

# Modelling continuous predictors with a 'spike' at zero: multivariable approaches

Carolyn Jenkner<sup>a\*</sup>, Eva Lorenz<sup>b</sup>, Heiko Becher<sup>b</sup>, Willi Sauerbrei<sup>a</sup>

In epidemiology and clinical research, predictors often consist of an amount of individuals with a value of zero while the distribution of the remaining ones is continuous (variables with a spike at zero). Examples in epidemiology are smoking or alcohol consumption and in clinical research laboratory measures, sometimes caused by a lower detection limit of the measurement. Recently, an extension of the fractional polynomial (FP) procedure was proposed to deal with such situations. To indicate whether or not a value is zero, a binary variable is added to the model. In a two-stage procedure, it is assessed whether the binary variable and/or the continuous FP function for the positive part is required (FP-spike). In univariate analyses, FP-spike leads to functional relationships which are easy to interpret.

If more than one spike variable is present, several approaches are possible. Possible methods of handling them are strongly dependent upon the bivariate or multivariable distribution of the zero and non-zero values. Within this paper, different statistical and distributional issues are analyzed and illustrated through three distinct datasets. Copyright © 2012 John Wiley & Sons, Ltd.

**Keywords:** regression modelling; spike at zero

## 1. Introduction

In clinical and epidemiological trials one often faces the problem of continuous covariates with a fraction of zero values. Typical examples are a measure for smoking e.g. cigarettes smoked per day or pack years, alcohol intake, or the number of positive estrogen receptors. This paper will focus on how the effect of such a variable on potential outcomes can be modeled. This situation was called differently in the literature so far e.g. a semi-continuous variable (e.g. Olsen, Schafer (2001) [6]), a spike at zero (e.g. Schisterman et al., 2006 [12]; Robertson, 1994 [8]), mass at zero (Lachenbruch 2001 [4]) or clumps of zero (e.g. Hallstrom 2010 [2]). So far, this situation which we will be calling spike at zero in the following was often ignored. In some cases the variable was dichotomized. That means, the continuous part is completely ignored. Sometimes, it was modeled including a binary indicator for the zero part but assuming the continuous part to be linear. In 2010, a method for the case with one spike variable was proposed by Royston et al [10] who included an indicator

<sup>a</sup> Institute of Medical Biometry and Medical Informatics, Freiburg University Medical Centre

<sup>b</sup> Epidemiology and Biostatistics Unit, Medical Faculty, University of Heidelberg

\*Correspondence to: Institute of Medical Biometry and Medical Informatics, Stefan-Meier-Str. 26, 79104 Freiburg. E-mail: jjenkner@imbi.uni-freiburg.de

Contract/grant sponsor: DFG Project XXX

for the spike but then model the continuous part using fractional polynomials (FP). Some theoretical investigations were performed in Lorenz (2010) [5] assessing the correct model under certain distributional assumptions with a spike at zero situation. The univariate procedure which was slightly modified in 2012 [1] now provided the basis for further research and for possible extensions to a multivariable setting and specifically to a setting with more than one spike variable present. These extension are not straightforward especially in the case of correlated spike variables. There are several possible ways of possible multivariable extensions of this procedure to find a suitable model. In a first step we will present possible bivariate methods to deal with situations in which there are two spike variables to analyze. The proposed methods will then be applied to real datasets and compared. Furthermore, we also propose a possible extension to situations with more than two spike, however, there are still a lot of unsolved questions and further research is needed.

## 2. Data

We will have a look at three different data sets as all of them have different properties concerning the spike variables. With the help of these three data sets, we will illustrate the proposed strategies. The specific situation of a spike at zero variable was ignored in all original analyses.

### 2.1. Study on Breast Cancer

The first data set that we will be using is from a randomized trial with 686 patients with primary node positive breast cancer. There were 299 events for recurrence free survival (RFS) and the recruitment took place from 1984 to 1989. Seven prognostic factors were assessed, but we will focus on the effect of the estrogen and progesterone receptor on RFS because they both have a spike at zero. In the data 11% of the patients take zero for the value of the estrogen receptor and 13 % for the value of the progesterone receptor. In 9% of the patients both receptors take value zero. In real data, we often face the situations that the variables of interest are correlated. The effect of hormonal treatment with tamoxifen was investigated in this old trial. For some years, it is generally known that there is a strong interaction between the estrogen receptor and tamoxifen. Thus, we will also consider whether the effect of estrogen is different in the two (yes or no) tamoxifen subgroups. For more details on the study see Schumacher et al. (1994) [13].

### 2.2. Study on laryngeal cancer

The second example that we will be investigating is a matched case-control study with 1026 observations. Of these, 257 were cases and 769 controls. The study assessed the possible risk factors for laryngeal cancer. Our analysis will focus on the two spike variables “smoking” and “alcohol”. The first spike variable “cigarette consumption” is measured in pack-years. The second spike variable “alcohol consumption” is measured in gram per day. The distribution is a bit different than in the first dataset. There are 21 % non-smokers in the study and 6 % who do not drink alcohol but there is only a small percentage of 1.4% who neither smoke nor drink alcohol. For further details on the study see Ramroth et al (2004) [7].

### 2.3. Study on lung cancer

In this case-control study with 1004 cases and 1004 controls the effect of occupational factors on lung cancer in women was assessed. We will have a look at three variables with a spike at zero, smoking, concentration of asbestos exposure and days in high-risk occupations. This situation brings up further challenges in the analysis. The spike percentages are relatively high here. Most of the patients, 82 % did not work in a high-risk occupation. There are 21 % non-smokers in the study and 68 % who had no asbestos exposition. For further details on the study see Jöckel et al. (1998) [3].

