Controlling for Continuous Confounders in Epidemiologic Research

Hermann Brenner[1] *and Maria Blettner*[2]

Multiple regression models are commonly used to control for confounding in epidemiologic research. Parametric regression models, such as multiple logistic regression, are powerful tools to control for multiple covariates provided that the covariate-risk associations are correctly specified. Residual confounding may result, however, from inappropriate specification of the confounder-risk association. In this paper, we illustrate the order of magnitude of residual confounding that may occur with traditional approaches to control for continuous confounders in multiple logistic regression, such as inclusion of a single linear term or categorization of the confounder, under a variety of assumptions on the confounder-risk association. We show 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. Alternative strategies to control for confounding, such as polynomial regression or linear spline regression, are a useful supplement to the more traditional approaches. (Epidemiology 1997;8:429–434)

Keywords: biostatistics, confounding, epidemiologic methods, logistic regression.

Regression models commonly employed in the analysis of epidemiologic studies assume a specific mathematical relation between levels of covariates and disease occurrence.[1,2] For example, a continuous logistic regression model assumes a linear relation between levels of continuous covariates and the log odds of disease. This assumption may often be violated, inasmuch as the true dose-response is usually unknown. Categorization of continuous covariates has been proposed to allow for more flexible modeling of the shape of covariate-risk association in such situations.[3] Inclusion of simple linear terms and categorization are currently the most widely used strategies to deal with continuous covariates in multiple regression models.

The literature on the pros and cons of categorization of continuous covariates in epidemiologic analysis is extensive. Most of this discussion has focused on issues of power, validity of estimation of the shape of the covariate-disease association, influence of outliers, illustrativeness of data presentation,[3–6] or selection of cut-points of categorical analysis.[7–11] Many of the disadvantages of categorical analysis are related to the fact that it does not make efficient use of within-category information.

During the past years, alternative strategies, such as polynomial regression or spline regression,[12,13] for dealing with continuous covariates in multiple regression models have received increased attention. These methods may partly overcome the limitations of the more commonly employed traditional approaches.

Most of the discussion on how to deal with continuous covariates in epidemiologic analyses has focused on estimating effects of the exposure variable of primary interest. In contrast, the question of how to deal with continuous confounders has received much less attention. This issue is of considerable importance, however, given that misspecification of the confounder-risk association may hinder adequate control for confounding. Cochran[14] assessed control for confounding through "subclassification" by categories of a continuous confounder when comparing the means of some continuous trait in two study groups. He showed that, under certain distributional assumptions and for monotonic relations of the confounder and the trait under investigation, the percentages of bias removed by subclassification are roughly 64%, 79%, 86%, 90%, and 92% for classifying the confounder into 2, 3, 4, 5, and 6 categories, respectively. The complementary proportion of bias not removed by imperfect control for confounding is commonly called "residual confounding." More recently, Becher[15] reported similar levels of residual confounding resulting from confounder classification in simulation studies for various types of regression models. Building on the work by Cochran, Rothman[3] argued that most of

From the Departments of Epidemiology, [1]University of Ulm, Ulm, Germany, and [2]German Cancer Research Center, Heidelberg, Germany.

Address correspondence to: Hermann Brenner, Department of Epidemiology, University of Ulm, Albert-Einstein-Allee 43, D-89081 Ulm, Germany.

Submitted April 29, 1996; final version accepted October 31, 1996.

Editors' note: See related commentaries on pages 453 and 457 of this issue.

the confounding from a given factor can be removed by a stratified analysis based on only two categories of a continuous variable and that it is rarely necessary to have more than about five categories. Nevertheless, more complete control for confounding may be necessary to prevent misleading results in the case of a strong confounder, particularly if the exposure has no or only a weak effect. Common examples are epidemiologic studies on the risk of respiratory disease due to environmental factors in the presence of confounding by cigarette smoking.[16] In such situations, more flexible regression models that warrant more complete control of confounding could be particularly useful.[13,17]

Here, we assess the efficacy of various strategies to control for continuous confounders, using a continuous logistic regression model under a variety of confounder-risk associations. The strategies we assess include using a single linear term or a quadratic and a cubic term along with a linear term of the confounder, and categorization by a varying number of categories, with or without inclusion of linear splines.

