# The use of fractional polynomials to model continuous risk variables in epidemiology

Patrick Royston,[a] Gareth Ambler[a] and Willi Sauerbrei[b]

**Background** The traditional method of analysing continuous or ordinal risk factors by categorization or linear models may be improved.

**Methods** We propose an approach based on transformation and fractional polynomials which yields simple regression models with interpretable curves. We suggest a way of presenting the results from such models which involves tabulating the risks estimated from the model at convenient values of the risk factor. We discuss how to incorporate several continuous risk and confounding variables within a single model. The approach is exemplified with data from the Whitehall I study of British Civil Servants. We discuss the approach in relation to categorization and non-parametric regression models.

**Results** We show that non-linear risk models fit the data better than linear models. We discuss the difficulties introduced by categorization and the advantages of the new approach.

**Conclusions** Our approach based on fractional polynomials should be considered as an important alternative to the traditional approaches for the analysis of continuous variables in epidemiological studies.

**Keywords** Continuous risk factors, model building, categorization, regression models, fractional polynomials, non-parametric models

**Accepted** 14 April 1999

A glance at any recent issue of an epidemiological journal will reveal that the traditional method of analysing continuous or ordinal risk factors in categories remains popular. A trend test in which ordinal scores (0, 1, 2, …) are assigned to the categories may also be performed.[1] The approach has some advantages.[2] The table of results is easy to understand and provides a natural way of communicating the findings, for example in terms of low, medium and high risk groups. Some robustness is achieved since no restrictive functional form (such as linearity) is imposed on the relation with the outcome. Attribution of extravagant risks to high values of the risk factor where data are sparse is usually avoided. Not least, the analysis is straightforward to perform.

However the final model, a risk step-function, is convenient but crude. It is unreasonable to postulate that risk suddenly increases as a category cutpoint is crossed. It is hard to interpret the random fluctuations in the estimates that inevitably occur across categories. Also, the results depend on the number and choice of cutpoints. From a statistical point of view, cutpoints should be chosen *a priori* to avoid data-driven inference and to make the analysis and results more comparable with similar studies. Since exploratory analysis is often needed, however, investigators may choose cutpoints which exhibit a convincing apparent effect on risk. The extreme is to seek so-called optimal cutpoints, but without the required *P*-value correction. The negative consequences are well-documented:[3,4] severe overestimation of effect sizes and *P*-values which are much too small. Furthermore, precision of estimates may be much reduced if a simple model which retains the continuous scale of the observations is not adopted.[5] As demonstrated in simulation studies, the loss of efficiency resulting from categorizing a quantitative exposure variable may be severe.[2,6] Problems with trend tests are discussed by Maclure and Greenland.[1] Finally, existing knowledge or strong prior belief based on scientific arguments may imply a monotonic (strictly increasing or decreasing) risk function. Incorporation of such knowledge is difficult using cutpoint-based approaches.

An alternative is a regression-based approach which avoids cutpoints. As with all types of regression modelling, the initial choice is the straight line $b_0 + b_1x$, where $x$ is the risk variable of interest. The scale on which the risk is linear in $x$ depends on the type of study. For a case-control study, for example, it is typically the log odds of being a case. Difficulties arise when the assumption of linearity is found to be untenable and a more appropriate model is sought. One possibility is to apply simple

[a] Imperial College School of Medicine, London, UK.

[b] Institute of Medical Biometry and Informatics, University of Freiburg, Germany.

Reprint requests to: Professor P Royston, Department of Medical Statistics & Evaluation, Imperial College School of Medicine, Hammersmith Hospital, Ducane Road, London W12 0NN, UK. E-mail: proyston@ic.ac.uk

transformations such as logarithm or square root to $x$. Alternatively, polynomial models may be used, but in practice they are often unsatisfactory (e.g.[7–9]).

There are two main issues here. (1) How do we *analyse* continuous risk variables—using categories or continuous functions? (2) How do we *report* the results—options include tabulating risk estimates in categories, presenting risk/exposure graphs, and reporting the mathematical formula for the risk function? If categories are used throughout, there is no conflict between the analysis and reporting of results. If we choose to analyse the risk variable on a continuous scale, there are many possibilities. A basic choice is between parametric and non-parametric models. Parametric models such as polynomials are easy to fit and the risk function may be written down concisely, but they may fit the data badly and give misleading inferences. On the other hand, non-parametric models (e.g. generalized additive models[7]) may fit the data well but be difficult to interpret due to fluctuations in the fitted curves. The risk function is usually impossible to write down concisely.

Seeking a single solution for these difficult issues may be inappropriate. Our aim is to present in the epidemiological context a simple but flexible parametric approach to modelling continuous risk factors called fractional polynomials.[8] We contrast it with the traditional analysis using categories and discuss some aspects of non-parametric modelling. We consider the presentation of models for continuous exposure variables. We give a detailed illustration of the modelling of a single continuous risk factor by applying our methods to data from the Whitehall I cohort study.[10] We also consider how to deal simultaneously with several continuous predictors, which may be risk factors or confounders. We use a multivariable procedure based on Royston and Altman[8] with refinements and extensions by Sauerbrei and Royston[11] to build these more complex regression models.

## Data

Whitehall I[10] is a prospective, cross-sectional cohort study of 18 403 male British Civil Servants employed in London. Its aim was to examine factors, particularly socioeconomic features, which influence death rates in a defined population. Identified causes of death included coronary heart disease (CHD), stroke and cancer. At entry to the study, each participant provided demographic details, filled in a health questionnaire and gave a blood sample for biochemical analysis. For present purposes we focus on the data from 10 years of follow-up. The sample consists of employees in departments other than the Diplomatic Service and the British Council, essentially the same as was analysed by Marmot *et al.*[10] We excluded four men with implausible systolic blood pressures of <85 mmHg and 106 with <10 years of follow-up, leaving 17 260 men of whom 1670 (9.7%) died. Such exclusion will not introduce bias if the censoring is non-informative. It allows us to use logistic regression analysis rather than Cox proportional hazards regression as the main analysis tool, which facilitates graphical exploration of the data. We consider cigarette smoking and systolic blood pressure as continuous risk factors for all-cause mortality, with age, plasma cholesterol concentration and Civil Service job grade as confounders. The response variable is binary, death within the 10-year follow-up period. The model estimates the log odds of dying as a function of the risk factor of interest.

