STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:127–141
Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2331

Dichotomizing continuous predictors in multiple regression: a bad idea

Patrick Royston[1,*,†], Douglas G. Altman[2] and Willi Sauerbrei[3]

[1]*MRC Clinical Trials Unit, 222 Euston Road, London NW1 2DA, U.K.*
[2]*Centre for Statistics in Medicine, University of Oxford, Wolfson College Annexe, Linton Road, Oxford OX2 6UD, U.K.*
[3]*Institute of Medical Biometry and Medical Informatics, University Hospital of Freiburg, Stefan-Meier-Str. 25, 79104 Freiburg, Germany*

SUMMARY

In medical research, continuous variables are often converted into categorical variables by grouping values into two or more categories. We consider in detail issues pertaining to creating just two groups, a common approach in clinical research. We argue that the simplicity achieved is gained at a cost; dichotomization may create rather than avoid problems, notably a considerable loss of power and residual confounding. In addition, the use of a data-derived 'optimal' cutpoint leads to serious bias. We illustrate the impact of dichotomization of continuous predictor variables using as a detailed case study a randomized trial in primary biliary cirrhosis. Dichotomization of continuous data is unnecessary for statistical analysis and in particular should not be applied to explanatory variables in regression models.

KEY WORDS: continuous covariates; dichotomization; categorization; regression; efficiency; clinical research

1. INTRODUCTION

'Why have researchers continued to ignore methodologists' advice not to dichotomize their measures?' [1]. Measurements of continuous variables are made in all branches of medicine, aiding in the diagnosis and treatment of patients. In medical research, such continuous variables are often converted into categorical variables by grouping values into two or more categories. It seems that the usual approach in clinical and psychological research is to dichotomize continuous variables, whereas in epidemiological studies it is customary to create several categories, often four or five, allowing investigation of a possible dose–response relation.

*Correspondence to: Patrick Royston, MRC Clinical Trials Unit, 222 Euston Road, London NW1 2DA, U.K.
†E-mail: patrick.royston@ctu.mrc.ac.uk

Contract/grant sponsor: Volkswagen-Stiftung

Received 4 January 2005
Accepted 11 May 2005

Although dichotomization is often done, its practice and implications have often been ignored in texts on medical statistics. In this paper we consider in detail the consequences of converting continuous data to two groups. We believe that dichotomization of continuous data is unnecessary for statistical analysis, and for most statisticians is not a natural way of analysing continuous data. It is done to make the analysis and interpretation of results simple. Furthermore, clinical decision-making often requires two classes, such as normal/abnormal, cancerous/benign, treat/do not treat, and so on. Although necessary and sensible in clinical settings, in a research context such simplicity is gained at a high cost, and may well create problems rather than solve them. As noted by Weinberg [2], 'alternative methods that make full use of the information at hand should indeed be preferred, where they make sense'. Such approaches include different types of splines, and fractional polynomials [3, 4].

In this paper, we discuss the impact of dichotomization of continuous predictor variables and present a detailed case study to illustrate the issues.

2. DICHOTOMIZING CONTINUOUS VARIABLES

Dichotomization is widespread in clinical studies [5], but the reasons for its popularity are largely a matter for speculation. There is to be a general need in clinical practice to label individuals as having or not having an attribute (such as 'hypertensive', 'obese', 'high' PSA), often preliminary to determining diagnostic or therapeutic procedures. Unfortunately, this attitude perhaps affects the way in which research is done. However, a similar liking for reducing data to two groups has been observed in other fields including psychology [6] and marketing [7].

As it is so common, many researchers may feel that this is in some sense the recommended approach. They may be inexperienced in analysing continuous variables, and may be unaware of the considerable range of suitable methods of analysis. Also, they may simply prefer more familiar and easier analyses. Additionally, among those who are more comfortable with regression there may be concerns about assuming a linear relation between the explanatory variable and the outcome. Such an automatic assumption may be wrong, and is neither necessary nor desirable.

2.1. Perceived advantages of dichotomizing

Various perceived advantages of dichotomizing continuous explanatory variables have been advanced, but they generally cannot be supported on statistical grounds [6]. The most common argument seems to be simplicity. Forcing all individuals into two groups is widely perceived to greatly simplify statistical analysis and lead to easy interpretation and presentation of results. A binary split leads to a comparison of groups of individuals with high or low values of the measurement, leading in the simplest case to a t test or χ^2 test and an estimate of the difference between the groups (with its confidence interval). In the context of a regression model with multiple explanatory variables the advantage is not as clear, although the regression coefficient (or odds ratio) for a binary variable may be felt easier to understand than that for a change in one unit of a continuous variable. Likewise the analysis of a single binary variable is much easier than that of a multi-category variable, which necessitates the creation of several dummy variables and for which there are several possible coding options and analysis strategies. Such

relative simplicity may be illusory, however. Even if there are good reasons to suppose that there is an underlying grouping, dichotomization at the median will not reveal it [6].

MacCallum *et al*. [6] considered various other weak or false arguments that may be put forward in support of dichotomization. For example, investigators may argue that because the analysis of a dichotomized variable is conservative, if a significant relation is found we can expect that the underlying relation is a strong one. They may also argue that dichotomization makes sense when the measurement is recorded imprecisely, and would provide a more reliable measure. This argument is incorrect—dichotomization will reduce the correlation with the (unknown) true values [6].