## 3. Methods

### 3.1. Fractional polynomials

The simplest way of modelling a continuous covariate is untransformed assuming a linear relationship. In most cases, however, assuming linearity is not flexible enough to adequately capture the relationship of two variables. Thus, one can extend the class of possible functional relationships e.g. allowing some power transformation models,  $\beta_1 x^p$ . Several choices of  $p$  are possible. Royston and Altman proposed a set of values  $S = \{-2, -1 - 0.5, 0, 0.5, 1, 2, 3\}$  of which the power term  $p$  is to be chosen ( $x^0$  denotes  $\log x$ ). This group of functions is called first degree fractional polynomials. An extension is a two-term FP2 function. FP2 functions with powers chosen from  $S$  are defined as  $\beta_1 x^{p_1} + \beta_2 x^{p_2}$ . In a function selections procedure (FSP), the best FP function is then chosen. It is illustrated in figure 1. First, according to a deviance criterion the best FP2 and FP1 function are fitted. Then in a closed test procedure it is assessed whether the most complex function, the best FP2 function is significantly better than the null model. If this test is not significant we stop concluding that the covariate does not have an influence. Otherwise, we continue testing the best FP2 against the linear model. If the FP2 is not significantly better, we stop concluding that the best model is the linear model. Otherwise, we will go on testing the best FP2 against the best FP1. If the test is not significant we will conclude that FP1 is the best model. Otherwise the final model will be FP2. It may also be sensible to only allow FP1 functions. Possible reasons are restriction to monotonic functions and an increased power if the effect is linear. For further details see Royston and Sauerbrei [9]. The usual Multivariable Fractional Polynomial (MFP) procedure combines variable selection using backward elimination with the above described function selection procedure.

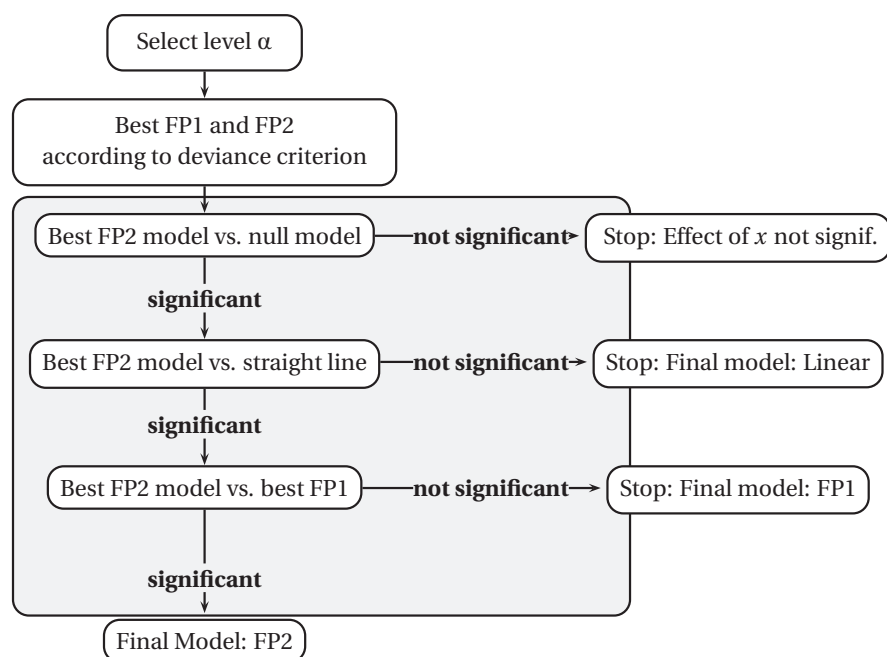


Figure 1. Function Selection Procedure of the FP Procedure (FSP)

### 3.2. One spike variable

The method which our extension is based on was proposed recently in Royston and Sauerbrei (2008) [9] and further explained in Royston et al. (2010) [10]. We consider an outcome variable  $y$  and a non-negative continuous covariate  $x$

which has a certain proportion of zero values. The FP-Spike procedure is an extension of the usual FP procedure but in addition it adds a binary variable  $z$  which distinguishes between  $x = 0$  ( $Z = 1$ ) or  $x > 0$  ( $Z = 0$ ) and selects the functional form in two stages. Based on theoretical considerations and results of a simulation study, Becher et al. (2012) [1] illustrate that a slightly modified version is preferable. This version will be used in the following. In a first stage, a fractional polynomial is selected including the additional binary variable  $z$  in all models. The function selection procedure is the same as described above in figure 1, but with two important differences. The binary variable  $z$  and the FP function is modeled only for the positive values ( $x > 0$ ). Because of  $z$ , the degrees of freedom of the likelihood ratio test statistic have to be adapted. The model is

$$\varphi(x, \beta) = \beta_0 \mathbf{1}_{\{x_1=0\}} + (\beta_1 f(x) +) \mathbf{1}_{\{x_1>0\}} + \beta_{cons},$$

where  $\varphi$  represents the outcome, thus it depends on what type of regression model we are working with. It can be either linear predictor of the odds or hazard ratio, or of a continuous covariate in a linear model. We chose to model  $f(x)$  with fractional polynomials, however is also possible to use any other model building technique for continuous covariates such as splines. In linear and logistic regression we usually have a constant term in the model, here called  $\beta_{cons}$ , but not in the Cox model. In the following notation we ignore this constant part.

The notation will be the following. For each positive variable  $x$  with a spike a binary indicator  $z$  is generated indicating if  $x$  is zero or positive.

$x$  = continuous variable

$z$  = binary indicator

$\tilde{x} := (x, z)$

The detailed procedure is as follows:

0. The maximum permitted complexity for the continuous part of the class of FP functions is chosen; e.g. FP2 or FP1 respectively. The suggested default is FP2. The procedure is now explained for FP2. Corresponding versions for higher or lower FP-classes are obvious.
1. All FP transformations are only applied to positive values. The transformation will be called  $\text{FP}_i^{\text{Spike}}$  where  $i$  is the degree of the FP transformation.
2. To select the 'best' function a nominal P-values  $\alpha$  is chosen. A typical value is  $\alpha = 0.05$ . Taking  $\alpha = 1$  selects the function with the lowest deviance from the class of the most complex permitted FP functions (no function selection procedure).
- 3 Function selection procedure for variables with a spike (FSP-Spike)
  - 3.1 All transformations for  $\tilde{x}$  are fitted. The best  $(\text{FP2}_{\text{Spike}}(x) + z) =: \text{FP2}^*$  and  $(\text{FP1}_{\text{Spike}}(x) + z) =: \text{FP1}^*$  are chosen according to deviance criterion. Correspondingly,  $(x + z) =: \text{Lin}^*$ .
  - 3.2 In the first stage,  $\text{FP2}^*$  is compared to the null model on five d.f. If the likelihood ratio test is not significant, the variable is considered to have no influence and the algorithm stops. Otherwise,  $\text{FP2}^*$  is compared to  $\text{Lin}^*$  (3 df). If the test is not significant, the algorithm stops choosing  $\text{Lin}^*$ . Otherwise,  $\text{FP2}^*$  is tested against  $\text{FP1}^*$ . If the test is not significant, the algorithm stops choosing  $\text{FP1}^*$ . Otherwise  $\text{FP2}^*$  will be the final function ( $\text{FP}^*$ ) in the first stage.
  - 3.3 In the second stage, the two components of  $\text{FP}^*$  are each tested for removal from the model. If both parts are significant, then the final model includes both; if one or both parts are non-significant, then the one with the larger p-value is removed. In the latter case, the final model comprises either the binary dummy variable  $z$  or the selected FP function  $(\text{FP2}_{\text{Spike}}(x), \text{FP1}_{\text{Spike}}(x), x)$ . If only an FP function is selected, then the spike at zero plays no specific part.