**Table 1** Fractional polynomial models to predict all-cause mortality from the number of cigarettes smoked per day. The deviance difference compares the fit with that of a straight line ($p = 1$). The maximum deviance difference is distributed approximately as $\chi^2$ with 1 d.f. A positive value indicates a model which fits better than a straight line

| Power $p$ | Deviance | Deviance difference |
|---|---|---|
| –2 | 10 744.16 | 7.47 |
| –1 | 10 731.14 | 20.49 |
| –0.5 | 10 720.42 | 31.21 |
| 0 | 10 712.49 | 39.14 |
| 0.5 | 10 720.23 | 31.40 |
| 1 | 10 751.63 | 0.00 |
| 2 | 10 842.51 | –90.89 |
| 3 | 10 907.79 | –156.16 |

## A Simple Approach to Modelling by Using Fractional Polynomials

Suppose that we have an outcome variable which is related to a single continuous or ordered-categorical risk factor $x$. Initially we assume that the influence of $x$ is monotonic, which is true for many risk/exposure relationships. We consider more complex relationships and confounding variables later.

The natural starting point, the straight line model $b_0 + b_1x$ is often adequate, but other models must be investigated for possible improvements in fit. We look for non-linearity by fitting a first-order fractional polynomial to the data.[8] The best power transformation $x^p$ is found, with the power $p$ chosen from candidates –2, –1, –0.5, 0, 0.5, 1, 2, 3, where $x^0$ denotes log $x$. For example, for $p = –2$ the model is $b_0 + b_1/x^2$. The set includes the straight line (i.e. no transformation) $p = 1$, and the reciprocal, logarithmic, square root and square transformations. Even though the set is small, the powers offer considerable flexibility. Including more powers usually offers only a slight improvement in model fit. In particular, there is a danger with including large negative powers, such as –3, that individual extreme observations will influence the fit too much. Only if a test of $p = 1$ against the best-fitting alternative model with $p \neq 1$ is statistically significant do we prefer the non-linear function. The test is performed by comparing the difference in model deviances with a $\chi^2$ distribution on 1 degree of freedom. The resulting $P$-value is approximate and is justified by statistical arguments.[8] We assume the usual $P$-value of 5% for significance testing (see further comments later).

Table 1 shows the deviance for each power transformation of *cigs* + 1 as a predictor of the log odds of death in a logistic regression model for all-cause mortality in the Whitehall I study.

The addition of 1 to *cigs* avoids zero values which would prevent the use of logarithms and negative power transformations, the value 1 being the smallest increment in cigarette consumption observed in the data.[8] The best-fitting power is $p = 0$ (logarithmic transformation). The model has a deviance 39.14 lower than a straight line so the hypothesis that $p = 1$ is rejected ($P < 0.0001$). The results are illustrated in Figure 1.

The 'o' symbols in Figure 1(a) are observed 10-year mortality rates (i.e. number dying divided by number at risk) for intervals of cigarette consumption from 1 to 39 in steps of 3, then 40–49 and 50–60 cigarettes/day. The area of a symbol is proportional

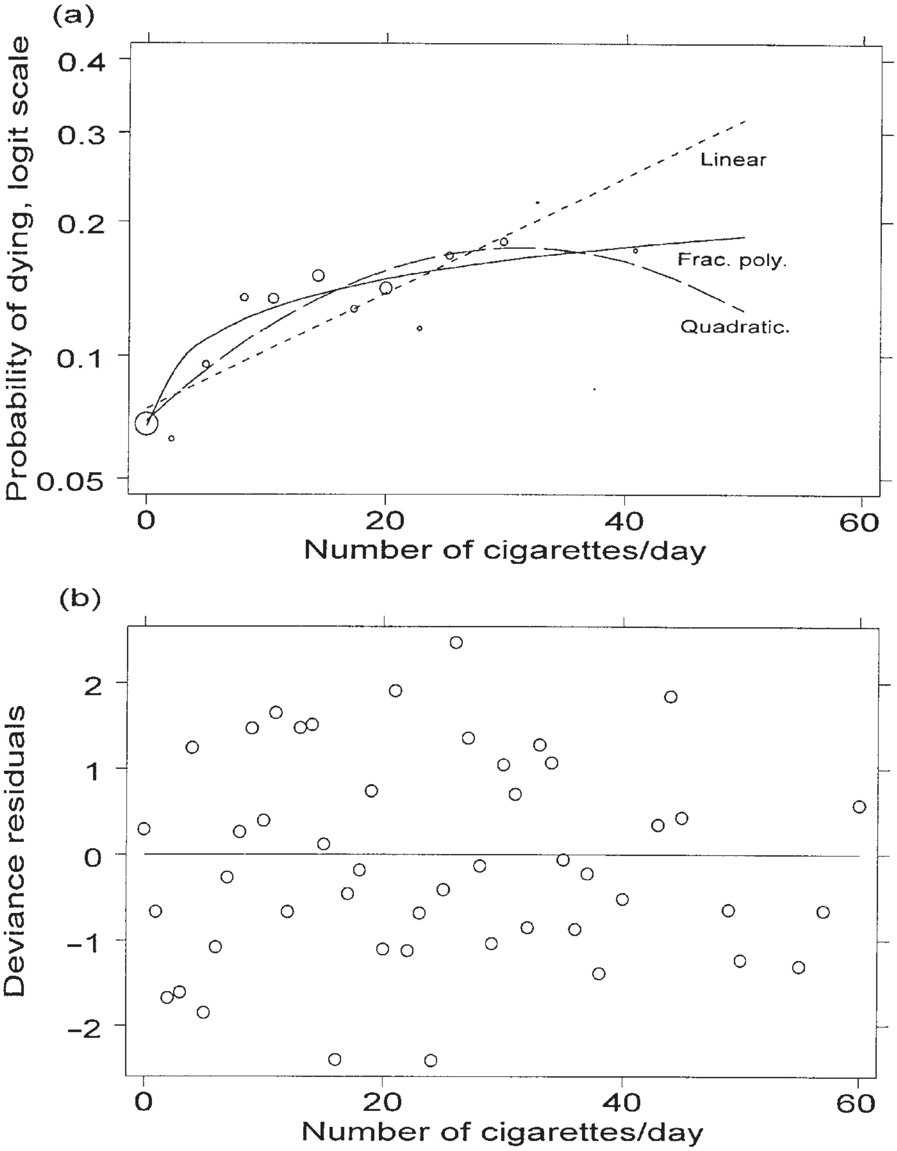

**Figure 1** All-cause mortality and cigarette consumption (*cigs*). (a) Observed rates and estimates from three logistic regression models. First-degree fractional polynomial (full line); linear (short dashes); quadratic (long dashes). The vertical axis is scaled as the log odds of dying. Size of symbols indicates precision of estimates. (b) Deviance residuals plotted against *cigs*

to the reciprocal of the variance of the estimate, so more reliable estimates have larger symbols. The largest symbol corresponds to non-smokers who comprise 58.5% of the sample. The fractional polynomial seems to capture the relationship quite well. The straight line model is a poor estimate of the true shape of the relationship as suggested, albeit noisily, by the observed rates. It underestimates the risk for low consumption and overestimates it for heavy smokers. The fitted curve from a quadratic polynomial, although a statistically significantly better fit than a straight line, suggests quite implausibly that the maximum risk occurs at about 30 cigarettes/day and that the risk at 60 cigarettes/day is about the same as at 3 cigarettes/day. Although the data contain little information on mortality rates of smokers of >30 cigarettes/day, we regard a model which predicts a lower death rate for very heavy smokers than for light ones as unacceptable.