Not only are many of the perceived advantages illusory, dichotomization comes at a cost, as discussed in the next section.

2.2. Disadvantages of dichotomizing

The disadvantages of grouping a predictor have been considered by many authors, including References [6–11]. Grouping may be seen as introducing an extreme form of rounding, with an inevitable loss of information and power. When a normally distributed predictor is dichotomized at the median, the asymptotic efficiency relative to an ungrouped analysis is 65 per cent [12]. Dichotomizing is effectively equivalent to losing a third of the data, with a serious loss of power to detect real relationships. If the predictor is exponentially distributed, the loss associated with dichotomization at the median is even larger (efficiency is only 48 per cent [12]). Discarding a high proportion of the data is regrettable when many research studies are too small and hence underpowered. It seems likely that many who do this are unaware of the implications [6]. Furthermore, dichotomization may increase the probability of false positive results [11].

When the true risk increases (or decreases) monotonically with the level of the variable of interest, the apparent spread of risk will increase with the number of groups used. With just two groups one may seriously underestimate the extent of variation in risk; see Reference [13, p. 92] and Figure 5 below. Put differently, when individuals are divided into just two categories, considerable variability may be subsumed within each group. Faraggi and Simon [14] demonstrate a substantial loss of power when a cutpoint model is used to estimate what is in fact a continuous relationship between a covariate and risk. Furthermore, the cutpoint model is unrealistic, with individuals close to but on opposite sides of the cutpoint characterized as having very different rather than very similar outcome. We would expect the underlying relation with outcome to be smooth but not necessarily linear, and usually but not necessarily monotonic. Using two groups makes it impossible to detect any non-linearity in the relation between the variable and outcome.

Lastly, if regression is being used to adjust for the effect of a confounding variable, dichotomization of that variable will lead to residual confounding compared with adjustment for the underlying continuous variable [15–17]. Further issues arise when more than one explanatory variable is dichotomized. Both of these issues are discussed below.

2.3. Choice of cutpoint for dichotomization

Several approaches are possible for determining the cutpoint. For a few variables there are recognized cutpoints which are widely used (e.g. $>25\,\mathrm{kg/m^2}$ to define 'overweight' based on body mass index). For some variables, such as age, it is usual to take a 'round number',

an elusive concept which in this context usually means a multiple of five or 10. Another possibility is to use the upper limit of the reference interval in healthy individuals. Otherwise the cutpoint used in previous studies may be adopted.

In the absence of *a prior* cutpoint the most common approach is to take the sample median. However, using the sample median implies that different studies will take different cutpoints so that their results cannot easily be compared. For example, in prognostic studies in breast cancer, Altman *et al.* [9] found 19 different cutpoints used in the literature to dichotomize S-phase fraction. The median cutpoint was used in 10 studies. The range of the cutpoints was 2.6–12.5 per cent cells in S-phase, whereas the range of 5 'optimal' cutpoints (discussed in the next section) was 6.7–15.0 per cent. (Incidentally, we note that moving the cutpoint to a higher value leads to higher mean values of the variable in both groups.)

2.4. 'Optimal' cutpoints

The arbitrariness of the choice of cutpoint may lead to the idea of trying more than one value and choosing that which, in some sense, gives the most satisfactory result. Taken to extremes, this approach leads to trying every possible cutpoint and choosing the value which minimizes the *P*-value (or perhaps maximizes an estimate such as the odds ratio [18]). In practice, the search may be restricted to, say, the central 80 or 90 per cent of observations [9, 19]. The cutpoint giving the minimum *P*-value is often termed 'optimal', but it is optimal only in a narrow sense, and is unlikely to be optimal beyond the sample analysed [9].

Because of the multiple testing the overall type I error rate will be very high, being around 25–50 per cent rather than the nominal 5 per cent [9, 19–21]. Also, the cutpoint chosen will have a wide confidence interval and will not be clinically meaningful. Crucially, the difference in outcome between the two groups will be over-estimated, perhaps considerably, and the confidence interval will be too narrow. It is possible to correct the *P*-value for multiple testing [9, 19–21]. In addition, different types of shrinkage factor can be applied to correct for the bias and confidence intervals with the desired coverage can be derived by bootstrap resampling [22, 23]. However, it is not clear which shrinkage factor is best, and the approach is complex and little used so far.

Almost all studies using optimal cutpoints derive the cutpoint using univariate analysis and then use the resulting binary variable in multivariable analysis. Unless adjustment is made the results will be severely misleading [9]. Mazumdar *et al.* [24] extend the method of searching for a cutpoint for one specific predictor by adjusting in a multivariable model for other predictors known to be important. In particular, if a model reduction algorithm is used, the dichotomized predictor may lead to other, more influential variables being displaced. This data-dependent approach to analysis should be avoided. The strategy has been used frequently in oncological research.

2.5. Twofold cross-validation method

To evaluate the significance level and the hazard ratio (HR) associated with an 'optimal' cutpoint, Faraggi and Simon [14] suggested an approach based on twofold cross-validation. The main feature is that the cutpoint used to classify an observation is 'optimally' selected from a subset that excludes the observation. The algorithm may be summarized as follows. The data set is divided at random into two approximately equal subsets. The 'optimal' cutpoint is determined within each subset and is used to dichotomize observations in the other

subset. With this procedure, three usually different 'optimal' cutpoints are estimated. The approach defines a single dichotomization for all patients and is used for calculating the HR and *P*-value. The 'optimal' cutpoint from the original data is retained for later use.

