The
35 first parameter then represents the basic association of exposure and the second parameter
36 represents the association of the levels of exposure among the exposed.
37
38
39
40
41
42
43
44
45

46 For general modeling of a continuous variable, the fractional polynomial approach (FP) has
47 become popular. Recently, the SAZ situation has been considered using an extended FP approach
48 in Royston et al. (6) and refined in Becher et al. (7). We have derived the theoretically correct
49 model under some specific assumptions on univariate continuous distributions (8). We have
50 expanded this by investigating the correct dose-response curve for a SAZ situation and univariate
51 normal, log normal and gamma distribution of the positive part of X (7). Extensions to the
52
53
54
55
56
57
58
59
60

1
2
3 bivariate case were published recently (9). However, in real data the exposure distribution often
4
5 takes no recognizable or standard form.
6
7

8
9
10 The main aim of this paper is to present and compare methods to model a continuous variable
11
12 with a SAZ and to give recommendations for practical applications. In section 2, we describe the
13
14 methods. In section 3, we introduce studies which included a variable with a SAZ. We re-analyze
15
16 the data, modeling the SAZ using the above alternative approaches and compare to the original
17
18 analysis. The discussion includes recommendations for practical application.
19
20
21

22 23 24 25 METHODS

26
27
28
29
30 In this section, we describe five methods to model the functional form of continuous variables
31
32 with a SAZ with regression models. These are summarized in Table 1.
33
34
35

36 37 (1) Categorization of the SAZ variable

38
39 The continuous variable X with a SAZ is transformed into k categories. The non-exposed group
40
41 defines the baseline. From the regression model we get $k-1$ regression coefficients which allow
42
43 direct estimation of the association of categories 2 to k relative to category 1. This method is still
44
45 commonly used in epidemiology. It corresponds to classical methods for the analysis of grouped
46
47 data. The criteria for choosing the number of categories and their limits are the same as for the
48
49 classical methods described in Becher et al. (10). Three to five groups seem to be used most
50
51 often.
52
53
54

55 56 57 (2) Modeling the SAZ variable assuming a linear association (untransformed)

1
2
3 The standard method which uses the full information is to include X untransformed into the
4 model without considering a binary indicator. It assumes linearity in the linear predictor. This
5 will be called “Linear” in the following.
6
7
8
9

10 (3) Fractional polynomial procedure

11 The FP approach has originally been suggested by Royston and Altman (11). The idea of FP is to
12 allow the variable to enter the model after it has been transformed, in order to allow non-linear
13 relationships. The transformation used is selected from a predefined set of eight different values
14 giving first degree FP functions (FP1). This set is defined as $H_1(x) = \beta_1 x^p$ with $p \in S = \{-2, -1, -$
15 $0.5, 0, 0.5, 1, 2, 3\}$ with x^0 being defined as $\log(x)$. More flexible second degree FP functions
16 (FP2) are defined by $H_2(x) = \beta_1 x^p + \beta_2 x^q$ with p and q taken from S . If $p=q$ the second
17 transformation is defined as $x^p \log(x)$ such that $H_2(x) = \beta_1 x^p + \beta_2 x^p \log(x)$. Royston and Altman
18 (11) showed that second degree FPs ($m=2$) cover a rich family of dose-response relationships
19 which is sufficient in most applications (11). To account for the fact that some of these
20 transformations cannot be performed for $X=0$, a small constant c is added to each observation.
21 We follow a common procedure to use $c=1$ (12), although it has been criticized since the result of
22 the modeling procedure depends on the choice of c and alternatives have been suggested (section
23 4.7 and 5.6 in (13)).
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44

45 For model selection, a closed test procedure, called FSP (function selection procedure), has been
46 suggested (13).
47
48
49

50 The user must choose the significance level (α) and the degree (m) of the most complex FP
51 model allowed. Typical choices are $\alpha = 0.05$ and FP2 ($m=2$). Furthermore, a default function for
52 X is required. The identity ($p=1$) is chosen. The FSP selects a function in several steps. First, the
53 best FP2 model is compared to the null model on 4 d.f. (one d.f. for parameters β_1 and β_2 and one
54
55
56
57
58
59
60

1
2
3 d.f. each for the choice of p and q). If the likelihood ratio (LR) test is not significant and selection
4 of variables is of interest (for example in the multivariable fractional polynomials (MFP)
5 procedure (13)) the algorithm stops. In the present context when estimation of the dose-response
6 is the main purpose, the (default) linear function is chosen and corresponding parameter estimates
7 are provided. A non-significant association in the first stage of the FSP does not imply that a
8 usual test for linearity is also non-significant (see 4.16 in Royston and Sauerbrei (13)). If the first
9 test is significant, the best FP2 model is compared to the default function (3 d.f.). If the test is not
10 significant, the algorithm stops choosing the default as final model ($R(X=x \text{ vs } X=0) = \exp(\beta x)$).
11 Otherwise, the best FP2 model is tested against the best FP1 model. If the test is not significant,
12 the algorithm stops choosing the best FP1 model. Otherwise the best FP2 model is chosen as final
13 function.
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