The is a large literature on goodness-of-fit which we cannot go into here. For example, Hosmer *et al.*[12] recently compared several goodness-of-fit tests for logistic regression models. They showed that the Hosmer-Lemeshow $\hat{C}$ 'deciles of risk' statistic[13] was a satisfactory general test. Here, it gives a *P*-value of 0.14, giving no evidence of lack of fit. In addition, Figure 1b shows the deviance residuals for the fractional polynomial model plotted against number of cigarettes smoked. If the model is correct, the deviance residuals will be approximately normally distributed (provided the expected frequencies are not too small) and show no non-random trends or patterns. This seems to be the case.

**Table 2** Conventional cutpoint analysis of relationship between all-cause mortality and cigarette consumption

| | Number | | Odds ratio of death | |
|---|---|---|---|---|
| **Cigarettes/day** | At risk | Dying | Estimate | 95% CI |
| 0 (referent) | 10 103 | 690 | 1.00 | – |
| 1–10 | 2254 | 243 | 1.65 | 1.41–1.92 |
| 11–20 | 3448 | 494 | 2.28 | 2.02–2.58 |
| 21+ | 1455 | 243 | 2.74 | 2.34–3.20 |

**Table 3** Presentation of results of the first-degree fractional polynomial model for mortality as a function of cigarette consumption, preserving the usual information from an analysis based on categories. The odds ratio (OR) of death is calculated from the final model log OR = –2.62 + 0.293 * log(*cigs* + 1)

| **Cigarettes/day** | | **No.** | | **OR (model-based)** | |
|---|---|---|---|---|---|
| Range | Ref. point | At risk | Dying | Estimate | 95% CI |
| 0 (referent) | 0 | 10 103 | 690 | 1.00 | – |
| 1–10 | 5 | 2254 | 243 | 1.69 | 1.59–1.80 |
| 11–20 | 15 | 3448 | 494 | 2.25 | 2.04–2.49 |
| 21–30 | 25 | 1117 | 185 | 2.60 | 2.31–2.91 |
| 31–40 | 35 | 283 | 48 | 2.86 | 2.52–3.24 |
| 41–50 | 45 | 43 | 8 | 3.07 | 2.68–3.52 |
| 51–60 | 55 | 12 | 2 | 3.25 | 2.82–3.75 |

## Presentation of Results from Regression Models

The optimal presentation of results from an epidemiological study will often depend on the goals of the study. For example, if the aim is to test a simple binary hypothesis about a putative hazard, a *P*-value and confidence interval for the effect in question may be all that is required. Here we have concentrated on continuous predictors and, by implication, on the general problem of dose/response estimation. The suggestions we make in this section should be seen in that light.

Table 2 shows a conventional analysis of the smoking/mortality data in four categories.

Although the relationship between smoking and the risk of dying may be discerned from Table 2, much more information can be gleaned from the data, as Figure 1 shows. In Table 3, we suggest how to display results from a parametric regression analysis, here the first-degree fractional polynomial model.

The category-based estimates have been replaced with those from the regression model at relevant exposures. For example, the relative risk for 15 cigarettes/day is estimated as 2.25, increasing to 2.60 for 25 cigarettes/day. It is important to see how many individuals fall into each category and how many die, so we have retained the information in Table 3. The 95% CI are calculated from the appropriate standard errors as follows. Suppose $x_{\text{ref}}$ ref is the referent. We wish to calculate the SE of the difference between the fitted value at $x$ and that at $x_{\text{ref}}$ according to a first-degree fractional polynomial model $b_0 + b_1x^p$. The difference equals $b_1$ $(x^p - x^p_{\text{ref}})$ and its standard error is SE($b_1$) $(x^p - x^p_{\text{ref}})$. SE($b_1$) is obtained in standard fashion from the regression analysis. The 95% CI for the odds ratio are therefore exp $([x^p - x^p_{\text{ref}}]\ [b_1 \pm 1.96\text{SE}\ (b_1)])$. A CI for the relative risk instead of the odds ratio requires the inverse logit transformation instead of the inverse logarithm. The confidence interval neglects uncertainty in the estimation of the power $p$.

As with all regression models, the width of the CI at a given value of $x$ is strongly determined by the model and by the distance from the mean value of $x$ and much less by the number of observations in the region around $x$. With the traditional categorical analysis the number of the observations in a category determines the width more directly. For example, the confidence interval at 55 cigarettes/week is only 50% wider than at 25 cigarettes/week, whereas the latter is supported by about 100 times more observations. The confidence interval from the regression model may therefore be unrealistically narrow. This issue has only recently attracted detailed attention from statisticians and is incorporated in a concept called 'model uncertainty'. For references, see Discussion.

In addition, the equation for the regression model together with the SE of the estimated regression coefficients must be presented. The fitted model (with SE in parentheses) is

$$\log \text{OR} = -2.625\ (0.038) + 0.293\ (0.018)^*\log(cigs + 1).$$

A graph of the fitted function and its confidence limits, with or without observed rates as in Figure 1, should also be produced when feasible.

## More Complex Models: Second-degree Fractional Polynomials

Whether or not goodness-of-fit tests show that a first-degree fractional polynomial provides an unsatisfactory fit to the data, it is worth considering second-degree fractional polynomials, which offer considerably more flexibility. In particular, many functions with a single turning point (a minimum or a maximum), including some so-called J-shaped relationships, can be accommodated. For example, Figure 2 shows four curves with first power –2 and different second powers.

It demonstrates some of the different shapes that are possible. For further details see references 8 and 11. For further discussion and examples, see Greenland.[9] The models are of the form $b_0 + b_1x^p + b_2x^q$ or, for the mathematical limit $p = q$, $b_0 + b_1x^p + b_2x^p \log x$. As before, $p$ and $q$ are chosen from among –2, –1, –0.5, 0, 0.5, 1, 2, 3. The best fit among the 36 combinations of such powers is defined as that which maximizes the likelihood. For significance testing at the nominal 5% level against the best first-degree model, the difference in deviances is compared with the 95th percentile of $\chi^2$ with 2 d.f. A first-degree model is nested within a second-degree one so that the deviance of the latter is guaranteed to be smaller. The significance test, however, is approximate since the deviance difference is not exactly distributed as $\chi^2$.[8] For the smoking example, the best-fitting model has powers (–2, –1) in *cigs* + 1. A $\chi^2$-test on 2 d.f. based on the deviance difference from the log(*cigs* + 1) model has $P = 0.10$, so again the simpler model seems adequate.