Mazumdar *et al.* [24] stressed that if the underlying clinical setting is truly multivariable, the cutpoint search should incorporate other important variables. The same point was made earlier by Faraggi and Simon [14]. In epidemiological language, one should adjust for such variables in some way. However, Mazumdar *et al.* [24] give no suggestions or comments on how to determine the adjustment model. Mazumdar *et al.*'s proposed modification of the Faraggi–Simon method is to search for the three cutpoints as before, but adjusting for these other variables. Assuming in a simulation study that other correlated variables influence the outcome, they show that their modification improves power and reduces bias in the estimated HR and the cutpoint when the true model has a cutpoint.

We will exemplify some properties and difficulties of this recent approach in an example data set.

2.6. Impact of dichotomizing more than one explanatory variable

In practice, there is often more than one continuous explanatory variable in a regression analysis. The effect of dichotomization of two X variables will depend on the correlation coefficients between them and the response (Y), and cannot easily be predicted. Under some conditions, the inclusion of two dichotomized correlated variables can lead to a spurious relation between an X variable and Y [1, 6]. It is especially likely to occur when the partial correlation between one X variable and Y is close to zero. Also, this scenario can lead to spuriously significant interactions between X variables [1].

These findings suggest that regression models with two or more dichotomized continuous explanatory variables could be seriously misleading, both in respect of which variables are significant in the model, and perhaps also with respect to the overall predictive ability. If some of the cutpoints were selected using a data-dependent method, problems would worsen.

3. ILLUSTRATIVE ANALYSES

3.1. PBC data set

We use for illustration data from a randomized controlled trial in patients with primary biliary cirrhosis [25]. Between 1971 and 1977, 248 patients were randomized to receive either azathioprine or placebo with follow up until 1983. After removing 41 (17 per cent) of cases with missing values or no patient follow-up, data on 207 patients (105 deaths) in the PBC data set were available for analysis. We considered as candidate predictors the covariates age, albumin, bilirubin, central cholestasis, and cirrhosis. Age, albumin and bilirubin were continuous measurements and the other two were binary. The data were analysed by Cox regression.

3.2. Multivariable analysis of continuous and categorical predictors

To build a model involving a mix of continuous and binary predictors, we used the multivariable fractional polynomial (MFP) algorithm [4, 26]. In brief, the aim is to keep continuous

predictors continuous in the model. To do this successfully, potentially non-linear relationships must be accommodated. One approach is by using fractional polynomial functions. Univariate fractional polynomial models (see Reference [27] for a short introduction) were extended Reference [26] to allow simultaneous estimation of fractional polynomial functions of several continuous covariates. The user must prespecify the maximum complexity (degree) of fractional polynomial for each continuous predictor (usually 2), and the nominal significance level for testing variables and functions (often 0.05). The algorithm removes uninfluential predictors by applying backward elimination at the predefined significance level. It proceeds cyclically. The significance and functional form of each continuous predictor in turn are determined univariately, adjusting for all continuous and categorical predictors currently in the model. Convergence occurs when no further changes to selected variables and their fractional polynomial transformations take place. Convergence typically requires two to three cycles.

3.3. Multivariable analysis of the PBC data

For comparison with cutpoint approaches, we developed a multivariable prognostic model for the PBC data by applying the MFP procedure just outlined. We took a second-degree fractional polynomial as the most complex permitted function, and selected variables and functions of continuous variables by using a nominal P-value of 0.05. All models were adjusted for randomized treatment. The Cox model selected by the MFP procedure comprised cirrhosis, central cholestasis, age (untransformed), and log bilirubin. Albumin was not statistically significant when tested in the form of its best-fitting second degree fractional polynomial function, and was eliminated. At the final cycle of the algorithm, the test of a second degree fractional polynomial for bilirubin *versus* a linear function had $\chi^2 = 27.9$ on 3 degrees of freedom (d.f.) ($P < 0.001$), clear evidence that a straight line was not an adequate fit for this variable. The test of a first degree fractional polynomial *versus* the second degree function had $\chi^2 = 0.1$ on 2 d.f. ($P = 0.9$), showing that the simpler (logarithmic) function was acceptable.

3.4. 'Optimal' cutpoint for age

We first consider deriving an 'optimal' cutpoint for age. The model χ^2 was found for each candidate cutpoint in the central 90 per cent of observations ranging from 41 to 69 years, first univariately, then adjusting for the other factors (cirrhosis, central cholestasis, log bilirubin and treatment) from the MFP model. We define a binary variable representing dichotomization of X at X^* as 0 if $X \leqslant X^*$ and 1 otherwise.

The top left panel of Figure 1 shows that the 'optimal' cutpoint in a univariate analysis is at 45 years, with a χ^2 of about 10. After adjusting for three variables and treatment (Figure 1, top right panel) the 'optimal' cutpoint shifts to 65 years and has a χ^2 value of about 20. The χ^2 values are very unstable and are not conventionally significant for several cutpoints. Note that 52 years is nearly as good as the 'optimal' cutpoint in the adjusted analysis. The estimated HR fluctuates widely across cutpoints, particularly in the adjusted analysis (Figure 1, bottom right panel). When the cutpoint on age is large, the risk for patients classified as 'old' is much increased, but only a few patients fall into such a subgroup. For example for a cutpoint of 65 years, only 14.5 per cent of patients would fall into the 'old' group.