### 3.3. One spike variable - adjustment for further covariates

In order to extend the model to a multivariable setting, there are several possibilities.

The first obvious multivariable extension is the consideration of one spike variable and a necessary adjustment for further covariates, summarized in an adjustment index. If the adjustment model is pre-specified or determined independently of the spike variable by a suitable selection procedure, the function for the spike variable can be determined by the usual univariate procedure, but incorporating the adjustment index in all steps of the spike procedure. Obviously, the adjustment index can have a severe influence on the selected function for the spike variable.

A straight forward way for first extensions to a model with further covariates is, to use FSP-Spike for variables with a spike and adjust for and, thus, include other variables which can be either pre-specified or selected using the normal FSP of the MFP algorithm (Extension of the MFP algorithm for the inclusion of single spike variables). In this case, the procedure is obvious. In the following we will ignore the consideration of other variables without a spike, but for procedures such as MFP or other stepwise approaches handling of adjustment variables is straightforward.

### 3.4. Two spike variables

If we have more than one spike variable, the situation gets a bit more complex, and there are several ways that could be used to handle them. We could

- a. consider the spike variables separately and use the FP-Spike procedure for each variable
- b. considering the correlation between spikes
  - if they are uncorrelated: see a.
  - if they are correlated: we could create a combination of dummy variables
- c. create new variable indicating that both are zero or at least one of the two variables is non-zero.
- d. consider submodels

These ideas are a non exhaustive collection of possible ways of the analysis of such a situation. Which method to chose may depend on the specific situation and especially the distributions of the two spike variables as indicated in table 1. Table 1 defines four categories of observations, e.g. observations in case A take value zero for both covariates. Not all of the proposed methods are suitable for all possible distributions in table 1. Some of the methods described in the following are e.g. more suitable in handling situations in which the number of observations in categories B and C are high. That means if we have several observations in which only one of the covariates takes value 0. Other approaches may be preferable in situations in which the amount of observations that take value 0 for both covariates or also a continuous value for both is relatively low, that means the number of observations in categories B and C are rather high. Four different methods are now described and compared.

**3.4.1. Consider spike variables separately** One could use the FP-Spike procedure separately for each variable, that means that as is the case with only one variable we use the usual MFP procedure for variables without spike and the FP-Spike procedure separately for each variable that has a spike. The FP-Spike procedure is applied for both variables separately. We will distinguish them using indices: Spike variable 1 will be coded as  $x_1$  and  $z_1$ , and Spike variable 2 as  $x_2$  and  $z_2$ .

0./1. Cf. FP Spike. Choose values for each variable separately.

2. Nominal P -values  $\alpha_1$  and  $\alpha_2$  for variable and function selection are chosen for each spike variable. Typical values are  $\alpha_1 = \alpha_2 = 0.05$ . Values may differ among variables. Taking  $\alpha_1 = 1$  for a given variable forces it into the model (no variable selection). Taking  $\alpha_2 = 1$  for a continuous variable forces the most complex permitted FP function to be fitted for it (no function selection).
3. For each spike variable  $x_1$  and  $x_2$  binary indicators  $z_1$  and  $z_2$  are generated.

## 4. Function selection procedure (FSP-MFP-Spike1)

- 4.1 A model is fitted including  $\tilde{x}_1$  and  $\tilde{x}_2$ . The visiting order of the predictors is determined according to the P -value for omitting either  $\tilde{x}_1$  or  $\tilde{x}_2$ . The most significant predictor is visited first and the least last. Assume that the variables  $\tilde{x}_1$  and  $\tilde{x}_2$  have been arranged in this order, which is retained in all cycles of the procedure.
- 4.2 Let  $c = 0$ , to initialize the cycle counter.
- 4.3 1. Cycle: FSP-Spike is applied to  $x_1$  at the  $\alpha = \alpha_1$  level adjusted for  $\tilde{x}_2$ .  $(x_1^*)_{c1}$  is the selected function for  $\tilde{x}_1$  (it may consist of only the binary indicator, only a continuous function, both or might be dropped if neither part is significant) in the first cycle, which is indicated with c1.
- 4.4 FSP- Spike is applied to  $\tilde{x}_2$  adjusted for  $(x_1^*)_{c1}$ .  $(x_2^*)_{c1}$  is the selected function for  $\tilde{x}_2$ .
- 4.5 2. Cycle: FSP Spike is again applied to  $\tilde{x}_1$  adjusted for  $(x_2^*)_{c1}$ . The selected function is  $(x_1^*)_{c2}$ . Then FSP Spike is applied to  $\tilde{x}_2$  adjusted for  $(x_1^*)_{c2}$  leading to  $(x_2^*)_{c2}$ .
- 4.6 The algorithm stops if  $(x_1^*)_{ci} = (x_1^*)_{c(i+1)}$  and  $(x_2^*)_{ci} = (x_2^*)_{c(i+1)}$ .

**3.4.2. Combination of dummy variables** If one wants to take a possible correlation into account one could consider a combination of dummy variables. These dummy variables distinguish 4 categories which are also indicated in table 1. Thus, they distinguish if both covariates take value zero, if only one of them equals zero and the other takes a continuous value or if both of them take a non-zero value. These three dummy variables take the part of the single binary indicator in the univariate case.