Table 4 shows the model $\chi^2$ values for all possible first- and second-degree fractional polynomial models with systolic blood pressure as the risk factor for all-cause mortality.

The deviance for each model has been subtracted from that of the straight line model for each entry. Positive values in the table indicate an improvement in fit compared with a straight

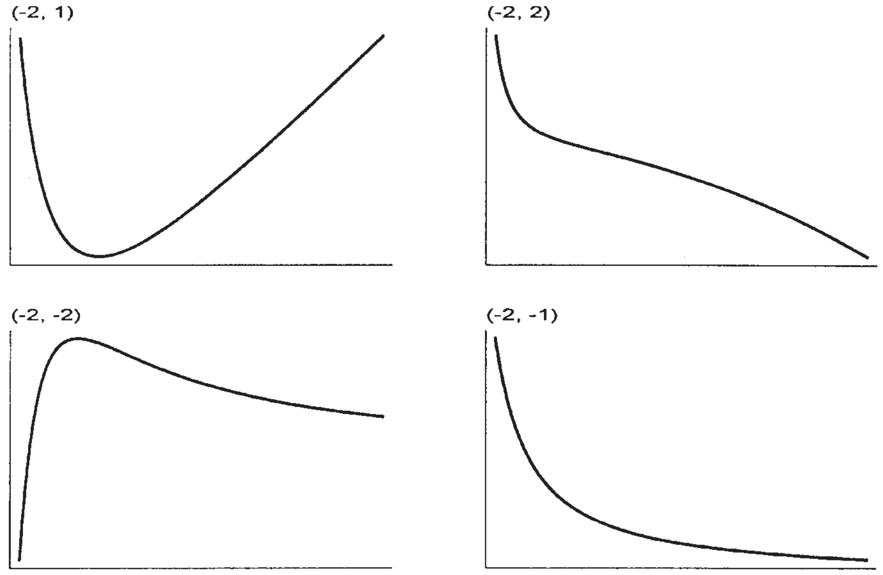

**Figure 2** Some examples of curve shapes possible with second-degree fractional polynomials

**Table 4** Fractional polynomial models for the relationship between systolic blood pressure and all-cause mortality. The deviance difference compares the fit with that of a straight line ($p$ = 1). The maximum deviance difference (underlined) is distributed approximately as $\chi^2$ with 1 d.f. or 3 d.f. for first- or second-degree models respectively. The best-fit first- and second-degree model powers are also underlined

| **Fractional polynomials for systolic blood pressure** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| First degree | | Second degree | | | | | | | | |
| Power | Deviance | Powers | | Deviance | Powers | | Deviance | Powers | | Deviance |
| *p* | difference | *p* | *q* | difference | *p* | *q* | difference | *p* | *q* | difference |
| –2 | –74.19 | –2 | –2 | 26.22 | –1 | 1 | 12.97 | 0 | 2 | 7.05 |
| –1 | –43.15 | –2 | –1 | 24.43 | –1 | 2 | 7.80 | 0 | 3 | 3.74 |
| –0.5 | –29.40 | –2 | –0.5 | 22.80 | –1 | 3 | 2.53 | 0.5 | 0.5 | 10.95 |
| 0 | –17.37 | –2 | 0 | 20.72 | –0.5 | –0.5 | 17.97 | 0.5 | 1 | 9.51 |
| 0.5 | –7.45 | –2 | 0.5 | 18.23 | –0.5 | 0 | 16.00 | 0.5 | 2 | 6.80 |
| 1 | 0.00 | –2 | 1 | 15.38 | –0.5 | 0.5 | 13.93 | 0.5 | 3 | 4.41 |
| 2 | 6.43 | –2 | 2 | 8.85 | –0.5 | 1 | 11.77 | 1 | 1 | 8.46 |
| 3 | 0.98 | –2 | 3 | 1.63 | –0.5 | 2 | 7.39 | 1 | 2 | 6.61 |
| | | –1 | –1 | 21.62 | –0.5 | 3 | 3.10 | 1 | 3 | 5.11 |
| | | –1 | –0.5 | 19.78 | 0 | 0 | 14.24 | 2 | 2 | 6.44 |
| | | –1 | 0 | 17.69 | 0 | 0.5 | 12.43 | 2 | 3 | 6.45 |
| | | –1 | 0.5 | 15.41 | 0 | 1 | 10.61 | 3 | 3 | 7.59 |

line. The best-fit fractional polynomials of degrees 1 and 2 (underlined) have powers 2 and (–2, –2) respectively. The best second-degree model fits significantly better than the best first-degree model (deviance difference 19.79, $P < 0.001$). Note that a quadratic, with powers (1, 2), has deviance only 6.61 lower than a straight line and is an inferior fit compared with the best second-degree model.

Figure 3a shows the five best second-degree fractional polynomial curves from Table 4.

All the models have comparable deviances (20.72 to 26.22 lower than a straight line) and a similar functional form, despite differences among the powers chosen. As with cigarette consumption (Figure 1a), the mortality rate has been estimated for small intervals of blood pressure, namely ⩽90, 91–95, 96–100, 101–199 by 3, 200–239 by 10 and ⩾240 mmHg. The area of the symbols indicates the precision as before. The dose-response relationship here is very strong, providing evidence of a J-shaped curve as discussed by Farnett *et al.*[14] The fractional polynomial analysis indicates an increased risk of death for systolic pressures around a nadir at about 110 mmHg, to some extent confirmed by the observed rates. Most individuals' blood pressures and hence the most reliable information lie between 100 and 180 mmHg, and considerable caution is required when making inferences outside this range. The greatest variation among the five fitted curves is in the regions with the sparsest data. Where there is plenty of data, they agree closely. An approach such as that of

Goetghebeur and Pocock[15] is needed to quantify the evidence for a J-shape specifically and sensitively. Figure 3b shows that the deviance residuals from the best-fitting fractional polynomial model have no apparent non-random trends or patterns. The Hosmer-Lemeshow $\hat{C}$ test gives a *P*-value of 0.68, showing no evidence of lack of fit.