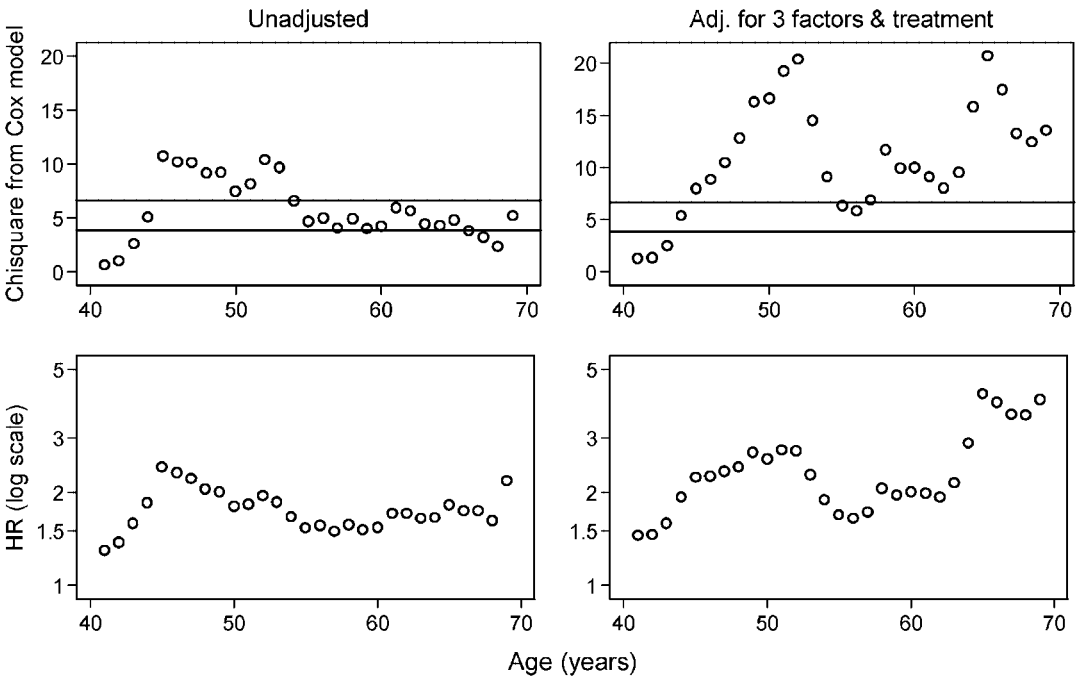

Figure 1. Derivation of the 'optimal' cutpoint for age, unadjusted (left panels) and adjusted for three other prognostic factors and treatment (right panels). Upper panels show the χ^2, the lower and upper horizontal lines denoting the critical values of the χ^2 distribution on 1 d.f. for testing significance at the 5 and 1 per cent levels, respectively. Lower panels show the HR for comparing 'old' with 'young' age by using dichotomization at the different ages shown.

3.5. Evaluation of the twofold cross-validation method

We applied Mazumdar *et al.*'s [24] extension of Faraggi and Simon's [14] twofold cross-validation procedure in 50 replicates to estimate the log HR and its 95 per cent confidence interval, adjusting for three prognostic variables and treatment. A different random number seed was used each time. The results are plotted ordered by the HR in Figure 2. The estimated HR has a large variance between replicates and a positively skew distribution, making it unclear how large the influence of age is. The median HR is 1.8, and this may be compared with the value of 4.2 for the 'optimal' cutpoint (see Figure 1, bottom right panel).

Figure 3 compares the cutpoints obtained in the two subsets across the 50 replications. About one half of the paired cutpoints are identical. Ignoring the arbitrary ordering between 'first' and 'second' cutpoints, most (41/50) of the paired cutpoints are very different, with one cutpoint, in the lower group, around 52 years and the other around 65 years. In only 1/50 replications was the same cutpoint (65 years) chosen in both halves. It is questionable whether estimates of HRs and *P*-values based on dichotomizations from such different cutpoints in the two halves have any merit.

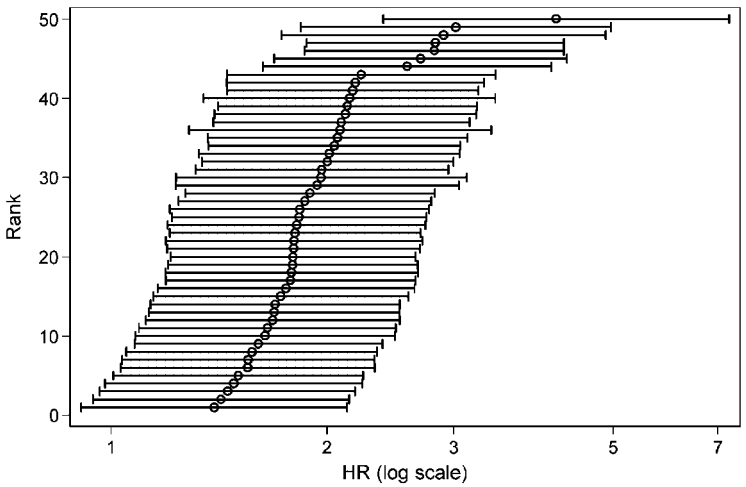

Figure 2. Ranked estimated HR for age in 50 replicate runs of the twofold cross-validation procedure, with 95 per cent confidence intervals.