32 We denote the result of the FP method using the closed test procedure as “FP”. “best FP1” refers
33 to the first degree FP with the smallest deviance among all eight FP1 functions, and “best FP2”
34 denotes the second degree FP with the smallest deviance among all 36 FP2 functions. The
35 principle of the FSP does not change if adjustment for further variables is required.
36
37
38
39
40
41
42
43
44
45
46
47
48
49

- (4) Modeling the SAZ variable assuming a linear association (untransformed) and including a binary indicator

50 Robertson et al. (5) described a method to include a binary indicator Z, which takes value 1 if
51 $X=0$ and 0 otherwise, into the model in addition to the untransformed dose variable. This
52 approach accounts for modeling unexposed individuals separately while the continuous part of
53 the variable is modelled assuming linearity. We refer to this method as “Linear+z”.
54
55
56
57
58
59
60

(5) Fractional polynomial procedure including a binary indicator for the SAZ

An FP procedure to model one continuous variable with a SAZ was described in Becher et al. (7). We refer to this version as “FP-spike”. In a SAZ situation, an additional coefficient β_0 is estimated which refers to the binary indicator Z which takes the value 1 if $X=0$ and 0 otherwise. The positive continuous variable ($X>0$) is modelled using FPs with transformations from the class of FP functions as defined in method (3). Adding a constant term c was obviated by applying the transformations to positive values only. The procedure is an extension to the standard FP procedure with two stages. In a first stage the FSP is applied in the same way, but all models include Z . The best FP2 model with Z (FP2+z) is compared to the null model on 5 d.f. (one d.f. for parameters β_1 and β_2 one d.f. each for the choice of p and q , and one for Z). If the LR test is not significant, the variable is considered to have no association at the specified α level. As before, the default (linear) function would be chosen as the result of the first stage of the procedure. Provided that the first test is significant, the best FP2+z is compared to the default function with Z (3 d.f.). If the test is not significant, the first stage ends. Otherwise, the best FP2+z is tested against the best FP1+z. If the test is not significant, the first part of the algorithm stops choosing the best FP1+z. Otherwise the best FP2+z will be the result of the first stage. Then, in a second stage, it is tested whether either Z or the selected FP can be removed from the model.

RESULTS

1
2
3 This section presents a comparison of these methods in three case-control studies. In addition to
4 the original analysis as presented in the publications, we re-analyzed the data, using the above-
5 described methods (2) to (5). Resulting dose-response functions are displayed graphically and
6 deviances and deviance differences which are needed for the FSP are presented.
7
8
9
10
11
12
13
14

15 Study on postmenopausal breast cancer

16
17
18
19
20

21 This is a large population-based case-control study on breast cancer conducted in Germany,
22 including 3,464 cases aged 50-74 years and 6,657 controls, frequency matched by region and age
23 (14). Here, duration of use of hormone therapy in years (HT) was considered as continuous risk
24 factor with a SAZ. Exposure was reported in 29.85% of cases and 21.95% of controls. The
25 distribution of the continuous part is given in Figure 1.
26
27
28
29
30
31
32
33

34 The original paper provides an analysis with duration of HT categorized into four groups and
35 adjusted for menopausal status, age at menarche, number of full-term pregnancies, ever breast-
36 feeding, number of mammograms, ever benign breast disease, body mass index, first-degree
37 family history of breast cancer and occupational status (14) (Figure 3 in original publication). The
38 same adjustment variables are used in all other approaches. Results are given in Table 2 and
39 Figure 2.
40
41
42
43
44
45
46
47

48 Compared to unexposed individuals, significantly elevated ORs were observed for duration of
49 use of HT of less than 5 years, 5-9 years, 10-14 years and 15 or more years. The risk started to
50 increase after 5 years of use but did not significantly increase further in the categories with a
51 longer duration of use (14).
52
53
54
55
56
57

58 The standard FP approach was applied adding the constant $c=1$ to the original observations of X .
59
60

1
2
3 The FSP yielded a FP1 as best model (Table 2, method 3) with power (p) as 0, i.e. transformation
4
5 $\log(x + 1)$. The fit is significantly better than the linear function (Table 2, method 2) and not
6
7 significantly worse than the best FP2 function (Table 2, method 3b). The FSP procedure is as
8
9 follows: The deviance difference of the best FP2 to the null model, 123.23 is much larger than
10
11 $\chi_{4,0.05}^2 = 9.49$ indicating an overall significant association of HT use. When comparing the best
12
13 FP2 model to the default (linear) function, the deviance difference, $123.23 - 105.09 = 18.14$ is larger
14
15 than $\chi_{3,0.05}^2 = 7.12$ indicating that the default function is not sufficient. Then, the best FP2 model
16
17 is compared to the best FP1 model. Here the deviance difference is $123.23 - 117.79 = 5.44 < \chi_{2,0.05}^2 =$
18
19 5.99 and thus the best FP1 model is the result of the FSP.