In Figure 3c we compare the best first- and second-degree fractional polynomials, a straight line and a cruder categoric model with cutpoints somewhat arbitrarily placed at 110, 130, 160, 200 and 240 mmHg. Differences exist only for very low or high measurements. The straight line and the first degree fractional polynomial curve appear to underestimate the risk markedly

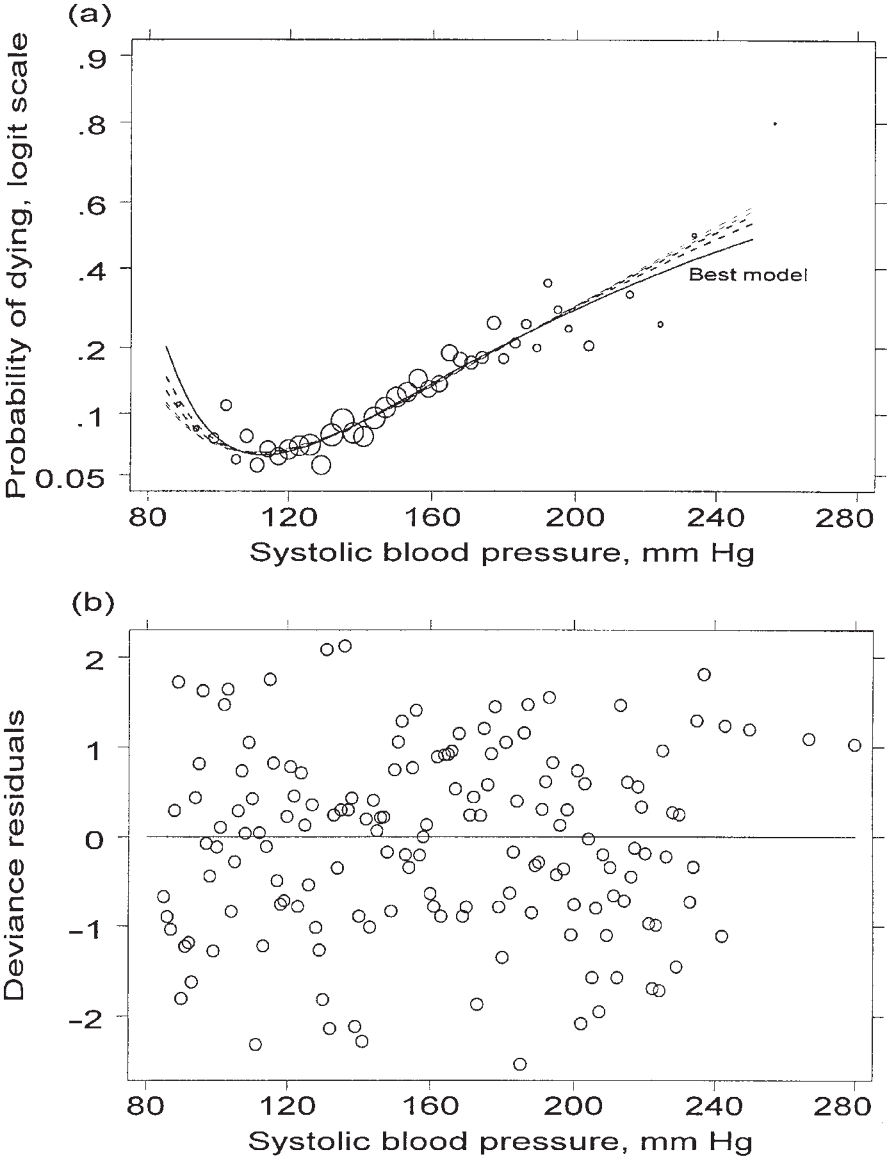

**Figure 3** All-cause mortality and systolic blood pressure: observed rates and estimates from logistic regression models based on second-degree fractional polynomials. The vertical axis in (a) and (c) is scaled as the log odds of dying

(a) The five best-fitting models are shown (see Table 4). Best fit (continuous line) has powers (–2, –2); others (broken lines) have powers (–2, –1), (–2, –0.5), (–2, 0) and (–1 –1).
Size of symbols indicates precision of estimates

(b) Deviance residuals plotted against systolic blood pressure

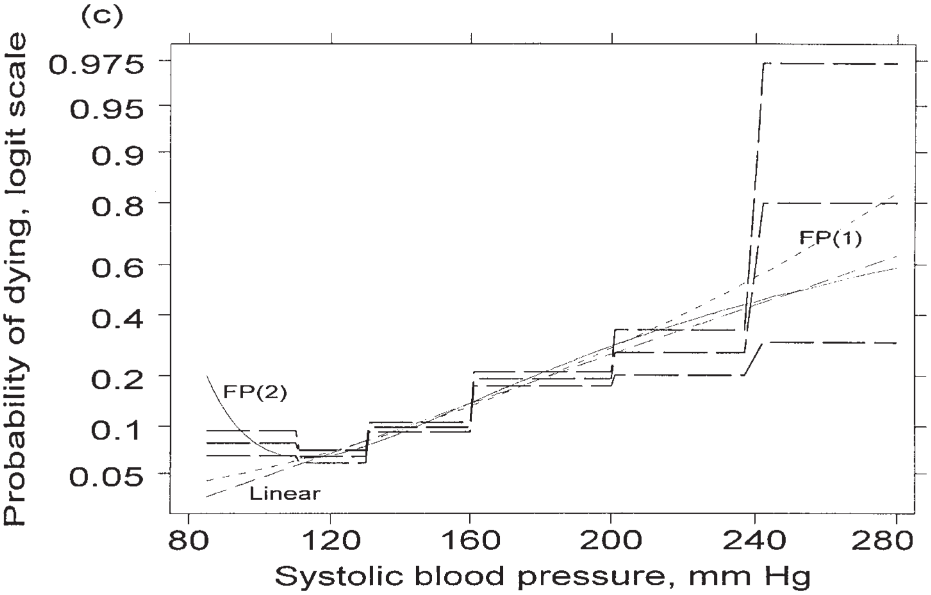

**Figure 3** (cont'd)

(c) As (a), but showing fits from other logistic regression models: best first- and second-degree fractional polynomials (short dashes and continuous line respectively), linear (medium dashes) and categoric (long dashes) with 95% CI

at very low blood pressures. The cutpoint model may be viewed as a rough approximation to the second-degree fractional polynomial model, though the risk estimates differ somewhat between 200 and 240 mmHg.

The best fractional polynomial has powers (–2, –2) (Table 4) and is of the form $b_0 + b_1x^{-2} + b_2x^{-2}\log x$. The standard errors (conditional on the powers) are calculated as follows. For convenience we write $w = x^{-2}$, $w_{\mathrm{ref}} = x_{\mathrm{ref}}^{-2}$, $z = x^{-2}\log x$, $z_{\mathrm{ref}} = x_{\mathrm{ref}}^{-2}\log x_{\mathrm{ref}}$, where $x_{\mathrm{ref}}$ is the referent. For the case of differing powers $p$ and $q$, we would have $w = x^p$, $w_{\mathrm{ref}} = x_{\mathrm{ref}}^p$, $z = x^q$, $z_{\mathrm{ref}} = x_{\mathrm{ref}}^q$. The difference between the fitted values at $x$ and $x_{\mathrm{ref}}$ is $b_1(w - w_{\mathrm{ref}}) + b_2(z - z_{\mathrm{ref}})$. The SE, which neglects uncertainty in $p$ and $q$, is

$$\sqrt{[(w - w_{\mathrm{ref}})^2\mathrm{var}(b_1) + (z - z_{\mathrm{ref}})^2\mathrm{var}(b_2) + 2(w - w_{\mathrm{ref}})(z - z_{\mathrm{ref}})\mathrm{cov}(b_1, \mathrm{b}_2)]}$$

where var and cov denote variance and covariance respectively. The variances are simply the squares of the SE and the convariance is obtained from the variance-covariance matrix of the estimated regression coefficients, $b_1$ and $b_2$. Table 5 shows the results of the second-degree fractional polynomial analysis for systolic blood pressures of 88, 95, 105 (referent), 115, 125, 135, 150, 170, 190, 220 and 250 mmHg.