## 0.-2. As in version: 2 Spike variables

3. 3 dummy variables are generated indicating if  $x_1$  and  $x_2$  are zero ( $d_1 = 1$  if  $x_1 = x_2 = 0$ , only  $x_1$  is zero ( $d_2 = 1$  if  $x_1 = 0$  and  $x_2 > 0$ ) or only  $x_2$  is zero ( $d_3 = 1$  if  $x_1 > 0$  and  $x_2 = 0$ ).
4. Function selection procedure (FSP-MFP-Spike2)
- 4.1 A model is fitted which includes  $d_1, d_2, d_3$  and the untransformed  $x_1$  and  $x_2$ . The visiting order of the predictors is determined according to the P -value for omitting either  $x_1$  or  $x_2$ . The most significant predictor is visited first and the least last. Assume that the variables  $x_1$  and  $x_2$  have been arranged in this order, which is retained in all cycles of the procedure.
- 4.2 Let  $c = 0$ , to initialize the cycle counter.
- 4.3 The normal FSP (cf. MFP Royston et al.) is applied to  $x_1$  at the  $\alpha = \alpha_1$  level adjusted for  $x_2$  (untransformed) and the three dummies.  $(x_1^*)_{c1}$  is the selected function for  $x_1$ .
- 4.4 FSP is applied to  $x_2$  adjusted for  $(x_1^*)_{c1}$ , and  $d_1, d_2, d_3$ .  $(x_2^*)_{c1}$  is the selected function for  $x_2$ .
- 4.5 In the 2nd cycle, FSP is again applied to  $x_1$  adjusted for  $(x_2^*)_{c1}$ . The selected function is  $(x_1^*)_{c2}$ . Then FSP Spike is applied to  $x_2$  adjusted for  $(x_1^*)_{c2}$  leading to  $(x_2^*)_{c2}$ .
- 4.6 The algorithm stops if  $(x_1^*)_{ci} = (x_1^*)_{c(i+1)}$  and  $(x_2^*)_{ci} = (x_2^*)_{c(i+1)}$ .

		Dummy			
$x_1$	$x_2$	Category	$d_1$	$d_2$	$d_3$
0	0	A	1	0	0
0	> 0	B	0	1	0
> 0	0	C	0	0	1
> 0	> 0	D	0	0	0

**Table 1.** The observations of two spike variables separated into 4 Categories with possible dummy coding

**3.4.3. Dummy 1** Another possibility is to create a new variable that indicates whether both variables are zero ( $d_1 = 1$  if  $X_1 = 0$  and  $X_2 = 0$ ,  $d_1 = 0$  otherwise) or at least one of them is greater than zero ( $Z = 1$  if  $X_1 = 0$  or  $X_2 = 0$ ,  $Z = 0$  otherwise).

otherwise). This methods uses only  $d_1$  of the three dummy variables. If it is reasonable or not strongly depends on interpretative issues because here we combine categories B, C, and D and only distinguish them from category A. Furthermore, with categories B and C we might still have the problem of a spike at zero. The detailed procedure is similar to the ones already described with only some slight changes:

0. -2. As in version: 2 Spike variables

3. A new variable  $d_1$  is generated indicating that both Spike variables take value zero-

4. Function selection procedure (FSP-MFP-Spike3)

4.1 cf. FSP-MFP-Spike 2, but replace 3 dummies with only  $d_1$ .

**3.4.4. Consider submodels** Another way of handling more than one spike variables is to consider the four categories in submodels if the sample size allows it. That means we build a spike submodel in each category as displayed in the following equation.

$$\begin{aligned} f(x) = & \beta_{0,0} * \mathbf{1}_{\{x_1=0\} \cap \{x_2=0\}} \\ & + (\beta_{0,1} + \text{FP}(x_2)) * \mathbf{1}_{\{x_1=0\} \cap \{x_2>0\}} \\ & + (\beta_{1,0} + \text{FP}(x_1)) * \mathbf{1}_{\{x_1>0\} \cap \{x_2=0\}} \\ & + \text{FP}(x_1, x_2) * \mathbf{1}_{\{x_1>0\} \cap \{x_2>0\}} \end{aligned}$$

Or in other words we include the following 7 variables in our model:

$x_2$	$x_1$	
	0	> 0
0	$z_1 = \begin{cases} 1, & \text{if } x_1 = 0, x_2 = 0 \\ 0, & \text{else} \end{cases}$	$z_2 = \begin{cases} 1, & \text{if } x_1 = 0, x_2 > 0 \\ 0, & \text{else} \end{cases}$ $z_4 = \begin{cases} x_1, & \text{if } x_1 > 0, x_2 = 0 \\ 0, & \text{else} \end{cases}$
> 0	$z_3 = \begin{cases} 1, & \text{if } x_1 > 0, x_2 = 0 \\ 0, & \text{else} \end{cases}$ $z_5 = \begin{cases} x_2, & \text{if } x_1 = 0, x_2 > 0 \\ 0, & \text{else} \end{cases}$	$z_6 = \begin{cases} x_1, & \text{if } x_1 > 0, x_2 > 0 \\ 0, & \text{else} \end{cases}$ $z_7 = \begin{cases} x_2, & \text{if } x_1 > 0, x_2 > 0 \\ 0, & \text{else} \end{cases}$

**Table 2.** Definition of the 7 variables used in the submodel approach

With these 7 variables all relevant odds ratios can be calculated. In terms of model building, it may be reasonable to restrict the choice of the FP-models for  $z_4$  and  $z_6$  in that they have to have the same functional form that means the same choice of powers for the FP models but we allow different coefficients in the respective groups. A possible advantage here is that some of the subgroups might be relatively small and thus it is easier to estimate only the coefficient and not having to determine the functional form. However, if the groups are big enough one can also allow that the FPs can be selected completely independent in the different groups. The respective odds ratios can be given as follows (cf. Becher et al (2012) [1]):

$$OR_{X_1=x_1^*, X_2=x_2^* \text{ vs } X_1=x_{10}, X_2=x_{20}} = \frac{f_1(x_1^*, x_2^*) / f_0(x_{10}, x_{20})}{f_1(x_{10}, x_{20}) / f_0(x_1^*, x_2^*)}$$

### 3.5. More than two Spike variables

If we face the situation that there are more than 2 spike variables the simplest approach is as in 3.3. Some extensions of the other approaches for 2 variables are also possible in principle. The key issue is the correlation structure of the variables.