Methods

We assumed that an investigator carries out a cohort study to assess the association of some continuous exposure variable X_1 with the risk of some dichotomous disease status Y (with values 1 and 0 for presence and absence of disease, respectively), while controlling for some continuous confounder X_2 by multiple logistic regression. To focus on the efficacy of control for confounding, we assumed that X_1 does not influence risk of disease, which means that, in the absence of other sources of bias, any apparent effect of X_1 is due either to confounding by X_2 or to random error. A necessary condition for X_2 being a confounder is that X_2 is related to both X_1 and the risk of disease (apart from its association with X_1).[3] In the following, we illustrate the performance of different strategies to control for confounding by X_2 under a variety of conditions.

Distribution and Association of Covariates

Many continuous covariates studied in epidemiologic research approximately follow a normal distribution in the population. For simplicity, we assumed that X_1 and X_2 follow a bivariate normal distribution with mean 0 and variance 1 for each component. We assessed the following levels of correlation between X_1 and X_2: 0.5, 0.7, and 0.9.

Covariate-Risk Associations

We assessed a variety of scenarios of the confounder-risk association, as outlined in Table 1. In all scenarios, we assumed disease risk to be affected by the confounder X_2 but not by X_1.

In scenario A, the risk at the mean level of X_2 ($X_2 = 0$) is 0.10, as reflected in the intercept of $\ln(1/9) = \ln(0.1/0.9)$, and the logit of disease risk follows a linear function of X_2. In this scenario, the logistic model is correctly specified with inclusion of a single linear term for X_2.

TABLE 1. Association between Confounder X_2 and Disease Risk R Assumed in the Scenarios Used for Numerical Illustration; P_i = ith Percentile of the Distribution of X_2

Scenario	Confounder-Risk Association
A	$\text{logit}(R\|X_2) = \ln(1/9) + \ln(2) \times X_2$
B	$\text{logit}(R\|X_2) = \ln(1/9) + \ln(2) \times X_2 + \ln(4.5) \times X_2^2$
C	$R = 0.05$ if $X_2 < P_{50}$, $R = 0.15$ if $X_2 \geq P_{50}$
D	$R = 0.05$ if $X_2 < P_{67}$, $R = 0.20$ if $X_2 \geq P_{67}$
E	$R = 0.05$ if $X_2 < P_{75}$, $R = 0.25$ if $X_2 \geq P_{75}$
F	$R = 0.05$ if $X_2 < P_{80}$, $R = 0.30$ if $X_2 \geq P_{80}$

In scenario B, the baseline risk at the mean level of the confounder ($X_2 = 0$) is the same as in scenario A (0.10), but the logit of the disease risk follows a linear combination of X_2 and the square of X_2. In this scenario, the logistic model is correctly specified with inclusion of both a linear and a quadratic term of X_2.

The relation between X_2 and disease risk in scenarios A and B is depicted in Figure 1 for levels of X_2 between -3 and $+3$ (this range encompasses 99.7% of the distribution of X_2). Whereas the association is monotonic in scenario A, a J-shaped relation emerges in scenario B. Such J-shaped relations have repeatedly been observed in epidemiology. Well-known examples include the associations of body mass index or alcohol consumption with all-cause mortality.[18,19]

Scenarios C, D, E, and F reflect situations with a threshold effect of X_2. Such threshold effects have repeatedly been described for noncarcinogenic agents, particularly in the field of occupational epidemiology.[20]

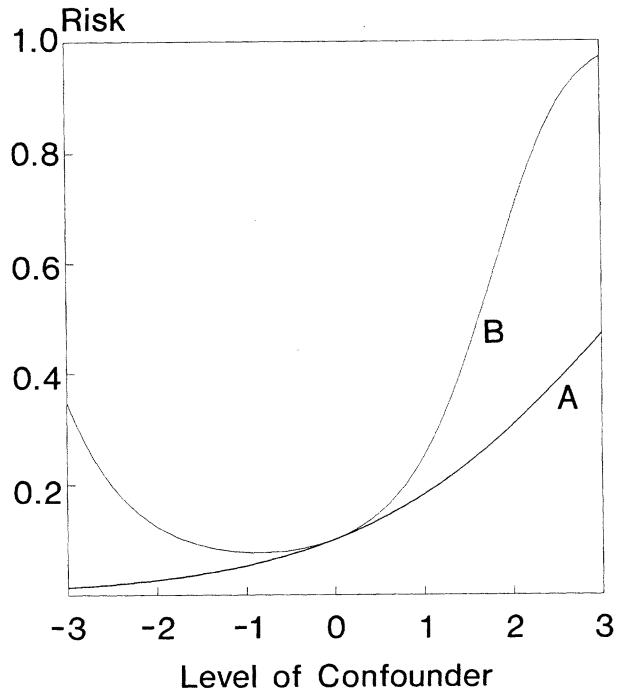

FIGURE 1. Confounder-risk relations assumed in scenarios A and B.