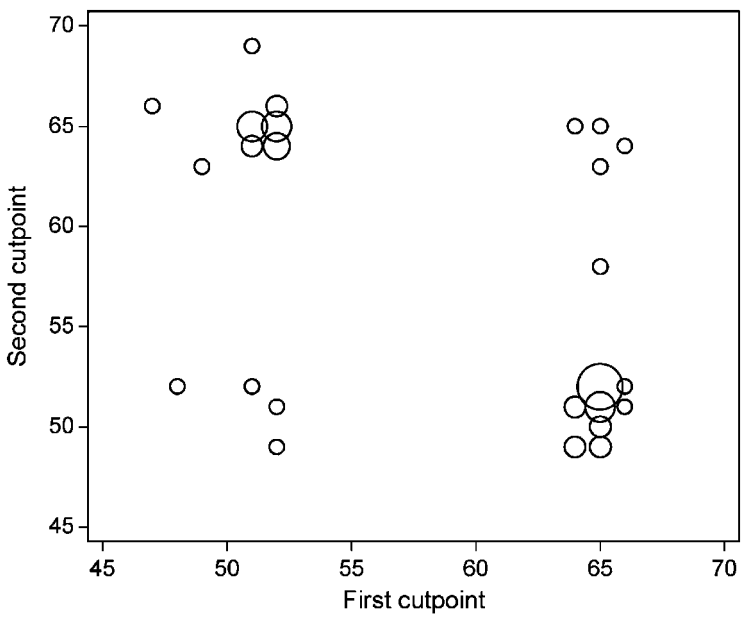

Figure 3. Pairs of 'optimal' cutpoints for age in random halves of the PBC data, adjusting for three prognostic variables and treatment. The area of each circle is proportional to the number of coincident observations plotted there.

3.6. Derivation of risk groups

There is no obvious reason to produce a prognostic model with one or more categorized continuous variables when the resulting linear predictor will still take many values. However, there is a real point in creating risk groups from such a model—not least, as an aid to making clinical decisions about therapy. Accordingly, we prefer first to derive a continuous risk score from a model in which all relevant covariates are kept continuous, and then to apply categorization at the final step. Patients are divided into several groups for clinical application by applying cutpoints to the risk score. Royston and Sauerbrei [28] suggest an approach to choosing a 'reasonable' number of risk groups loosely based on the idea that the HR between neighbouring groups should be statistically significantly different from 1. In the present example, it turns out that four groups is the maximum that may be entertained to maintain such separation of the hazard between neighbouring groups. Figure 4 shows Kaplan–Meier survival curves for four groups with equal numbers of events in each, derived from a risk score calculated from the MFP model. The patients separate nicely into low, low intermediate, high intermediate and high risk groups, the probability of surviving 3 years ranging from about 25 to 90 per cent.

3.7. Adjustment of a treatment effect

The PBC data originate from a randomized controlled trial of azathioprine *versus* placebo. In PBC, serum bilirubin concentration is a powerful predictor of survival time, and even slight imbalance in this factor between randomized groups could induce bias in the estimated treatment effect. There was indeed a small imbalance between the groups in log bilirubin of 0.23 SD units. Estimating the treatment effect within a multivariable model is the usual approach to adjusting for the imbalance. We assessed the robustness of the estimated treatment effect to alternative ways of modelling the prognostic effect of bilirubin. Table I shows the

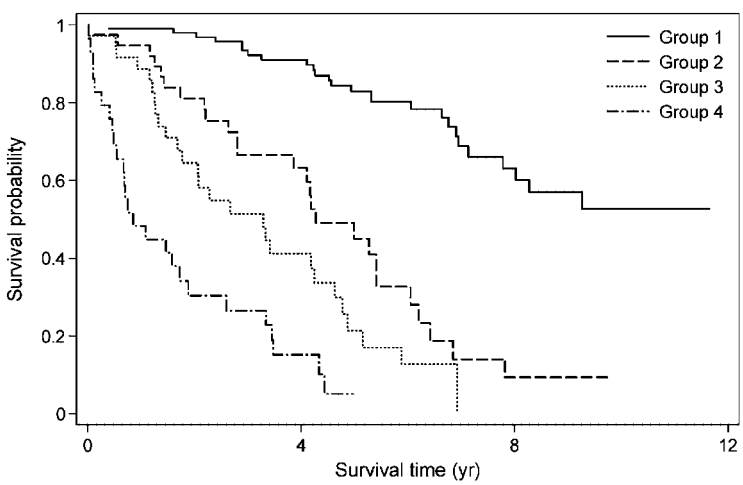

Figure 4. Prognostic groups for PBC data based on categorizing the risk score from the MFP model. Groups 1–4 contain 103, 39, 36 and 29 patients, respectively, with 26, 26, 27 and 26 events (deaths).

Table I. Estimated treatment effect in PBC data using different confounder models.