20
21 The FP-spike approach yielded a FP1 function with power $p=0$ as the best model (Table 2,
22
23 method 5). The procedure runs as follows: The deviance difference of the best FP2+z to the null
24
25 model, 125.52, is larger than $\chi_{5,0.05}^2 = 11.07$ indicating an association of HT use. In the next step,
26
27 the best FP2+z model is compared with the default (linear) function with Z. The deviance
28
29 difference, $125.52 - 109.10 = 16.42$ is larger than $\chi_{3,0.05}^2 = 7.12$ indicating that the default function
30
31 with Z does not appropriately describe the dose-response. In the next step, the best FP2+z model
32
33 is compared to the best FP1+z model. Here the deviance difference, $125.52 - 121.52 = 4.00$
34
35 $< \chi_{2,0.05}^2 = 5.99$ and thus the best FP2+z model is not superior to the best FP1+z model. The result
36
37 of the first stage is thus an FP1+z. In the second stage Z and the FP1 component are each tested
38
39 for removal. The deviance difference to the model with Z only is $121.52 - 92.03 = 29.49 > \chi_{2,0.05}^2 =$
40
41 5.99 indicating that the continuous term cannot be omitted. The deviance difference to this model
42
43 without Z is $121.52 - 121.514 = 0.006 < \chi_{1,0.05}^2 = 3.84$ indicating that Z only marginally improves
44
45 the fit and therefore can be omitted. The resulting function for $OR(X=x \text{ vs } X=0)$ in the second
46
47 stage is $\exp(0.24\log(x))$ is displayed in Figure 2. The other approaches which are more limited in
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 terms of possible shapes of the dose-response show a considerably worse fit (Table 2, methods 2,
4
5 3b, 4).
6
7
8
9

10
11
12 Dose-response curves derived by all five approaches are very different. The categorical analysis
13 may be useful as a first step in the analysis. However, results depend strongly on chosen
14 categories and may be severely misleading. The categorical analysis indicated a small risk for
15 short exposure, and a rather constant elevated risk for longer exposures. This shape cannot be
16 approximated well by any of the other approaches. The results of the FP and the FP-spike
17 procedure are similar in terms of deviance and functional form.
18
19
20
21
22
23
24
25
26
27
28
29

30 Laryngeal cancer case-control study 31 32 33 34

35 A case-control study with 257 cases (236 males, 21 females) and 769 population controls (702
36 males, 67 females) 1:3 frequency matched by age and sex was conducted in southwest Germany
37 (15). We consider the lifetime hours of occupational exposure to cement dust as a single risk
38 factor with SAZ. Only a small number of patients with positive values of cement dust exposure
39 was reported (35 (13.62 %) cases and 37 (4.81 %) controls). The distribution of the continuous
40 part is given in Figure 3. Median lifetime exposure hours were 3,410 in exposed cases and 3,080
41 in exposed controls.
42
43
44
45
46
47
48
49
50
51
52
53

54 The results are presented in Table 3 and Figure 4. A logarithmic scale is used for the x-axis to
55 better illustrate the association at the low exposure level. The original paper provides an analysis
56
57
58
59
60

1
2
3 with the categorized variable (Table 3, method 1), stratified for age and gender (15). Compared to
4 unexposed individuals, significantly elevated and rather similar ORs were observed for exposure
5 levels of 1-3,000 and 3,000 and more lifetime working hours, indicating that the level of
6 exposure has no differential association. The standard FP approach yielded a FP1 with
7 transformation $(x+1)^{-2}$ as the best model (Table 3, method 3a). The fit is considerably better than
8 the default (linear) function (Table 3, method 2) with a deviance difference of 20.44-
9 10.11=10.33. The best FP2 was obtained for transformations $(x + 1)^{-2}$ and $(x + 1)^3$ (Table 3,
10 method 3c). This, however, improved the fit only slightly with deviance difference 22.04-
11 20.07=1.97 and the FSP yielded the FP1 function as the best model.
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26

27 The FP-spike approach yielded a final model with only the binary indicator Z (Table 3, method
28 5a). In the first stage, a linear function with Z was chosen. The second stage, investigating
29 whether one or the other component can be removed from the final model, showed that Z already
30 sufficiently described the functional relationship. The corresponding deviance difference is
31 20.44-20.07=0.37, indicating that the additional linear term does not contribute to model
32 improvement. Removing Z, however, significantly worsens the fit with deviance difference
33 20.44-10.11=10.33.
34
35
36
37
38
39
40
41
42