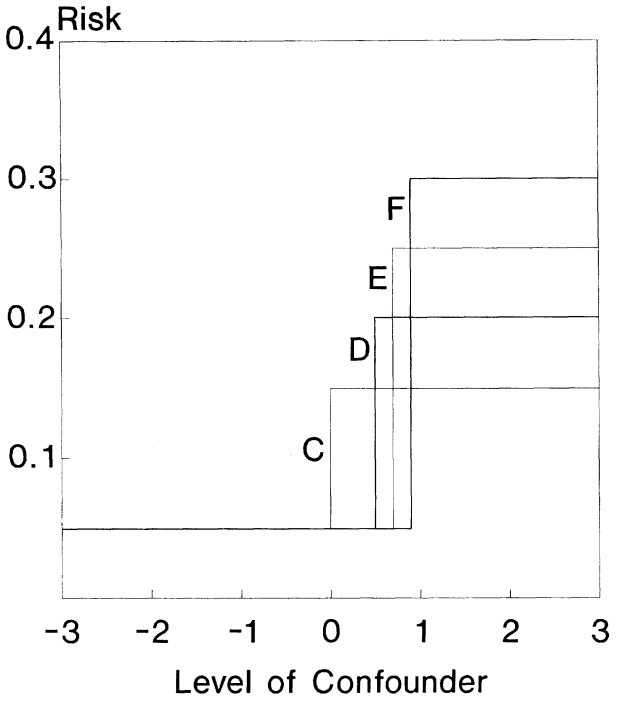

FIGURE 2. Confounder-risk relations assumed in scenarios C, D, E, and F.

Scenarios C, D, E, and F are further examples of situations in which the assumption of a linear increase of the logit of disease risk with X_2 that is implicitly made when including a single linear term of X_2 in the logistic model is clearly violated. Below the threshold, a baseline risk of 5% is assumed throughout. The risk threshold is varied between the 50th (scenario C), 67th (scenario D), 75th (scenario E), and 80th (scenario F) percentile of X_2. The risk above the threshold is varied in such a way that the overall risk of the population equals 0.1 (the risk that pertains to the mean level of X_2, $X_2 = 0$, in scenarios A and B). The relation between the confounder X_2 and disease risk assumed in scenarios C, D, E, and F is depicted in Figure 2.

Generating of Datasets

We generated hypothetical cohort studies with 1,000,000 study participants who were each assigned

- levels of the exposure X_1 and the confounder X_2, using random observations from the bivariate normal distribution with mean 0 and variance 1 for each component and correlation coefficients 0.5, 0.7, and 0.9.
- presence or absence of disease, using a random observation B from the uniform distribution on the interval (0,1); random observations were generated by the SAS function RANUNI[21]; individuals were assumed to be diseased if $B < R$ and undiseased otherwise, where R is the risk of disease derived from the confounder-risk associations listed in Table 1 (note that this procedure is equivalent to generating random binomial deviates for the occurrence of disease).

Types of Analysis

We carried out all analyses by fitting logistic regression models to the cohort data in which the exposure variable X_1 was either

- included as continuous variable, or
- included as a dichotomous variable (with cutpoint 0, the mean level of X_1).

We derived odds ratios with pertinent standard errors for the exposure variable X_1 by exponentiation of the regression coefficients of X_1 in logistic models. For this large population (N = 1,000,000), the variation that arises through the generation of the dataset (assignment of X_1, X_2, and Y) is essentially negligible for practical purposes. The odds ratios can be considered as true odds ratios for this specific population or as expected odds ratios if sampling from the total population were performed. For models including X_1 as a continuous variable, odds ratios refer to an increase of X_1 by 1 unit. For models including X_1 as a dichotomous variable, odds ratios refer to comparisons of individuals with $X_1 \geq 0$ to individuals with $X_1 < 0$. Note that the correctly specified logistic model uses X_1 as a continuous variable; inclusion of X_1 in the regression models as a dichotomous variable is not optimal (see introduction). We nevertheless include this approach, because it is commonly employed in epidemiologic studies. In the specific examples used for illustration in this paper, the correct odds ratio should always equal 1.0 regardless whether the exposure is included as a continuous variable or a dichotomous variable in the model.