The highest risk category only contains five individuals and the estimated odds ratio at 250 mmHg is very imprecise. As mentioned for the smoking example (earlier), the width is still underestimated. The strong risk gradient even among individuals all of whom would be classed as severely hypertensive is shown by comparing the estimated odds ratios of 8.24 and 4.54 at 220 and 190 mmHg respectively.

**Table 5** Presentation of results of the second-degree fractional polynomial model for mortality as a function of systolic blood pressure, preserving the usual information from an analysis based on categories. The odds ratio (OR) is calculated from the final model log OR = $2.92 - 5.43x^{-2} - 14.30 * x^{-2}\log x$ . The referent for the fractional polynomial model is 105 mmHg

| **Systolic blood pressure (mmHg)** | | **No. of men** | | **OR (model-based)** | |
|---|---|---|---|---|---|
| Range | Ref. point | At risk | Dying | Estimate | 95% CI |
| ≤90 | 88 | 27 | 3 | 2.47 | 1.75–3.49 |
| 91–100 | 95 | 283 | 22 | 1.42 | 1.21–1.67 |
| 101–110 | 105 | 1079 | 84 | 1.00 | – |
| 111–120 | 115 | 2668 | 164 | 0.94 | 0.86–1.03 |
| 121–130 | 125 | 3456 | 289 | 1.04 | 0.91–1.19 |
| 131–140 | 135 | 4197 | 470 | 1.25 | 1.07–1.46 |
| 141–160 | 150 | 2775 | 344 | 1.77 | 1.50–2.08 |
| 161–180 | 170 | 1437 | 252 | 2.87 | 2.42–3.41 |
| 181–200 | 190 | 438 | 108 | 4.54 | 3.78–5.46 |
| 201–240 | 220 | 154 | 41 | 8.24 | 6.60–10.28 |
| 241–280 | 250 | 5 | 4 | 15.42 | 11.64–20.43 |

## Handling Several Simultaneous Predictors

Usually in epidemiological studies, several predictors (risk factors and/or confounders) must be handled simultaneously. The univariate fractional polynomial approach may be applied to each risk factor in turn, adjusting for categorized versions of the other factors. However, this would give a separate model for each factor. It is preferable to build a single model which keeps all continuous factors continuous. The final model incorporates the 'best transformations' of the predictors, some of which may be no transformation ($p = 1$), i.e. a straight line. An approach to building such a model is described by Sauerbrei and Royston[11] who illustrated its use in two examples to obtain a diagnostic and a prognostic model where several continuous and categorical predictors were considered. Backward elimination is combined with an adaptive algorithm which selects the best fractional polynomial transformation for each continuous variable in turn. Depending on the *P*-value associated with the best transformation, one or more predictors may be excluded from the final model.

In epidemiological studies adjustment for confounders plays a central role, so the distinction between a continuous risk factor and a continuous confounder is important. If a model for the confounders was available *a priori* through subject matter knowledge, the situation for a single continuous risk factor would in principle be as just described, the only difference being choice of a univariate fractional polynomial adjusted for the given 'confounder model'. Often, however, decisions on the selection and modelling of confounders must be taken. Although other choices are possible, we propose to use *P*-values to select final models. So far we have used a nominal *P*-value of 5% for variable selection. However, the nominal *P*-value for a confounder should be larger than for a risk factor. This may ensure negligible 'residual confounding' and is similar in spirit to the conventional use of many categories, but avoids the disadvantages associated with the risk step-function. When many confounders are considered, the fractional polynomial approach would give a small final model, whereas use of, say, five categories for each confounder would result in many parameters.

With many continuous predictors as candidates in the model, a large nominal *P*-value will lead to substantial overfitting. For confounders this may not matter, but for risk factors it may be inappropriate, so a lower *P*-value of 5% or even 1% is preferable. The choice may depend on the number of predictors and whether hypothesis generation is an important aim of the study. The multivariable fractional polynomial procedure allows one to distinguish between risk factors and confounders.

We illustrate the effect of confounding using the multivariable fractional polynomial approach.[11] We use nominal *P*-values of 0.05 and 0.2 (Discussion) for the fractional polynomial terms for the risk factor and the continuous confounders respectively. Confounders considered in the smoking example are age at entry, systolic blood pressure, plasma cholesterol and Civil Service job grade (four groups). The same first-degree fractional polynomial power of 0 for *cigs* is obtained as before. The powers chosen for the confounders are –1 for age, (–2, –2) for systolic blood pressure and 3 for cholesterol, the latter transformation being just significant at $P = 0.2$. After adjusting for confounders, the regression coefficient for log(*cigs* + 1) changes from 0.293 to 0.270, indicating some positive bias in the unadjusted estimate. Regression diagnostics reveal no peculiarities in the final model. An analysis with the 'usual' categorization approach (we chose five categories) for each continuous confounder does not alter the fractional polynomial function chosen for blood pressure.

Systolic blood pressure, plasma cholesterol concentration and age may also be considered as risk factors. Using now a nominal *P*-value of 0.05 for fractional polynomial terms in the multivariable approach, the best transformations for each of these predictors (adjusting for Civil Service grade and smoking) are unchanged, except that cholesterol is now linear. The relationships between the log odds of mortality and systolic blood pressure and age show significant non-linearities ($P = 0.001$ and $P = 0.005$ respectively).

## Discussion

Disadvantages of the traditional approach of categorizing continuous variables according to cutpoints have been pointed out several times.[2,3,8,9] In a recent editorial, Weinberg[16] stated that approaches based on fractional polynomials or regression splines merit a greater role in epidemiology and should have a lasting influence on epidemiological practice. Our experience is that a one-term fractional polynomial (a power transformation of *x*) will work quite often. The approach is transparent, informative, flexible and more realistic than estimation using categories. Special software is unnecessary in that only eight regression models must be tried for each continuous variable. A major advantage is that the estimated response curve, $b_0 + b_1x^p$, is expressed by only three numbers: the power *p* and the regression coefficients $b_0$ and $b_{1.}$ As noted by Greenland[9] with $b_1 > 0$ values of $p < 1$ represent relationships which increase more rapidly than a straight line at low values of *x* and more slowly at high values, and vice versa for $p > 1$. It is not difficult to estimate[17] a general power of *x* unrestricted to the special set of values. Usually the extended model does not improve the fit much, but it may reassure one that a better power transformation has not been missed.