43 Methods 1, 3a and 5a as well as 3c and 5b give almost the same result in terms of deviance and
44 shape of the function. This is because the exposure is associated with an increased risk, however
45 with very little dose-dependence. Model 3a-c force the OR function continuously through the
46 value 1 for $x \rightarrow 0$, and similarities between 3a and 5a are obtained since $(x+1)^{-2} \rightarrow 0$. Among
47 exposed, the lowest recorded exposure value is 100 hours and for this exposure the estimated OR
48 is $\exp(-1.14((100 + 1)^{-2} - 1)) = \exp(-1.14 \times (-0.999)) \approx \exp(1.14) = 3.13$. The
49
50
51
52
53
54
55
56
57
58
59
60

dataset and Stata program used for this example are provided in Web Dataset 1 and Web Program

1.

Lung cancer case-control study

A hospital-based case-control study with 1,004 lung cancer cases and 1,004 population controls matched for region, sex and age was conducted in two areas in Germany (16). Here, we consider lifetime hours of asbestos exposure as continuous risk factor with a SAZ. A relatively high number of cases (65.44%) and controls (71.81%) reported zero asbestos exposure.

The continuous part is approximately log normally distributed in the 347 cases and 283 controls (Figure 5).

The original paper provides an analysis with the categorized variable, adjusted for smoking in four categories (16). The exposure variable was categorized in the original values according to the tertiles of the distribution among exposed individuals. Results are presented in Table 4 and Figure 6.

Compared to unexposed individuals, elevated ORs were observed for levels of cumulative lifetime working hours with asbestos exposure 1-940, 940-5,280 and more than 5,280 (Table 4, method 1). The standard FP approach using the closed test procedure yielded a linear function as the best model (Table 4, method 3a) with $\exp(0.18 \times 10^{-4}x)$ as the OR function ($se(\beta) = 7.06 \times 10^{-6}$). The more complex FP1 and FP2 result in different functions, but are not significantly better according to the FSP with $\alpha=0.05$ (Table 4, methods 3b and 3c).

1
2
3 The FP-spike approach using the closed test procedure yielded a linear function x without Z as
4 the best model (Table 4, method 5a). The resulting function is thus the same as derived by
5 standard FP. In the first stage of the selection procedure with Z included by default it was shown
6 that there is no significant association of the exposure variable, and the default (linear) function
7 for the exposure variable was chosen. The corresponding deviance difference of the best FP2+z
8 and the null model was 10.67 ($p = 0.06$). The second stage, when investigating whether both Z
9 and the selected (linear) function of the continuous part are needed for a suitable model, showed
10 that Z can be removed and the linear function sufficiently describes the functional relationship.
11 The best FP1 and the best FP2 function (Table 4, methods 5b and 5c) had a better model fit than
12 the linear function; however, the deviance difference was not significant. In this example,
13 therefore, both the FP and the FP-spike procedure yielded the same result.
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

33 DISCUSSION

34
35
36 We described five procedures to model exposure variables with a SAZ and applied these to data
37 from three case-control studies. While these datasets do not cover all possible practical data
38 situations, they provide substantial insight into some strengths and weaknesses of the methods.
39
40
41
42

43 A natural goal in dose-response analysis is to give the best possible answer how the response
44 depends on the exposure variable for a given dose. A categorical analysis is not satisfactory since
45 it implies (i) an arbitrary definition of cut-points, (ii) a jump of the risk estimate at these cut-
46 points which is biologically implausible, and (iii) an arbitrary choice of the number of categories.
47
48
49
50
51
52

53 In addition, categorization may introduce residual confounding and results in loss of power (17).
54

55 Nevertheless, we consider categorization with three to five categories as a useful first step in the
56 analysis; however, it should not be considered as the final result.
57
58
59
60

1
2
3
4
5 All other approaches yield continuous dose-response functions. These are parsimonious functions
6
7
8 estimated with a parametric approach. The method including a binary indicator has the principal
9
10 property that it may result in a risk function which does not start at one for a dose close to zero.
11
12 We have shown in a theoretical paper that this is a direct consequence under certain distributional
13
14 assumptions (9), however for risk extrapolation to low doses this property is certainly unwanted
15
16 and such a situation requires further research. Common procedures to adjust for confounders can
17
18 be applied. In the examples given, the same variables were included as in the original
19
20 publications.
21
22
23
24
25
26

27 To illustrate potential applications, we re-analyzed data from three case-control studies in which
28
29 the SAZ variable was categorized in the published analysis. We conclude that alternative
30
31 approaches would have been a better choice; however, it is more difficult to choose the best of
32
33 these.
34
35
36
37
38