Nb. of spikes	Multiv.	Method	Summary
1	no	FP-Spike	Inclusion of binary indicator
1	yes	FP-Spike + index	The FP-Spike procedure is used including an prognostic index in every step
1	yes	FP-Spike adjusted	Further variables for adjustment predefined or determined by usual FSP for variables without spike at zero are added to the model, FP-Spike for variables with spike
2	yes	Separately	Correlation structure is not considered. FP-Spike is used separately for both variables
2	yes	Combination of dummies	Three dummy variables distinguish between 4 categories of observations and thus replace the binary indicator in the univariate case. All 3 dummies are kept in the model,
2	yes	Dummy 1	Only the first dummy indicating that both variables are zero is kept in the model
2	yes	Submodels	For each of the 4 categories a separate functional relationship is estimated. 3 dummies are included plus additional continuous variables which take a positive value only if the other variable is zero. Restrictions, that all functions must have the same power terms are possible.
>2	yes	Log Linear model	Build a loglinear model to investigate the correlation structure of the dichotomized variables (0, > 0). If only two-way relationships, use the above described strategies

**Table 3.** Summary of FP Spike approaches for one or more spike variables

$x_1$	$x_2$	Separately		Dummy		New		Submodels	
0	0	$z_1$	$z_2$	$d_1$	$d_1$	$d_1$	$d_1$	$z_1$	$z_1$
> 0	0	$FP_{Spike}(x_1)$	$z_2$	$d_2/FP(x_1)$	$d_2$	$FP(x_1)$	$FP(z_4)$	$z_3$	
0	> 0	$z_1$	$FP_{Spike}(x_2)$	$d_3$	$d_3/FP(x_2)$		$FP(x_2)$	$z_2$	$FP(z_6)$
> 0	> 0	$FP_{Spike}(x_1)$	$FP_{Spike}(x_2)$	$FP(x_1)$	$FP(x_2)$	$FP(x_1)$	$FP(x_2)$	$FP(z_5)$	$FP(z_7)$

**Table 4.** Overview of differences between multivariable Spike methods for 2 variables with a spike

To gain insight we propose to investigate the correlation structure of the binary variables (0, > 0) for all variables with a spike at zero. This can be done by building a log-linear model to test for interactions between the covariates. Using this information, we can try to choose a suitable strategy of the already proposed methods. Loglinear models are a method for the evaluation of dependence structure in multi-way tables. They model the probability of the cell frequencies. A more detailed description will be given using an example in 4.4.

### 3.6. Summary

Table 3 and 4 summarize the proposed methods. Table 4 tries to give an overview which observations have an influence on which coefficient in the different models that means which observations influence the estimation of the respective coefficients. The left column distinguishes the four different types of observations. In the adjacent columns the 4 methods for two spike variables are documented. Here,  $FP(x_1)$  and  $FP_{Spike}(x_2)$  stand for the coefficients in the continuous part of the model the  $z$  and  $d$  variables for the coefficients of the dummy variables.

Table 3 gives an overview of all different methods described in this paper. In the following chapter, these methods will be illustrated and compared.

<b>a) ALL PATIENTS</b>					
<b>First stage</b>					
	Deviance	Dev.Diff.	d.f.	P	Power
FP2+SPIKE	3552.5				-2, -1
Null		23.8	5	<0.001	
Linear+SPIKE		14.1	3	0.003	1
FP1+SPIKE		1.9	2	0.391	-0.5
<b>Second stage</b>					
FP1 + SPIKE	3554.4				
Dropping SPIKE		14.4		<0.001	
Dropping FP1		14.5		<0.001	
<b>b) TAM SUBGROUP</b>					
<b>First stage</b>					
	Deviance	Dev.Diff.	d.f.	P	Power
FP2+SPIKE	915.2				-2, -2
Null		20.8	5	<0.001	
Linear+SPIKE		5.3	3	0.148	1
FP1+SPIKE		1.9	2	0.101	3
<b>Second stage</b>					
Linear + SPIKE	920.6				
Dropping SPIKE		14.9		<0.001	
Dropping Linear		<0.1		0.876	

**Table 5.** Breast Cancer Study: Details to derive the FP-Spike models for Estrogen receptor. a) all patients. b) subgroup with hormonal treatment

## 4. Results

To illustrate the different procedures we will analyze the three datasets described above. In all three datasets we have at least two Spike variables. The situations we will be dealing with are very diverse, however. Each dataset will bring up new issues for possible ways of analysis.

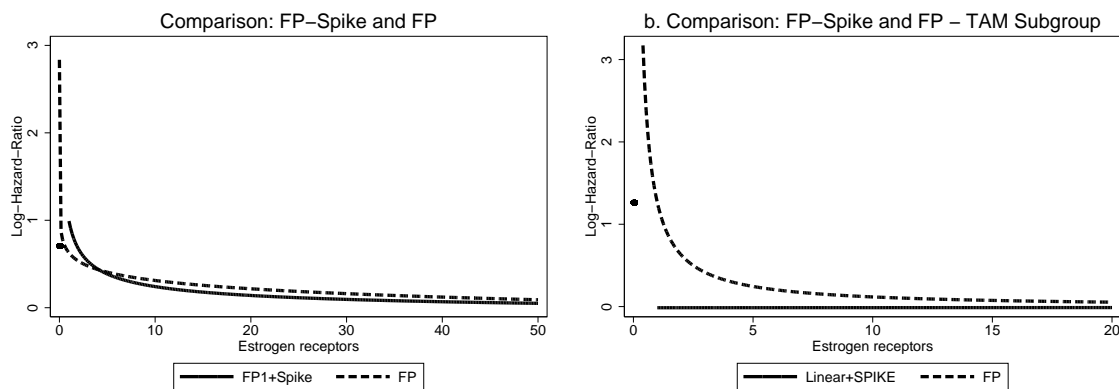
### 4.1. One spike variable

To illustrate the advantages of the FP-Spike procedure, we will compare different models for the estrogen receptor in the breast cancer study. In table 5, the results and details of the two stage procedure of FP-Spike are given. In the first stage, the best FP model including the SPIKE variable is chosen using the FSP. At the significance level  $\alpha = 0.05$  FP2+Spike is significantly better than the null model and the linear+SPIKE model but not significantly better than FP1+Spike. So the FP1 function with the SPIKE dummy variable is chosen in the first stage. In the second stage, both the SPIKE and the continuous FP1 part are tested for removal. Here, both parts cannot be left out without receiving a significantly worse model and, thus, both are kept in the final model.