Model no.	Adjustment for bilirubin	HR for treatment	95 per cent CI	*P*-value for treatment
1	None	0.83	0.57, 1.22	0.348
2	Median cutpoint	0.76	0.51, 1.12	0.170
3	'Optimal' cutpoint	0.72	0.49, 1.06	0.094
4	Four groups	0.67	0.45, 0.99	0.046
5	Eight groups	0.58	0.38, 0.87	0.009
6	Linear	0.61	0.41, 0.91	0.015
7	Quadratic	0.58	0.39, 0.86	0.007
8	FP1	0.61	0.41, 0.90	0.014
9	FP2	0.60	0.40, 0.89	0.011
10	Spline with 4 e.d.f.	0.59	0.39, 0.87	0.008
11	Multivariable (MFP)	0.61	0.41, 0.90	0.014

The strong prognostic factor bilirubin is handled differently in each model, whereas identical adjustment for age, cirrhosis and central cholestasis is applied. See text for details.

results of the investigation with 11 different adjustment models. In model 1 no adjustment is applied. In models 2–10, adjustment is done for age (linear), cirrhosis and central cholestasis together with various transformations of bilirubin: median cutpoint (32 mmol/l), 'optimal' univariate cutpoint (45 mmol/l), four and eight equal-sized groups, linear, quadratic, FP1, FP2 and spline functions. In model 11, adjustment is by a multivariable model derived by the MFP approach. The notation 'FPm' denotes a fractional polynomial function of degree m, i.e. with m terms. The best FP1 and FP2 models for bilirubin were $\beta_1 \ln X$ and $\beta_1 X^{0.5} + \beta_2 X^{0.5} \ln X$, respectively. The spline model was a generalized additive model (GAM) [3] using a cubic smoothing spline for bilirubin with four equivalent d.f. Model 1, with no adjustment, gives the smallest estimated treatment effect. Models 2–4, with adjustment using categorization models, give HRs for comparing treatments rather closer to 1 than models 5–11, which have adjustment for bilirubin with many (8) groups or on a continuous scale. The treatment effects agree closely between models 5 and 11, even when the misspecified linear function is used for bilirubin. Both cutpoint adjustment models perform quite poorly. Even four groups (model 4) are not enough to abolish the effect of the imbalance in bilirubin. The large differences between the unadjusted model and the 'successfully' adjusted models 5–11 indicate that this study is a rather extreme example of a trial in which randomization did not completely balance the two treatment groups with respect to disease severity. The effect is analogous to residual confounding in epidemiological studies [15–17]. The unadjusted treatment effect has a *P*-value of 0.35, whereas the adjusted effects for models 5–11 have *P*-values of around 0.01.

3.8. Loss of information due to dichotomization

We compared the information content and ability to discriminate outcomes between three models for the PBC data. All models included the two continuous and two binary prognostic factors identified by the MFP procedure, and treatment. In model 1 both age and bilirubin were dichotomized at the median. In model 2 'optimal' cutpoints, determined univariately, replaced median cutpoints. Model 3 was an MFP model with all continuous variables retained as continuous (see Section 3.3). Table II shows the model χ^2 statistic (calculated from differences in twice the log partial likelihood), c-index, D measure of separation [28] and its associated

Table II. Quantifying the loss of information in two cutpoint models for the PBC data, compared with model 3 in which continuous variables were retained as continuous.

Measure	Model 1 Median cutpoint	Model 2 'Optimal' cutpoint	Model 3 Continuous
Model χ^2	94.6	99.2	136.8
c-index	0.778	0.774	0.814
D	1.91	2.02	2.55
R_D^2	0.465	0.494	0.608

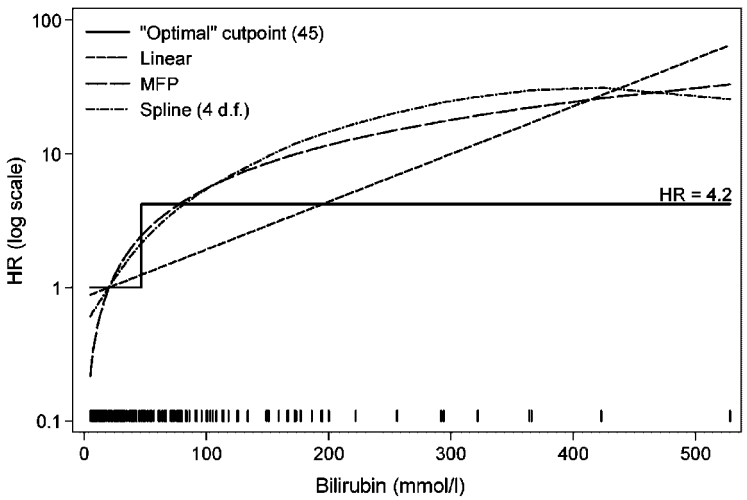

Figure 5. Functional form of the effect of bilirubin on the relative hazard according to the 'optimal' cutpoint of 45 mmol/l (determined univariately) and three continuous models (adjusted for three other prognostic factors and treatment). Functions are standardized such that the HR is 1 at the mean bilirubin (61.9 mmol/l). The short vertical lines on the horizontal axis indicate the values of bilirubin measurements.

R_D^2 measure of explained variation. The increase in model χ^2 for the continuous model 3 compared with both cutpoint models is >37, and the variance explained by model 3 is much higher. The c-index does not show the differences between models so clearly. The loss of information due to dichotomization is slightly less with 'optimal' cutpoints. The estimated treatment effect within models 1–3 is 0.74, 0.76 and 0.61, respectively.