39 In the breast cancer study a non-linear FP1 function without a binary indicator was selected with
40
41 the FP-spike procedure. The standard FP method also selected a FP1 function. In both cases the
42
43 log transformation was selected, and both resulted in a similar fit in terms of deviance. The dose-
44
45 response curves are very similar for a large range of observed values. Larger differences exist for
46
47 very short duration of use which is the result of the omitted constant added to the variable before
48
49 transformation in the FP-spike procedure. The linear approach, both with and without the binary
50
51 indicator, yield poorer fits, and cannot be recommended for this example.
52
53
54
55
56
57
58
59
60

1
2
3 The laryngeal cancer study was considerably smaller, and a very large proportion of 86.38 % in
4 cases and 95.19 % in controls had zero exposure for the SAZ variable. This limits the power to
5 detect complex dose-response functions and selection of the default (linear) function is often a
6 consequence. In the published analysis two exposure groups and the non-exposed baseline group
7 were considered, and the ORs were significantly elevated and similar in both exposure groups,
8 indicating that the association exists but independent of dose. Not surprisingly, the FP-spike
9 procedure selected the model with the binary indicator only as the final result. The standard FP
10 procedure selected a FP1. The selected function has a very high slope at low dose and
11 approximates a constant OR similar to the FP-spike procedure for higher doses. This is seen in
12 the smaller window in Figure 4. In terms of biological plausibility, the result of the standard FP
13 is to be preferred. We acknowledge that these are risk estimates for an unobserved dose range
14 only and must be interpreted very carefully.
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33

34 In the lung cancer study, a different situation was observed. A problem here is the more severe
35 skewness of the distribution of the exposure variable. When occupational exposure was analyzed
36 as a categorical variable in 3 distinct categories, no increased risk was observed for the low
37 exposure group, and increased risks, similar and both significant, for the middle and high
38 exposure categories. From these results one would expect an S-shaped dose-response curve
39 which is difficult to detect. Moreover, the risk estimates given from finer categories show an
40 irregular pattern, indicating that the random variation in the data is large. Consequently, both
41 standard FP and FP-spike yielded the linear default model. Apparently this function does not
42 match with the categorical estimates. The best FP1 and the best FP-spike yield a better fit in
43 terms of deviance with a steep slope at low dose, and these functions would possibly have been
44 selected if the sample size had been larger. This example also indicates a certain drawback of the
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 FP-spike procedure. The deviance difference in the initial step did not reach statistical
4
5 significance ($p=0.06$), caused by comparing deviances of the best FP2+z function with the null
6
7 model by using a test with 5 degrees of freedom. For the standard FP procedure it is well-known
8
9 that FSP loses some power if the underlying true function is linear (Royston and Sauerbrei (13),
10
11 chap 4.16). That may be considered as the price to pay if the true function is linear but the analyst
12
13 uses the closed test procedure to investigate whether non-linear functions fit the data better. For a
14
15 correct interpretation of the empirical standard error of the estimates also see (Royston and
16
17 Sauerbrei (13), chap 4.16)
18
19
20
21
22
23

24 The examples have shown that the analysis of a SAZ variable is complex and that general
25
26 recommendations are difficult to provide since it depends on the main goal of the analysis. If a
27
28 low dose extrapolation is needed where no observations are available, for example to set
29
30 acceptable lower limits of exposure from a study where exposed individuals had high exposure
31
32 levels, then neither the categorical analysis nor the FP-spike procedure can be used since the
33
34 function should continuously go through the origin which is one for dose zero. Then, the standard
35
36 FP method appears appropriate. If the goal is rather to provide a best estimate for, the common
37
38 exposure range, the FP-spike procedure seems appropriate.
39
40
41
42
43
44
45

46 ACKNOWLEDGEMENTS

47
48
49 Author affiliations: Institute of Public Health, Medical Faculty, University of Heidelberg, Im
50
51 Neuenheimer Feld 324, 69120 Heidelberg, Germany (Eva Lorenz, Heiko Becher); IMBI,
52
53 Freiburg University Medical Center, Stefan-Meier-Str. 26, 79100 Freiburg, Germany (Carolin
54
55
56
57
58
59
60

Jenkner, Willi Sauerbrei); and Institute of Medical Biometry and Epidemiology, University Hospital Hamburg-Eppendorf, Germany (Heiko Becher).

This work was supported by the German Research Foundation (grant BE 2056/10-2 and SA 580/7-2).

We thank Dr. Hermann Pohlabein, Prof. Wolfgang Ahrens and Prof. Karl-Heinz Jöckel for providing the data of the lung cancer study. We thank Prof. Jenny Chang-Claude and Dr. Anja Rudolph for providing the data of the postmenopausal breast cancer study and Ms. Ursula Eilber for her valuable technical assistance. We thank PD Dr. Heribert Ramroth for help with the data of the laryngeal cancer study.