Figure 2a. shows the functional forms of both the usual FP and the FP-Spike. One can see that the usual MFP tends to infinity close to zero which makes it hard to get a satisfying interpretation. The FP-Spike model produces a point estimate for all zero values and is, therefore, very easy to interpret.

The corresponding functional relationships can easily be given. The analysis with the usual FP method leads to the following:

$$\varphi_{FP}(x) = -0.137(\ln(x) - 4.577)$$



**Figure 2.** Breast Cancer Study: FP-Spike vs. FP for Estrogen receptor, univariate. In the TAM subgroup, the final model does not include the continuous part. It is only included for illustrative purposes here.

The problem of zero values is solved by transformation of the initial dataset which can be seen in the formula (the data is shifted so that the problem does not occur any more). For this model the deviance is 3557.2. We receive the following functional relationship if we use the FP-Spike procedure:

$$\varphi(x) = 0.706 * \mathbf{1}_{SPIKE=0} + 1.096 * (x^{-0.5} - 0.0956) * \mathbf{1}_{SPIKE=1}$$

Here, the deviance is 3554.4, thus, we have a slight decrease in deviance. The interpretative advantage of the results of the FP-Spike model becomes even clearer if we have a look at a subgroup of the study population, namely the patients that received a hormonal treatment with tamoxifen. Estrogen receptor positive or not has an influence on survival, however the value of the estrogen receptor does not provide further information. Table 5 shows that in the first stage a linear function including a binary spike variable is chosen. In the second stage, however we can see that dropping the linear part of the model does not significantly worsen the fit and, thus, our final model will only consist of the SPIKE, the binary indicator.

This can also be seen in figure 2. Here, the usual FP is plotted together with the FP-Spike function. For illustrative purposes, the linear part is still included, however, in the final model it is not included, as only the binary indicator is significant.

## 4.2. One spike variable - multivariable adjustment

We will now illustrate, the methods stated in section 3.3 with a data example. We'll again have a look at the estrogen receptor level and then compare the univariate FP-Spike analysis, the FP-Spike analysis adjusted for a binary progesterone receptor indicator and in a further model adjusted for a pre-specified prognostic index, which was calculated out of the originally published model in Sauerbrei et al (1999) [11] and includes the prognostic factors progesterone receptor, age, tumor grade, number of positive lymph nodes and hormonal therapy. In the two latter models, we investigate the additional prognostic value of the estrogen receptor as variable with spike.

In the unadjusted analysis, the estrogen receptor value is highly significant as a prognostic factor. If we include a further covariate, e.g. the progesterone receptor value, here only included as a binary indicator, the prognostic effect of estrogen is lower. In the final model, only the continuous part is selected whereas the binary indicator is dropped. Adjusting for further factors combined in an index data-dependently derived in Sauerbrei et al. (1999) [11], estrogen completely loses its prognostic value. This example is a nice illustration of the effect that neglecting multivariate analyses can lead to biased results. In the following, we will present the results of the analyses of two spike variables.

	Unadjusted		Adjusted for			
			PgR (binary)		Index	
	P	FP-Powers	P	FP-Powers	P	FP-Powers
First stage						
FP2+ Spike		(-2,-1)		(-0.5,3)		(-2,3)
Null	< 0.001		0.017		0.614 (*)	
LIN+SPIKE	0.003		0.008		0.470	
FP1+SPIKE	0.391 (*)	(-0.5)	0.546 (*)	(-0.5)	0.436	3
Deviance (1. st.)	3554.4		3544.2		3422.2	
Second stage						
drop SPIKE	<0.001		0.463		-	
drop FP-part	<0.001		< 0.001		-	
Deviance (Final)	3554.4		3544.6		3423.2	

**Table 6.** Breast Cancer Study: Details of FP-Spike for estrogen receptor with and without adjustment for further covariates.(\*) Model selected in first stage.

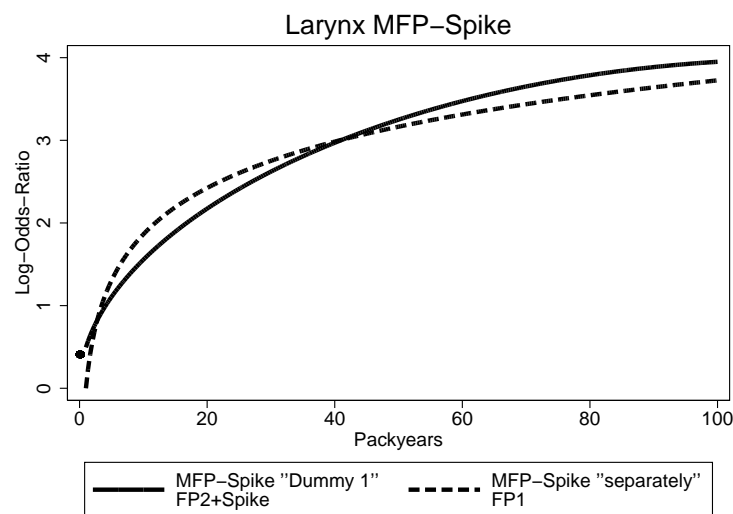
	Coef.	Std.Err.	P
<b>1. Separately</b>			
$z_{\text{pack}}$	dropped	(p=0.51)	
Packyears (0)	0.809	0.079	<0.001
$z_{\text{alc}}$	dropped	(p=0.07)	
Alcohol (1)	0.005	0.001	<0.001
Deviance	851.3		
<b>2. Dummy Combination</b>			
d1	0.709	1.117	0.525
d2	0.620	0.349	0.076
d3	0.303	0.519	0.560
Packyears(0)	0.842	0.107	<0.001
Alcohol (1)	0.006	0.001	<0.001
Deviance	847.7		
<b>3. Dummy 1</b>			
d1	0.410	1.080	0.704
Packyears (0.5)	0.495	0.500	< 0.001
Packyears (2)	-0.0001	0.324	0.001
Alcohol (1)	0.005	0.001	< 0.001
Deviance	847.1		

**Table 7.** Laryngeal cancer study: Multivariable Spike Model - three different methods of analyzing

## 4.3. Two spike variable

In the case of more than one spike variable, the Laryngeal cancer dataset will be used to illustrate every method so that we have a direct comparison of all of them.