3.9. Functional form

It is of interest to compare a cutpoint model for bilirubin with the continuous functions estimated by methods retaining continuous predictors as continuous. Figure 5 compares the univariate 'optimal' cutpoint (45 mmol/l) with linear and spline functions, and the function (here, log) selected by MFP. Clearly, the effect of bilirubin according to the cutpoint model is unrealistic. Also, the associated HR of 4.2 seems greatly to underestimate the range of

hazards seen with the continuous functions. Furthermore, none of the estimated continuous functions offers any justification for the data-driven choice of 45 mmol/l as a cutpoint. For most values of bilirubin above the mean, the straight line model probably underestimates the hazard. The MFP and spline models generally agree closely, but the spline model suggests a biologically implausible reduction in hazard for very high values of bilirubin.

4. DISCUSSION

It is well recognized in the methodological literature that dichotomization of continuous variables introduces major problems in the analysis and interpretation of models derived in a data-dependent fashion. Nevertheless, dichotomization of continuous variables is widespread in clinical research. Problems include loss of information, reduction in power, uncertainty in defining the cutpoint, arriving at a biologically implausible step function as the estimate of a dose–response function, and the impossibility of detecting a non-monotonic dose–response relation. Uncertainty in how to select a 'sensible' cutpoint to group a continuous variable into two classes has led researchers to use either the median or an 'optimal' cutpoint. The latter approach gives a highly inflated type 1 error probability, together with biased parameter estimates and variances that are too small [9, 11]. Although some remedies for these difficulties have been developed [9, 21–23], none of the authors of these papers actually recommends the use of 'optimal' cutpoints with their proposed corrections. In general, the situation seems hardly to have improved since the advice in 1993 of Maxwell and Delaney [1] to avoid dichotomization, quoted at the beginning of this paper.

Faraggi and Simon [14] put forward a method in which 'optimal' cutpoints are determined in three samples (overall and in two subsamples). The cutpoint determined in the overall sample is meant to be used in general applications. These authors showed by simulation that a realistic *P*-value and a nearly unbiased estimate of the HR are obtained by twofold cross-validation. 'Optimal' cutpoints are found separately in each subset, and the cutpoint from one subset is then used to classify patients in the other subset. Because the cutpoint used to dichotomize the patients in a given subset is determined independently of these patients, the *P*-values and HR estimates from dichotomized data in the overall sample are claimed to be approximately valid. Unfortunately, this ingenious idea for coping with problems in statistical analysis introduces fresh difficulties in interpretation and general use. In 41/50 replicate runs of the procedure in the PBC study, we obtained an 'optimal' cutpoint for age in one subset of about 52 years and a corresponding value of around 65 years in the other. A patient aged 60 would be classified as 'young' in one subset and 'old' in the other. The 'optimal' cutpoint in the overall sample is 65, but the χ^2 value for the cutpoint 52 is nearly as large (20.3 *versus* 20.7).

Recently, Mazumdar *et al.* [24] extended Faraggi and Simon's approach by finding the 'optimal' cutpoint for a variable of interest in a multivariable setting with adjustment for other factors. Assuming an underlying cutpoint model and a multivariate correlation structure between several continuous variables, they showed by simulation that the power was increased and estimates of the HR and the cutpoint were less biased when compared with the univariate approach. They also compared it with a split-sample approach. In the latter, an 'optimal' cutpoint is determined in a 50 per cent random subsample and used to classify patients in the complementary half. Compared with cross-validation, the split-sample method had lower

power and more biased estimates of the HR and cutpoint. The findings are as expected, since in the split-sample method, the sample size for the estimates and tests is reduced by half (see Reference [29] for further arguments against such an approach). Unfortunately, Mazumdar *et al*. [24] do not mention how to define the multivariable model. In an example in which identification of a group of patients at high risk of relapse from prostate cancer was required, they note that the search for a cutpoint for lactate dehydrogenase (LDH) gave different results in the univariate and multivariable settings. In the penultimate sentence of their paper, they stress that 'to incorporate the new markers in the decision-making process, categorization of these variables is essential'. We feel that this statement contradicts their own simulation results in which they demonstrate a substantial loss of power when a cutpoint model is used in cases where a smooth relationship exists between a continuous covariate and the outcome.

Instead of dichotomizing a continuous variable, we prefer to obtain a prognostic index by methodology which combines selection of variables with selection of functions for continuous variables [4, 26]. As stated in an editorial [2] in an epidemiological journal a decade ago, 'these elegant approaches [fractional polynomials and splines] merit a larger role in epidemiology.' Clinical researchers should in general avoid dichotomization at the model-building stage and adopt more powerful methods. In our analysis of the data from the PBC study, we compared several different approaches to creating a prognostic index. Explained variation was smallest for the model based on the median cutpoint, 6 per cent higher for the index derived with the 'optimal' cutpoint and 31 per cent higher for the MFP model. Although these figures will be slight over-estimates because no allowance has been made for data-dependent model-building, the advantage of using full information is obvious. We agree that medical decision-making often requires categorization of data, e.g. to define a high-risk group of patients for a clinical trial, as in Reference [24] example. However, categorization should be applied to the prognostic index, not to the original prognostic variables. Not to do so risks a loss of discrimination through inefficient use of the full information available with a continuous prognostic index.