We derived odds ratios for the exposure variable (OR) from logistic regression models in which the confounder variable X_2 was either

- not controlled at all (OR_{un})
- included as a single linear term (OR_l)
- included as a quadratic term (OR_{lq}) or both a quadratic and a cubic term (OR_{lqc}) in addition to a linear term
- included as a k-level categorical variable (using $k - 1$ dummy variables); results are presented with X_2 categorized into either two categories of equal size (OR_{cat2}), tertiles (OR_{cat3}), quartiles (OR_{cat4}), or quintiles (OR_{cat5})
  or
- controlled by linear spline regression as outlined in Appendix 1, using two to five equal-sized categories (OR_{spl2}, OR_{spl3}, OR_{spl4}, OR_{spl5}). In contrast to traditional categorical analysis, spline regression makes use of within-category variation of disease risk.

Results

In this section, we illustrate the efficacy of control for confounding by the various strategies of data analysis for the various scenarios of the covariate-risk association.

TABLE 2. Odds Ratios, OR [with Exponentiated Standard Errors of ln(OR)], for an Increase of Exposure X_1 by 1 Standard Deviation as a Function of the Type of Confounder Adjustment

Type of Adjustment	Scenario					
	A	B	C	D	E	F
None						
OR_{un}	1.61 (1.003)	1.31 (1.002)	1.37 (1.003)	1.54 (1.003)	1.66 (1.003)	1.76 (1.003)
Single linear term						
OR_{l}	1.01 (1.004)	1.00 (1.003)	1.00 (1.005)	1.01 (1.005)	1.00 (1.005)	1.01 (1.005)
Linear and quadratic term						
OR_{lq}	1.01 (1.004)	1.01 (1.004)	1.00 (1.005)	1.01 (1.005)	1.00 (1.005)	1.01 (1.005)
Linear quadratic, and cubic term						
OR_{lqc}	1.01 (1.004)	1.01 (1.004)	1.00 (1.005)	1.01 (1.005)	1.00 (1.005)	1.01 (1.005)
Categorization						
OR_{cat2}	1.31 (1.004)	1.11 (1.003)	1.00 (1.004)	1.19 (1.004)	1.32 (1.004)	1.43 (1.004)
OR_{cat3}	1.20 (1.004)	1.03 (1.003)	1.05 (1.004)	1.01 (1.004)	1.13 (1.004)	1.23 (1.004)
OR_{cat4}	1.14 (1.004)	0.98 (1.003)	1.00 (1.004)	1.05 (1.004)	1.01 (1.005)	1.11 (1.005)
OR_{cat5}	1.11 (1.004)	0.96 (1.003)	1.02 (1.005)	1.04 (1.005)	1.04 (1.005)	1.01 (1.005)
Spline regression						
OR_{spl2}	1.01 (1.004)	1.01 (1.004)	1.00 (1.005)	1.01 (1.005)	1.00 (1.005)	1.01 (1.005)
OR_{spl3}	1.01 (1.004)	1.01 (1.004)	1.00 (1.005)	1.01 (1.005)	1.00 (1.005)	1.01 (1.005)
OR_{spl4}	1.01 (1.004)	1.01 (1.004)	1.00 (1.005)	1.01 (1.005)	1.00 (1.005)	1.01 (1.005)
OR_{spl5}	1.01 (1.004)	1.01 (1.004)	1.00 (1.005)	1.01 (1.005)	1.00 (1.005)	1.01 (1.005)

Results are shown for a correlation coefficient of 0.7 between X_1 and X_2. The general patterns were very similar for the other levels of correlation between X_1 and X_2 we examined (except for the fact that the magnitude of confounding increased with increasing correlation).

Continuous Exposure Variable

Table 2 shows the odds ratios for an increase of exposure X_1 by 1 unit in models in which X_1 was included as a continuous variable. In all scenarios that we examined, unadjusted odds ratios erroneously suggested a moderate association between exposure and risk of disease. Control for confounding was therefore essential.

In scenario A, in which the logit of disease risk increases linearly with the confounder, confounding was effectively removed with inclusion of the confounder as single linear term. There was strong residual confounding, however, if the traditional categorical analysis was used. While confounding was removed only very unsatisfactorily with the inclusion of the dichotomized confounder variable in the model ($Or_{cat2} = 1.31$), control for confounding remained far from perfect even with as many as five confounder categories ($OR_{cat5} = 1.11$).