Conflict of interest: none declared

REFERENCES

1. Altman DG, Lausen B, Sauerbrei W, et al. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 1994;86(11):829-835.
2. Vickers AJ, Lilja H. Cut-points in clinical chemistry: time for fundamental reassessment. *Clinical chemistry* 2009;55(1):15-17.
3. Barendregt JJ, Veerman JL. Categorical versus continuous risk factors and the calculation of potential impact fractions. *Journal of epidemiology and community health* 2010;64(3):209-212.
4. Jedrychowski W, Becher H, Wahrendorf J, et al. Effect of tobacco smoking on various histological types of lung cancer. *Journal of cancer research and clinical oncology* 1992;118(4):276-282.
5. Robertson C, Boyle P, Hsieh CC, et al. Some statistical considerations in the analysis of case-control studies when the exposure variables are continuous measurements. *Epidemiology (Cambridge, Mass.)* 1994;5(2):164-170.
6. Royston P, Sauerbrei W, Becher H. Modelling continuous exposures with a 'spike' at zero: a new procedure based on fractional polynomials. *Statistics in medicine* 2010;29(11):1219-1227.
7. Becher H, Lorenz E, Royston P, et al. Analysing covariates with spike at zero: a modified FP procedure and conceptual issues. *Biometrical journal*. 2012;54(5):686-700.
8. Becher H. The concept of residual confounding in regression models and some applications. *Statistics in medicine* 1992;11(13):1747-1758.

- 1
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
 - 8
 - 9
 - 10
 - 11
 - 12
 - 13
 - 14
 - 15
 - 16
 - 17
 - 18
 - 19
 - 20
 - 21
 - 22
 - 23
 - 24
 - 25
 - 26
 - 27
 - 28
 - 29
 - 30
 - 31
 - 32
 - 33
 - 34
 - 35
 - 36
 - 37
 - 38
 - 39
 - 40
 - 41
 - 42
 - 43
 - 44
 - 45
 - 46
 - 47
 - 48
 - 49
 - 50
 - 51
 - 52
 - 53
 - 54
 - 55
 - 56
 - 57
 - 58
 - 59
 - 60
9. Lorenz E, Jenkner C, Sauerbrei W, et al. Dose–response modelling for bivariate covariates with and without a spike at zero: theory and application to binary outcomes. *Statistica Neerlandica* 2015;69(4):374-398.
10. Becher H. *Analysis of continuous covariates and dose-effect analysis* In: W. Ahrens, I Pigeot (Eds) *Handbook of epidemiology*. 1057-1086. 2nd ed.: Berlin, Germany: Springer-Verlag; 2014.
11. Royston P, Altman DG. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 1994;43(3):429-467.
12. Ambler G, Royston P. Fractional polynomial model selection procedures: investigation of type I error rate. *Journal of Statistical Computation and Simulation* 2001;69(1):89-108.
13. Royston P SW. *Multivariable Model-building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables*. Wiley: Chichester; 2008.
14. Flesch-Janys D, Slinger T, Mutschelknauss E, et al. Risk of different histological types of postmenopausal breast cancer by type and regimen of menopausal hormone therapy. *International journal of cancer*. 2008;123(4):933-941.
15. Dietz A, Ramroth H, Urban T, et al. Exposure to cement dust, related occupational groups and laryngeal cancer risk: results of a population based case-control study. *International journal of cancer*. 2004;108(6):907-911.
16. Jöckel KH, Ahrens W, Jahn I, et al. Occupational risk factors for lung cancer: a case-control study in West Germany. *International journal of epidemiology* 1998;27(4):549-560.
17. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine* 2006;25(1):127-141.

Figure 1. Distribution of duration of hormone therapy (HT) in cases and controls in exposed individuals, MARIE Study, Germany, 2002–2005.

Figure 2. Dose-response curves of hormone therapy duration and postmenopausal breast cancer under different modeling procedures as given in Table 2. Circles denote odds ratio (OR) estimates for fine categories relative to exposure zero. The size of the circles indicate the number of individuals per category. The function resulting from the original analysis is displayed as solid black line, from method Linear as dotted black line, from method FP as short dashed black line, from method Linear+z as medium dashed black line and from method FP-spike as long dashed black line. MARIE Study, Germany, 2002–2005.

Figure 3. Distribution of the continuous part of lifetime exposure to cement dust in cases and controls separately, RHEIN-NECKAR-LARYNX Study, Germany, 1998–2000.