Regression diagnostics, such as detection of influential points and plots of residuals, may reveal lack of fit or other peculiarities of a first-degree fractional polynomial model, indication the need for a more complex function. A second-degree fractional polynomial is a candidate for such a model and allows the possibility of at most one turning point in the curve. The approximate test proposed by Royston and Altman[8] may be used to see whether a second-degree model is a significantly better fit than a first-degree one. Occasionally, a third-degree model with up to two turning points, a natural extension discussed by Royston and Altman,[8] may be needed. Our belief as that given the amount of 'noise' which typically obscures risk relationships in epidemiology, first- or second-degree fractional polynomials will provide sufficiently accurate approximations to an unknown reality for most purposes. We exclude time series, spatial modelling and other situations which require specialized mathematical models.

As with most types of data-dependent model-building, the results of hypothesis tests cannot be interpreted in the strict sense, parameter estimates may be biased and the 'naive' standard errors of fitted fractional polynomial curves are underestimated. Currently we cannot make a clear statement about the seriousness of the problem. However, it is relevant to many other situations in which a complex, data-dependent model-building process determines the proposed model. This well-known problem has received recent attention. As Chatfield[18] and other authors have pointed out, estimates conditional on the final model are almost always much too small because they take no account of so-called model uncertainty. Buckland *et al.*[19]

and Draper[20] have suggested some statistical approaches to the problem. The use of the bootstrap[21] may offer insight into its seriousness.

Adjustment for continuous confounders may be done by categorizing them and applying fractional polynomial analysis to the main risk factor. To avoid excessive bias caused by residual confounding, at least four categories are needed.[22,23] If several confounders are required in the final model, the categorization approach may result in too many terms. Including each confounder as a continuous variable offers some advantages. Breslow and Day[24] noted that 'effective control of confounding is often obtainable by inclusion of a few polynomial terms in the regression equation, thus obviating the need for stratification.' Brenner and Blettner[23] state that the inclusion even of a linear term often provides satisfactory control of confounding. Since the linear assumption may be far away from the true functional relationship, we believe that better control of confounding may be achieved by determining the best fractional polynomial transformation. We propose to use a larger nominal *P*-value to build a 'confounder model', because possible overfitting of the functional relationship for a confounder is acceptable. We chose to use 20% which is consistent with statements by Dales and Ury[25] and Mickey and Greenland[26] who favour higher *P*-values for significance testing of confounders. Other aspects such as maintaining comparability with previous studies may influence the choice between adjusting by fractional polynomial transformation or by categorization. For example, certain cutpoints may already be widely accepted in the literature. In our example use of categorized or continuous confounders did not effect the functional form for the risk factor, but this may not be the case in other datasets. The determination of functional form is particularly important in studies where a variable may have a dual role as a confounder or a risk factor.

For continuous risk factors, overfitting may be more serious and we prefer smaller nominal *P*-values of 5% or 1%. As a result a straight line relationship will often be postulated and should be accepted unless the data offer convincing evidence of non-linearity. In any case the final model should be considered in the light of prior evidence and scientific reasoning. A more detailed discussion of the importance of the nominal *P*-value in the model building process is given by Saurebrei.[27] A selection procedure to determine a fractional polynomial model in a multivariable context was proposed by Sauerbrei and Royston.[11] With this test-based approach, control for confounding factors and modelling of risk factors may be done simultaneously in a simple way.

The dataset (Whitehall I) we have used as a source of examples of modelling continuous variables is large and has many events. We may therefore expect high power to detect non-linear relationships, such as that between cigarette smoking and the log odds of all-cause mortality. Often several models have similar deviances and fit the data about equally well. The final model is chosen in a data-dependent fashion and may be a matter of chance. The stability of its functional form should be investigated, for example by bootstrap replication,[11] Most datasets in epidemiology are smaller than Whitehall I, with consequently less power to detect non-linearity. We would expect fractional polynomial analysis to select linear functions in many such cases. However, linear functions are to be preferred to categorized models in smaller datasets because precision will be improved by using simple models in such circumstances. Provided the power is adequate (for example, the dataset is not too small), the choice of a linear function following fractional polynomial analysis will reassure the analyst that no important non-linearity has been missed.

An alternative approach (e.g.[28]) for variables such as the number of cigarettes smoked is to include a binary variable for non-exposure in the model and to fit a continuous dose-response relationship among those exposed. The argument is that the unexposed may differ from the exposed in ways other than just the exposure level (unmeasured covariates). A drawback is that we will obtain a discontinuous dose-response relationship that may be poorly estimated for individuals with very low exposure where the data may be sparse. More detailed discussion is given by Greenland and Poole[29] who concluded that both approaches should be tried when subject matter knowledge suggests no clear preference, and Robertson *et al.*[30] who considered the implications of differential distributions of the exposure variable among cases and controls. Using a non-smoker variable with our smoking data, the best-fitting fractional polynomial power for *cigs* is $p = -0.5$ (reciprocal square root transformation). The deviance is 14.0 lower than a straight line model ($P < 0.001$) and 3.3 lower than the previous fractional polynomial model. In terms of goodness-of-fit there is little to choose between the models in this instance. The fitted risk function is log OR = –2.62 for non-smokers, $-1.32 - 1.82 \times cigs^{-0.5}$ for smokers.

As already noted, parametric models such as polynomials and fractional polynomials have disadvantages. Lack of flexibility may lead to a poor fit. The analyst may be misled both by the shape of the fitted function and by the apparent precision with which it is estimated. The estimate at a given point may be affected by observations a long way away, leading to local bias. The fit may be poor or inappropriate at extreme predictor values (end effects). So-called local regression models such as splines and kernel methods are intended to overcome such difficulties. Of these, the generalized additive model or GAM[7] is gaining popularity. The degree of smoothness of the fitted function, usually a cubic smoothing spline, is controlled by a user-selected parameter known as the equivalent degrees of freedom. Abrahamowicz *et al.*[31] use GAM to model the relation between CHD mortality and serum cholesterol concentration and find significant non-linear relationships. They criticise the use only of straight lines as subject to potentially serious bias (a view with which we agree), but surprisingly do not consider parametric alternatives to linearity. An alternative approach is regression splines, which are polynomial segments (linear, quadratic or cubic) joined smoothly at points known as knots whose position and number are user-determined. A detailed description of quadratic splines is given by Greenland.[9] A recent epidemiological application of splines is Carey *et al.*[32]