By estimating the treatment effect in the PBC data within different adjustment models, we showed that the method used to adjust for an unbalanced, strongly prognostic variable can influence the result. Adjustment for bilirubin dichotomized at the median cutpoint does not fully correct for imbalance. Epidemiologists would state that there was residual confounding. Use of more groups or the full information from the continuous variable further reduces residual confounding and results in larger estimates of the treatment effect. This finding agrees with simulation studies in the epidemiological literature on the ability to reduce residual confounding by categorized variables [15, 17, 30]. Brenner and Blettner [17] state that 'inclusion of the confounder as a single linear term often provides satisfactory control for confounding even in situations in which the model assumptions are clearly violated. In contrast, categorization of the confounder may often lead to serious residual confounding if the number of categories is small.' The most extreme situation of two categories seems to have been abandoned in epidemiological studies.

For model building with continuous data, software is available for methods such as multivariable fractional polynomials [31, 32] and GAMs (e.g. in S-plus and R). Royston and Sauerbrei [33] demonstrated in a detailed resampling study that over-fitting and the resulting instability need not be a serious issue with MFP models. Even the use of conventional polynomials would in general improve on dichotomization. These methods should replace analyses using dichotomized continuous variables. Preference for one particular approach should be

guided by parsimony, an important criterion for selecting the simplest adequate descriptor of a functional form [2].

ACKNOWLEDGEMENTS

We are grateful for the support provided by the Volkswagen-Stiftung (RiP-program) at the Mathematisches Forschungsinstitut, Oberwolfach, Germany. Some of the work was carried out at Oberwolfach during a visit in 2004.

REFERENCES

1. Maxwell SE, Delaney HD. Bivariate median-splits and spurious statistical significance. *Psychological Bulletin* 1993; **113**:181–190.
2. Weinberg CR. How bad is categorization? *Epidemiology* 1995; **6**:345–347.
3. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman & Hall: New York, 1990.
4. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics* 1994; **43**(3):429–467.
5. Del Priore G, Zandieh P, Lee MJ. Treatment of continuous data as categoric variables in obstetrics and gynecology. *Obstetrics and Gynecology* 1997; **89**:351–354.
6. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. *Psychological Methods* 2002; **7**:19–40.
7. Irwin JR, McClelland GH. Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research* 2003; **40**:366–371.
8. Cohen J. The cost of dichotomization. *Applied Psychological Measurement* 1983; **7**:249–253.
9. Altman DG, Lausen B, Sauerbrei W, Schumacher M. The dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 1994; **86**:829–835.
10. Harrell Jr FE. Problems caused by categorizing continuous variables. http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/CatContinuous, 2004. Accessed on 6.9.2004.
11. Austin PC, Brunner LJ. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine* 2004; **23**:1159–1178.
12. Lagakos SW. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statistics in Medicine* 1988; **7**:257–274.
13. Breslow NE, Day NE. *Statistical Methods in Cancer Research*, vol. 1. IARC Scientific Publications: Lyon, 1980.
14. Faraggi D, Simon R. A simulation study of cross-validation for selecting an optimal cutpoint in univariable survival analysis. *Statistics in Medicine* 1996; **15**:2203–2213.
15. Becher H. The concept of residual confounding in regression models and some applications. *Statistics in Medicine* 1992; **11**:1747–1758.
16. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968; **24**:295–313.
17. Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. *Epidemiology* 1997; **8**:429–434.
18. Wartenberg D, Northridge M. Defining exposure in case-control studies: a new approach. *American Journal of Epidemiology* 1991; **133**:1058–1071.
19. Miller R, Siegmund D. Maximally selected chi-square statistics. *Biometrics* 1982; **38**:1011–1016.
20. Lausen B, Schumacher M. Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Computational Statistics and Data Analysis* 1996; **21**:307–326.
21. Hilsenbeck SG, Clark GM. Practical *P*-value adjustment for optimally selected cutpoints. *Statistics in Medicine* 1996; **15**:103–112.
22. Schumacher M, Holländer N, Sauerbrei W. Resampling and cross-validation techniques: a tool to reduce bias caused by model-building? *Statistics in Medicine* 1997; **16**:2813–2827.
23. Holländer N, Sauerbrei W, Schumacher M. Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint. *Statistics in Medicine* 2004; **23**:1701–1713.
24. Mazumdar M, Smith A, Bacik J. Methods for categorizing a prognostic variable in a multivariable setting. *Statistics in Medicine* 2003; **22**:559–571.
25. Christensen E, Neuberger J, Crowe J, Altman DG, Popper H, Portmann B, Doniach D, Ranek L. Tygstrup N, Williams R. Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis: final results of an international trial. *Gastroenterology* 1985; **89**:1084–1091.

26. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society Series A* 1999; **162**:71–94 (Corrigendum: *Journal of the Royal Statistical Society*, *Series A* 2002; **165**:399–400).
27. Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine* 2004; **23**:2509–2525.
28. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Statistics in Medicine* 2004; **23**:723–748.
29. Hirsch RP. Validation samples. *Biometrics* 1991; **47**:1193–1194.
30. Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology* 1995; **6**:450–454.
31. StataCorp. *Stata Reference Manual*, *Version 8*. Stata Press: College Station, TX, 2003.
32. Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. Multivariable regression model building by using fractional polynomials: description of SAS, Stata and R programs. *Computational Statistics and Data Analysis*, 2005, submitted.
33. Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Statistics in Medicine* 2003; **22**:639–659.