Figure 4. Dose-response curves lifetime hours of cement dust exposure and laryngeal cancer under different modeling procedures as given in Table 3. For values >15 fractional polynomial (FP) result coincides with FP-spike approach. Not visible in the plot. The function resulting from

1
2
3 the original analysis is displayed as solid black line, from method FP as short dashed black line
4 and from method FP-spike as long dashed black line. In Figure 4A, the dose-response curve is
5 plot using a linear scale. In Figure 4B, the low dose range is displayed using a logarithmic scale.
6 RHEIN-NECKAR-LARYNX Study, Germany, 1998–2000.
7
8
9

10
11 **Figure 5.** Distribution of the continuous part of the risk factor lifetime hours to asbestos exposure
12 in cases and controls separately, Lung Cancer Study, Germany, 1988-1993.
13
14
15

16 **Figure 6.** Dose-response curves of lifetime hours of asbestos exposure and lung cancer under
17 different modeling procedures as given in Table 4. Circles denote odds ratio (OR) estimates for
18 fine categories relative to exposure zero. The size of the circles indicate the number of
19 individuals per category. The function resulting from the original analysis is displayed as solid
20 black line, from method Linear as dotted black line, best FP1 as short dashed black line, best FP2
21 as medium dashed black line, from method Linear+z as long dashed line, best FP1+z as medium
22 dashed and dotted line and best FP2+z as long dashed and dotted black line. Lung Cancer Study,
23 Germany, 1988-1993.
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1. Summary of five methods to investigate a continuous covariable with a spike at zero (SAZ variable).

Model	Method	Description	Properties
1	Categorization	SAZ variable modeled in categories of groups of exposure values	yields simple final models, robust to outliers, easy to apply, natural baseline exists in case of SAZ variables yields a step-function for the risk, arbitrary choice of cut-points and number of categories, loss of power, prone to residual confounding
2	Linear	SAZ variable modeled continuously and untransformed	uses full information of a continuously measured variable linearity of the log odds ratio may be an invalid assumption, spike is ignored
3	FP	SAZ variable modeled using the FP method	uses full information of a continuously measured variable, wide range for shape of dose-response functions, low-dose extrapolation possible spike is ignored, arbitrary constant needs to be added before FP transformations
4	Linear+z	SAZ variable modeled continuously and untransformed including a binary indicator for the SAZ	enables modeling variables whose distribution has a discrete and a continuous component, uses full information of a continuously measured variable linearity of the log odds ratio may be an invalid assumption, limited option for low dose extrapolation
5	FP-spike	SAZ variable modeled using the FP method, including a binary indicator for the SAZ	enables modeling variables whose distribution has a discrete and a continuous component, uses full information of a continuously measured variable, wide range for shape of dose-response functions limited option for low dose extrapolation

Abbreviations: FP, fractional polynomial; SAZ, spike at zero

Table 2. Comparison of Dose-Response Analyses for Duration of Continuous Combined HT and Postmenopausal Breast Cancer Risk, MARIE Study, Germany, 2002–2005.

Model	Method	dose-response function OR(X=x vs X=0)	Deviance	Dev. diff. to null model	Dev. diff. to best FP2/ 1st stage model	d.f.	P ^a
0	Null model ^b		12254.19	0			
1	Categorization (original analysis)	$\exp(0.099 I_{X \in (1,5)}(x) + 0.59 I_{X \in [5,10)}(x) + 0.61 I_{X \in [10,15)}(x) + 0.55 I_{X \geq 15}(x))$	12129.84	124.35		4	<0.001
2	Linear	$\exp(0.042 x)$	12149.10	105.09		1	<0.001
<i>standard FP</i>							
3a	Linear (default)	$\exp(0.042 x)$	12149.10		18.14	1	<0.001
3b	FP1 ^c	$\exp(0.22 (\log(x + 1)))$	12136.40		5.44	2	0.66
3c	FP2	$\exp(1.80 ((x + 1)^{-2} - 1) - 2.58 ((x + 1)^{-1} - 1))$	12130.96	123.23		4	<0.001
4	Linear+z	$\exp(0.15 + 0.03 x)$	12145.09	109.10		2	<0.001
<i>FP-spike</i>							
<i>First stage</i>							
5a	Linear+z (default)	$\exp(0.15 + 0.03 x)$	12145.09		16.42	2	0.001
5b	FP1+z ^c	$\exp(-0.0054 + 0.24 \log(x))$	12132.67		4.00	3	0.136
5c	FP2+z	$\exp(-0.29 + 0.25 x - 0.069 x \log(x))$	12128.48	125.52		5	<0.001
<i>Second stage</i>							
5d	FP (dropping z) ^c	$\exp(0.24 \log(x))$	12132.67		0.006	1	0.95
	z (dropping FP)	$\exp(0.38)$	12162.16		29.49	1	<0.001

Abbreviations: d.f., degrees of freedom; Dev. diff., Deviance difference; FP1, first degree fractional polynomial; FP2, second degree fractional polynomial; HT, hormone therapy; OR, odds ratio.

^a P value referring to the given deviance difference from the previous columns

^b Deviance of the model without X but including all other covariates.

^c Results of the function selection procedure of the FP resp. FP-spike method

Table 3. Comparison of Dose-Response Analyses for Cement Dust Exposure and Laryngeal Cancer Risk, RHEIN-NECKAR-LARYNX Study, Germany, 1998–2000.