The strength of local models is their flexibility and the fact that their confidence intervals are generally wider and probably more realistic than those of parametric models. They better represent the amount of data near a given value of the risk factor. They may be used in exploratory data analysis and in helping one to select appropriate parametric models.[8] However, we do not think they are suitable as definitive models in epidemiology for the following reasons. The mathematical expressions for the curves are often very complex, so reporting of results must be by graphs or by extensive tabulation. The situation is unsatisfactory when similar studies are to be compared, and impossible if meta-analysis is intended. Data-dependence of the final model is more marked than for parametric models and the curves may be more difficult to interpret. Although procedures based on the Akaike Information Criterion or on generalized cross-validation may help, it is not necessarily clear what degree of smoothness to impose on the data. Too much smoothing leads to appreciable bias and too little to artefacts in the fitted curve. A typical example is a dose-response relationship which is expected *a priori* to be monotonic, but whose fitted curve contains artefacts. For an extreme example of overfitting resulting in implausible curves, see[33]. Monotonic dose-response relations may be modelled non-parametrically by using monotone splines.[34] It would be useful to perform simulation studies to gain better insight into the seriousness of the aspects just discussed, and to compare our fractional polynomial approach with spline models and other alternatives.

On balance, we feel that parametric models which retain the continuous scale of the observations have much to offer the epidemiologist. They are easy to handle in practice. Objectivity is greater because arbitrary categories, or far worse, data-driven categories, are avoided. There is complete freedom as to how to display results, for example by choosing relevant summary points for presenting risk estimates (Table 3), and/or by showing the entire fitted curve. With some additional calculation one can determine quantities such as the absolute risk gradient and its confidence interval, for example, the estimated difference in all-cause mortality between smokers of 40 cigarettes/day and non-smokers. Elswick *et al.*[35] offer a related method for interpreting the odds ratio from logistic regression after simple transformation of a covariate. Finally, a continuous parametric model predicts smoothly changing risk and, provided it fits the data adequately, is therefore more plausible than any cutpoint-based model.

## Acknowledgements

We thank Michael Marmot and Martin Shipley for making the data from the Whitehall I study available to us for analysis. We are grateful to Maria Blettner and a referee for helpful comments on an earlier version of the manuscript.

The research received financial support from project grant number 045512/Z/95/Z from The Wellcome Trust.

## References

[1] Maclure M, Greeland S. Tests for trend and dose response: misinterpretations and alternatives. *Am J Epidemiol* 1992;**135:**96–104.

[2] Zhao LP, Kolonel LN. Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. *Am J Epidemiol* 1992;**136:**464–74.

[3] Altman DG, Lausen B, Sauerbrei W, Schumacher M. The dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* 1994;**86:**829–35.

[4] Schulgen G, Lausen B, Olsen JH, Schumacher M. Outcome-oriented cutpoints in analysis of quantitative exposures. *Am J Epidemiol* 1994; **140:**172–84.

[5] Morgan TM, Elashoff RM. Effect of categorizing a continuous covariate on the comparison of survival time. *J Am Stat Assoc* 1986;**81:** 917–21.

[6] Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology* 1995;**6:**450–54.

[7] Hastie TJ, Tibshirani RJ. *Generalized Additive Models.* New York: Chapman and Hall, 1990.

[8] Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Appl Stat* 1994;**43:**429–67.

[9] Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology* 1995;**6:**356–65.

[10] Marmot MG, Shipley MJ, Rose G. Inequalities in death—specific explanations of a general pattern? *Lancet* 1984;**i:**1003–06.

[11] Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *J R Stat Soc, Ser A* 1999;**162:**71–94.

[12] Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit test for the logistic regression model. *Stat Med* 1997; **16:**965–80.

[13] Hosmer DW, Lemeshow S. *Applied Logistic Regression*. New York: Wiley, 1989.

[14] Farnett L, Mulrow CD, Linn WD, Lucey CR, Tuley MR. The J-curve phenomenon and the treatment of hypertension: is there a point below which pressure reduction is dangerous? *J Am Med Assoc* 1991; **265:**489–95.

[15] Goetghebeur EJT, Pocock SJ. Detection and estimation of J-shaped risk-response relationships. *J R Stat Soc, Ser A* 1995;**158:**107–21.

[16] Weinberg CR. How bad is categorization? (Editorial). *Epidemiology* 1995;**6:**345–47.

[17] Box GEP, Tidwell PW. Transformation of the independent variables. *Technometrics* 1962;**4:**531–50.

[18] Chatfield C. Model uncertainty, data mining and statistical inference (with discussion). *J R Stat Soc, Ser A* 1995;**158:**419–66.

[19] Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. *Biometrics* 1997;**53:**603–18.

[20] Draper D. Assessment and propagation of model uncertainty (with discussion). *J R Stat Soc, Ser B* 1995;**57:**45–97.

[21] Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993.

[22] Becher H. The concept of residual confounding in regression models and some applications. *Statistics in Medicine* 1992;**11:**1747–58.

[23] Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. *Epidemiology* 1997;**8:**429–34.

[24] Breslow NE, Day NE. *Statistical Methods in Cancer Research. Vol. I. The Analysis of Case-control Studies.* Lyon International Agency for Research on Cancer, 1980, pp.94–97.

[25] Dales LG, Ury HK. An improper use of statistical significance testing in studying covariables. *Int J Epidemiol* 1978;**4:**373–75.

[26] Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol* 1989;**129:**125–37.

[27] Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *Applied Statistics*. In press.

[28] Clayton D, Hills M. *Statistical Models in Epidemiology.* Oxford: Oxford University Press, 1993, pp.253–57.

[29] Greenland S, Poole C. Interpretation and analysis of differential exposure variability and zero-exposure categories for continuous exposures. *Epidemiology* 1995;**6:**326–28.

[30] Robertson C, Boyle P, Hsieh C-C, Macfarlane GJ, Maisonneuve P. Some statistical considerations in the analysis of case-control studies

when the exposure variables are continuous measurements. *Epidemiology* 1994;**5:**164–70.

[31] Abrahamowicz M, du Berger R, Grover SA. Flexible modeling of the effects of serum cholesterol on coronary heart disease mortality. *Am J Epidemiol* 1997;**145:**714–29.

[32] Carey VJ, Walters EE, Colditz GA *et al.* Body fat distribution and risk of non-insulin-dependent diabetes mellitus in women. The nurses study. *Am J Epidemiol* 1997;**145:**614–19.

[33] Zhao LP, Kristal AR, White E. Estimating relative risk functions in case-control studies using a non-parametric logistic regression. *Am J Epidemiol* 1996;**144:**598–609.

[34] Ramsay JO, Abrahamowicz M. Binomial regression with monotone splines: a psychometric application. *J Am Stat Assoc* 1989;**84:**906–15.

[35] Elswick RK Jr, Schwartz PF, Welsh JA. Interpretation of the odds ratio from logistic regression after a transformation of the covariate vector. *Stat Med* 1997;**16:**1695–703.