**4.3.1. Consider spike variables separately** In the easiest case, one would analyze all spike variables separately. The FP-Spike procedure described is used for all spike variables with additional rules for the visiting order, that means for each variable with a spike at zero we will include a binary indicator, then fit the best model and then check whether we need both the indicator and the continuous part in the model. The Larynx study was a matched case-control study. Thus, conditional logistic regression will be used for the analysis. In the separate analysis, we can see that for both variables the FP1 or the linear part are significant. The indicators do not improve the fit here and, therefore are dropped. In table 7, the details of the specified model are given.



**Figure 3.** Laryngeal cancer study: Visual comparison of two possible multivariable analyses: “separately” and “Dummy 1”. The dot represents the coefficient of the binary variable, thus, the point estimate for non-smokers.

Control(0)	Alk	
	>0	0
Smok		
>0	540(70.2%)	26 (3.4%)
0	190 (24.7%)	13 (1.7%)
Case(1)	Alk	
	>0	0
Smok		
> 0	231(89.9%)	17(6.6%)
0	8 (3.1%)	1 (0.4%)

**Table 8.** Lung Cancer Study: distribution of the 4 categories in cases and controls

**4.3.2. Combination of dummy variables** In this analysis, the model for the continuous part is pretty similar to the model in the case in which we analyzed both variables separately. The same functional forms are chosen for smoking and alcohol intake and the coefficients are nearly identical. The advantage of having 3 dummy variables combining some of the categories instead of including to separate ones might be that possible correlation effects can be caught.

**4.3.3. Dummy 1** In figure 3, one can see the difference between the functional relationships modelling the variables separately. The detailed model can be found in table 7. The functional relationship is very similar. We can see that the dummy  $d_1$  is not significant. This is also visible in figure 3 in which the value of the indicator could also be more or less the extension of the continuous model. The amount of observations which take zero for both *alcohol* and *smoking* is very low (only 1.7 % in controls and 0.7% in cases), thus, there are two few observations to draw conclusions. This approach might therefore not be suited for such distributional situation but might be preferable in cases in which there is a strong relationship between the two spikes (that means the amount of observations for which both variables take value zero is relatively high).

**4.3.4. Consider submodels** In the following example, we choose  $x_1$  = Smoking and  $x_2$  = Alcohol intake and will have a look at the study on laryngeal cancer. The variables for the analysis are defined as in table 4. With the submodel method we now fit a functional form for every category using conditional logistic regression. The results can be found in the following table.

	Coefficient	Std. Dev	P
$z_1$	0.775	1.121	0.489
$z_2$	0.341	0.582	0.559
$z_3(1)$	2.753	0.771	< 0.001
$z_4$	0.020	0.018	0.253
$z_5(1)$	0.007	0.005	0.176
$z_6(0)$	0.861	0.111	< 0.001
$z_7(1)$	0.006	0.001	< 0.001
Deviance	847.1		

**Table 9.** Larynx dataset: Analysis with the submodel approach. Due to a relatively small sample size in each category the approach might not be suitable in this setting.

Risk	Asbestos	Smok	n(%)	Exp. (Ind.)	Cases	Controls
0	0	0	396 (19.7)	213	250 (24.9)	146 (14.6)
0	0	1	19 (1.0)	38	3 (0.3)	16 (1.6)
0	1	0	351 (17.5)	466	203 (20.2)	148 (14.7)
0	1	1	36 (1.8)	84	6 (0.6)	30 (3.0)
1	0	0	188 (9.4)	321	91 (9.1)	97 (9.7)
1	0	1	27 (1.3)	58	3 (0.3)	24 (2.4)
1	1	0	766 (38.1)	701	389 (38.7)	377 (37.5)
1	1	1	225 (11.2)	127	59 (5.9)	166 (16.5)

**Table 10.** Lung Cancer study: Distribution of categories, absolute and relative frequencies in the dataset and in cases and controls, number of expected observations under independence

	Binary variable			x-variable		
	All	Case	Control	All	Case	Control
Corr (R,A)	0.36	0.39	0.31	0.39	0.41	0.34
Corr (R,S)	0.19	0.18	0.17	0.20	0.14	0.19
Corr (A,S)	0.15	0.15	0.14	0.17	0.13	0.17

**Table 11.** Lung cancer study: Spearman correlation coefficients in cases and controls

This conditional logistic regression model describes every category with a kind of submodel. That means that we do not have to collapse categories if we want to model the continuous part, as we had to do using the dummy variables. Thus, this approach is more precise. However, if there are only few observations in the subcategories the chosen models might not be very stable. One possibility here is to chose a functional relationship for the biggest category which will in most cases be the category where both variables are positive and then restrict the FP form of the other categories, namely the power terms, to those selected before and only estimate new beta coefficients.

#### 4.4. More than two spike variables

In the study on lung cancer, there are three variables with a spike at zero. The simplest approach is to consider the spike variables separately and proceed with the analysis in the usual way. However, such an approach is unsuitable if the variables are correlated. For pairwise considerations we have proposed some strategies in the preceding chapter. For three or more spike variables we propose a preliminary step to investigate the correlation structure of the dichotomized (0,0) variables in a log-linear model. For variables being independent of the others we can proceed as in the separately approach. For pairwise correlated variables we can use any of the other four approaches. For more complex correlation structures (three or higher dimensional interaction) further extensions would be required. Such a situation will be left for future research.