Model	Method	dose-response function OR($X=x$ vs $X=0$)	Deviance	Dev. diff. to null model	Dev. diff. to best FP2/ 1st stage model	d.f.	P^a
0	Null model ^b		1086.77	0			
1	Categorization (original analysis)	$\exp(1.13 I_{X \in (1,3000]}(x) + 1.15 I_{X > 3000}(x))$	1066.69	20.08		2	0.001
2	Linear	$\exp(0.85 \times 10^{-4} x)$	1076.66	10.11		1	0.003
<i>standard FP</i>							
3a	Linear (default)	$\exp(0.85 \times 10^{-4} x)$	1076.66		11.93	2	0.008
3b	FP1 ^c	$\exp(-1.14((x+1)^{-2} - 1))$	1066.70		1.97	2	0.374
3c	FP2	$\exp(-1.043((x+1)^{-2} - 1)) + 5.63 \times 10^{-14}((x+1)^3 - 1))$	1064.73	22.04		4	<0.001
4	Linear+z	$\exp(1.030 + 0.19 \times 10^{-4} x)$	1066.33	20.44		2	<0.001
<i>FP-spike</i>							
<i>First stage</i>							
5a	Linear+z (default) ^c	$\exp(1.030 + 0.19 \times 10^{-4} x)$	1066.33		3.37	2	0.339
5b	FP1+z	$\exp(1.043 + 5.63 \times 10^{-14} x^3)$	1064.73		1.76	3	0.414
5c	FP2+z	$\exp(1.25 - 9.58 \times 10^{-9} x^2 + 4.26 \times 10^{-13} x^3)$	1062.97	23.81		5	<0.001
<i>Second stage</i>							
5d	Linear (dropping z)	$\exp(0.85 \times 10^{-4} x)$	1076.66		10.33	1	0.001
	z (dropping Linear) ^c	$\exp(1.14)$	1066.70		0.37	1	0.55

Abbreviations: d.f., degrees of freedom; Dev. diff., Deviance difference; FP1, first degree fractional polynomial; FP2, second degree fractional polynomial; HT, hormone therapy; OR, odds ratio.

^a P value referring to the given deviance difference from the previous columns

^b Deviance of the model without X but including all other covariates.

^c Results of the function selection procedure of the FP resp. FP-spike method

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Table 4. Comparison of Dose-Response Analyses for Lifetime Hours of Asbestos Exposure and Lung Cancer Risk, Lung Cancer Study, Germany, 1988-1993.

Model	Method	dose-response function OR(X=x vs X=0)	Deviance	Dev. diff. to null model	Dev. diff. to best FP2/ 1st stage model	d.f.	P ^a
0	Null model ^b		951.94	0			
1	Categorization (original analysis)	$\exp(-0.025 I_{X \in (1,940]}(x) + 0.36 I_{X \in (940,5280]}(x) + 0.38 I_{X > 5280}(x))$	943.52	8.42		3	0.038
2	Linear	$\exp(0.18 \times 10^{-4}x)$	945.20	6.74		1	0.012
<i>standard FP</i>							
3a	Linear (default) ^c	$\exp(0.18 \times 10^{-4}x)$	945.20		2.83	1	0.42
3b	FP1	$\exp(0.0037((x+1)^{0.5}-1))$	943.03		0.66	2	0.719
3c	FP2	$\exp(0.0044((x+1)^{0.5}-1) - 1.71 \times 10^{-15}((x+1)^3-1))$	942.37	9.57		4	0.048
4	Linear+z	$\exp(0.14 + 0.14 \times 10^{-4}x)$	944.03	7.91		2	0.019
<i>FP-spike</i>							
<i>First stage</i>							
5a	Linear+z (default) ^c	$\exp(0.14 + 0.14 \times 10^{-4}x)$	944.03		2.752	2	0.43
5b	FP1+z	$\exp(0.014 + 0.0036x^{0.5})$	943.01		1.74	3	0.42
5c	FP2+z	$\exp(0.032 - 1.12x^{-2} + 0.0035x^{0.5})$	941.27	10.67		5	0.06
<i>Second stage</i>							
5d	Linear (dropping z) ^c	$\exp(0.18 \times 10^{-4}x)$	945.20		1.17	1	0.28
	z (dropping Linear)	$\exp(0.23)$	947.82		3.60	1	0.058

Abbreviations: d.f., degrees of freedom; Dev. diff., Deviance difference; FP1, first degree fractional polynomial; FP2, second degree fractional polynomial; HT, hormone therapy; OR, odds ratio.

^aP value referring to the given deviance difference from the previous columns

^bDeviance of the model without X but including all other covariates.

^cResults of the function selection procedure of the FP resp. FP-spike method