We observed similar patterns in scenario B, although the overall extent of confounding in the crude analysis and in the analyses using the categorized confounder variable was more limited. Despite the J-shaped relation between X_2 and disease risk (see Figure 1), confounding was controlled very effectively in this situation with inclusion of a linear term of X_2, whether or not an additional quadratic or cubic term of X_2 or linear splines were included.

Control for confounding by including the confounder as single linear term was also very effective in the examined scenarios assuming a threshold effect of the confounding variable (scenarios C–F). In these scenarios, odds ratios for X_1 hardly changed if quadratic and cubic terms or linear splines were added. Performance of control for confounding with the categorized confounder variable depends on how well the threshold of risk increase coincides with one of the cutpoints used for categorization. For example, confounding could be effectively controlled for with the 2-level and 4-level confounder variable in scenario C ($OR_{cat2} = 1.00$, $OR_{cat4} = 1.00$), as these variables have a cutpoint at $X_2 = 0$, the assumed risk threshold in scenario C. Control for confounding was likewise effective with the 3-, 4-, and 5-level confounder variable in scenarios D, E, and F owing to the coincidence of the risk threshold with those variables' cutpoints at the 67th, 75th, and 80th percentiles of X_2, respectively. Residual confounding can be severe, however, if none of the cutpoints for categorization is close to the threshold of disease risk. This result was most evident for the dichotomized confounder variable in scenario F. In general, the possibility that none of the cutpoints is close to the threshold of disease risk and, hence, the potential for residual confounding, decreases with the number of categories. Overall, using a continuous variable X_2 performed much better than categorizing the confounder variable in the situations that we examined.

Dichotomous Exposure Variable

For the most part, similar results were obtained in the analyses using a dichotomized exposure variable (Table 3). Control for confounding with a single linear term of the confounder was less complete, however, in scenarios B, C, D, E, and F. While residual confounding could be effectively removed by inclusion of an additional quadratic term in scenario B (as it should, since the model

TABLE 3. Odds Ratios, OR [with Exponentiated Standard Errors of ln(OR)], for Individuals with Exposure Levels $X_1 \geq 0$ Compared with Individuals with Exposure Levels $X_1 < 0$ as a Function of the Type Confounder Adjustment

Type of Adjustment	Scenario					
	A	B	C	D	E	F
None						
OR_{un}	2.14 (1.007)	1.62 (1.004)	1.76 (1.007)	2.14 (1.007)	2.35 (1.007)	2.52 (1.007)
Single linear term						
OR_{l}	1.01 (1.008)	1.08 (1.005)	1.10 (1.008)	1.11 (1.008)	1.07 (1.009)	1.04 (1.009)
Linear and quadratic term						
OR_{lq}	1.01 (1.008)	1.01 (1.007)	1.07 (1.008)	1.11 (1.008)	1.10 (1.009)	1.10 (1.009)
Linear quadratic, and cubic term						
OR_{lqc}	1.01 (1.008)	1.02 (1.007)	1.03 (1.008)	1.03 (1.009)	1.01 (1.009)	1.00 (1.009)
Categorization						
OR_{cat2}	1.37 (1.007)	1.19 (1.005)	1.01 (1.008)	1.29 (1.008)	1.46 (1.008)	1.60 (1.008)
OR_{cat3}	1.19 (1.008)	1.08 (1.006)	1.10 (1.008)	1.00 (1.009)	1.16 (1.009)	1.28 (1.009)
OR_{cat4}	1.12 (1.008)	1.03 (1.006)	1.01 (1.008)	1.08 (1.009)	1.00 (1.009)	1.11 (1.009)
OR_{cat5}	1.08 (1.008)	1.02 (1.006)	1.04 (1.008)	1.06 (1.009)	1.06 (1.009)	1.00 (1.009)
Spline regression						
OR_{spl2}	1.01 (1.008)	1.05 (1.007)	1.08 (1.008)	1.13 (1.009)	1.13 (1.009)	1.11 (1.009)
OR_{spl3}	1.01 (1.008)	1.03 (1.007)	1.00 (1.008)	1.03 (1.009)	1.06 (1.009)	1.09 (1.009)
OR_{spl4}	1.01 (1.008)	1.02 (1.007)	1.00 (1.008)	1.00 (1.009)	1.02 (1.009)	1.06 (1.009)
OR_{spl5}	1.01 (1.008)	1.02 (1.007)	1.01 (1.008)	1.00 (1.009)	0.99 (1.009)	1.02 (1.009)

was correctly specified with a linear and a quadratic term), an additional cubic term was required in scenarios C, D, E, and F to achieve more complete control of confounding. Again, traditional categorical analyses were most error-prone with the use of a dichotomized exposure variable (except for scenario C, in which the threshold of disease risk coincides with the cutpoint for categorization), and, in some instances, yielded unsatisfactory control for confounding even with as many as four or five categories of the confounder. In contrast, control for confounding by linear spline regression was satisfactory in most scenarios that we examined with more than three categories of the confounder.