The correlation structure of the three variables can be found in table 11. For the three dichotomized variables notation

Symbol	Loglinear model	df	$\chi^2$	$\chi^2$ Controls	$\chi^2$ Cases
(A,S,R)	$\log \mu = \lambda + \lambda^A + \lambda^S + \lambda^R$	4	376.6	145.4	212.5
(R, AS)	$\log \mu = \lambda + \lambda^A + \lambda^S + \lambda^R + \lambda^{AS}$	3	297.8	112.9	170.5
(A, SR)	$\log \mu = \lambda + \lambda^A + \lambda^S + \lambda^R + \lambda^{SR}$	3	273.0	104.6	161.1
(S, AR)	$\log \mu = \lambda + \lambda^A + \lambda^S + \lambda^R + \lambda^{AR}$	3	90.8*	37.1	46.5
(AR,AS)	$\log \mu = \lambda + \lambda^A + \lambda^S + \lambda^R + \lambda^{AS} + \lambda^{AR}$	2	45.9	18.4	18.6
(AR,SR)	$\log \mu = \lambda + \lambda^A + \lambda^S + \lambda^R + \lambda^{AR} + \lambda^{SR}$	2	18.0*	9.1	9.3
(AR,SR,AS)	$\log \mu = \lambda + \lambda^A + \lambda^S + \lambda^R + \lambda^{AR} + \lambda^{AS} + \lambda^{SR}$	2	0.015*	0.01	0.45
(ASR)	$\log \mu = \lambda + \lambda^A + \lambda^S + \lambda^R + \lambda^{AR} + \lambda^{AS} + \lambda^{SR} + \lambda^{ASR}$	0	0.0	0.0	0.0

**Table 12.** Labeling of loglinear models and results for the whole dataset and separately for cases and controls, star indicates the selected model in the respective step

and results of the eight log-linear models can be found in table 12. For the model with main effects only the number of expected cases deviate severely from the number of observed cases in the eight cells, clearly illustrating that correlations between the three binary variables exists. (Table 11). We propose a step-up procedure to derive a suitable log-linear model. We use  $\alpha=0.05$  as the significance level. In the first step the risk-asbestos interaction is added.  $X^2$  improves from 377 to 91. Adding the smoke-asbestos interaction leads to a significant improvement ( $X^2=46$ ). Adding the smoke-risk interaction results in a nearly perfect fit of the model ( $\chi^2 = 0.015$ ), so there is no three dimensional interaction in these data, but all three pairs of interactions are highly significant. As a strategy for model building we propose to consider the pair with the strongest interaction first (here asbestos-risk). We can use any of the strategies for two variables, but propose to adjust for the third variable (here smoke) as described for one variable ( $\tilde{x}$ ). The result of this step determines the adjustment model for investigation of the next pair AS and so on.

Using this strategy we get a final model(AR, SR, AS) including all twoway interactions. Now we can use the proposed strategies for the respective couples.

The model that fits best is the one including all two-way interactions. Now, we have to decide how to incorporate that in our model. One idea is to extend the dummy construction of the two spike variable case to 4 dummy variables.

## 5. Discussion/Further research

We proposed several new ideas for the analysis of variables with a spike at zero. In general, it is important to use the full information of the continuous variables that is available. The situation of a spike at zero occurs frequently in clinical research.

In the univariate case, results of the proposed FP-Spike method are easier to interpret than the usual FP results as could be seen in figure 2. In the case of two spike variable, we presented several ways of handling such situations. However, the proposed methodologies strongly depend on the distribution of zero and non-zero values. With the proposed methods for more than two spike variables, it is also possible to consider interaction of the spike variables.

It is now necessary to assess the properties of the different methods. A simple comparison using deviance and degrees of freedom is possible. However, as not all models are nested further comparisons are more complicated. The key components for comparison will be independence or dependence and the relative and absolute sample size in categories B and C. Our conclusions so far are that the separate analysis can be used if the two variables are more or less independent. As two separate dummies are used, we can not account for correlation other than the adjustment for the further variables.

In the case of correlated variables the combination of dummy variables might thus be more suitable. These three dummies differentiate between the four types of observations and thus take a relationship into consideration. If the amount of observations in categories B and C is relatively low, it might be reasonable to only include an indicator if both variables take value zero. In the case of different relationships for observations for which both variables are positive or only one of

them is positive, the sumbmodel approach is the most flexible one as it allows to use different functional relationships and different coefficients in the respective subgroups. To formalize these findings and to give more detailed guidance of when to use which strategy, further research is needed. It is also still a challenge to find suitable ways of graphical presentation for some of the proposed strategies. The “separately” method and “dummy 1” can be plotted as shown in the results part because the effects can be separated for the individual variables. For the other approaches three dimensional graphs will be needed.

## Acknowledgments

C. Jenkner was supported by the German Research Foundation (DFG), SA 2056/10-1. E. Lorenz was supported by the German Research Foundation (DFG), BE 2056/10-1. We thank Dr. Hermann Pohlabein, Prof. Wolfgang Ahrens and Prof. Karl-Heinz Jöckel for providing the data of the lung cancer study.

## References

1. Becher, H., Lorenz, E., Royston, P., and Sauerbrei, W. (2012). Analysing covariates with spike at zero: a modified FP procedure and conceptual issues. *Biometrical Journal*.
2. Hallstrom AP. (2010). A modified Wilcoxon test for non-negative distributions with a clump of zeros. *Statistics in Medicine*, 29:391-400.
3. Jöckel KH, Ahrens W, Jahn I, Pohlabein H, Bolm-Audorff U. (1998). Occupational risk factors for lung cancer: a case-control study in West Germany. *Int J Epidemiol*, 27(4):549-60.
4. Lachenbruch PA. (2001). Power and sample size requirements for two-part models. *Statistics in Medicine*, 20:1235-8.
5. 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.
6. Olson, M.K., Schafer, J.L. (2001). A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. *Journal of the American Statistical Association*, 96:730-745.
7. Ramroth, H., Dietz, A. and Becher, H. (2004). Interaction Effects and Population-attributable Risks for Smoking and Alcohol on Laryngeal Cancer and Its Subsites. *Methods in Medicine*, 43:499-504.
8. Robertson, C., Boyle, P., Hsieh C-C., Macfarlane GJ. 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.
9. Royston, P. and Sauerbrei, W. (2008). *Multivariable Model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. John Wiley & Sons.
10. 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(11):1219-1227.
11. Sauerbrei, W., Royston, P., Bojar, H., Schmoor, C., Schumacher, M., and for the German Breast Cancer Study Group (1999). Modelling the effects of standard prognostic factors in node-positive breast cancer. *British Journal of Cancer*, 79(11-12):1752-1760.
12. Schisterman EF, Reiser, B, Faraggi, D. (2006). ROC analysis for markers with mass at zero. *Statistics in Medicine*, 25:623-38.
13. Schumacher, M., Bastert, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, RLA. and Rauschecker, HF. (1994). Randomized 2x2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12(10):2086-2093.