Discussion

This paper illustrates that there is no single strategy to control for continuous confounders that universally yields the most effective control for confounding. This finding particularly applies to the most commonly used strategies, inclusion of confounders as single linear terms or as categorical variables in multiple regression models.

In theory, control for confounding with categorized variables should be increasingly effective if an increasing number of categories is employed. Inclusion of too many categories will decrease precision, however, particularly if several covariates have to be considered simultaneously, which is the rule rather than the exception in epidemiologic research. On the other hand, our analyses confirm and extend previous findings[14,15] that control for confounding can be very ineffective with classification of individuals into five or less categories. It therefore appears that categorization may often be inadequate when controlling for continuous confounders. The potential advantage of traditional categorical analysis over other options in confounding control that occurs when cutpoints of categorization coincide with eventual thresholds of disease risk appears to be restricted to rather unusual, rare situations.

In the situations that we examined, control for confounding by inclusion of single linear terms tended to be a reasonable choice for a broad variety of confounder-risk associations, even if these associations deviated from the specific shape assumed in the regression model (such as the linear increase of the logit of disease risk with the level of covariates assumed in multiple logistic regression). These results may not hold in general, as the examples were restricted to a small region of the parameter space. Even within this restricted parameter space, control for confounding by inclusion of a single linear term remained suboptimal in some situations (see scenarios B, C, D, E, and F in Table 3). Therefore, various approaches to improve control for confounding may be worthwhile.

Two general types of such approaches have been addressed in this paper. The first approach is to include additional nonlinear terms of the confounder. Inclusion of a quadratic or both a quadratic and a cubic term along with a linear term assessed in this paper are simple examples of such polynomial regression. The second approach is control for confounding by spline regression. The type of spline regression assessed in this paper can be considered a refinement of both control for confounding by a single linear term (in that it allows for changes of the slope of the confounder-risk association between categories of the confounder) and control for confounding by the traditional categorical analysis (in that it makes use of within-category variation of disease risk). Our analyses suggest that both polynomial and spline regression approaches may considerably reduce residual confounding in some, but not all, instances. It

may be worthwhile to use such strategies more often in epidemiologic research, particularly in situations in which weak exposure effects are assessed and potentially strong confounders have to be controlled for. A typical example would be studies on the effects of air pollution or radon on the risk of lung cancer in the presence of confounding by smoking. The commonly employed crude classification of smoking in such studies may often be insufficient for effective control of smoking effects.

In this paper, we restricted examples to bivariate normal distributions of covariates. Clearly, other distributions are more relevant in specific situations. Likewise, the covariate-risk associations assessed in this paper reflect special (and partly unrealistic) situations. They were chosen to facilitate illustration of the conditions under which control for confounding by various strategies may or may not work. The general approach introduced in this paper may be easily adapted by investigators to assess the performance of various strategies to control for confounding under the specific circumstances in their areas of research.

Examples in this paper were also restricted to particularly simple applications of polynomial regression and spline regression to facilitate illustration of the main principles. These methods may be further refined (for example, by inclusion of fractional and inverse powers of covariates in polynomial regression or by inclusion of quadratic terms in spline regression) to attain more complete control of confounding.[22] Our examples indicate, however, that such refinement may rarely be necessary in practice. Inclusion of too many terms would tend to decrease the precision and may even be harmful rather than beneficial if the size of the study population is limited and several confounders have to be controlled for.

Application of polynomial regression and spline regression has been proposed for more flexible estimation of exposure-risk associations in epidemiology. The price for the gain in flexibility, however, is a less parsimonious description of the exposure-disease association, which may hinder communication of main results. More flexible modeling of the confounder-risk association allows for more complete control of confounding without sharing this potential drawback (provided that the confounder-risk association is not of primary interest). Control for confounding may therefore be a particularly attractive field of application of techniques like polynomial regression or spline regression.

References

1. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: John Wiley and Sons, 1989.
2. Cox DR. Regression models and life tables. J R Stat Soc 1972;34B:187–220.
3. Rothman KJ. Modern Epidemiology. Boston: Little, Brown, 1986.
4. Zhao LP, Kolonel LN. Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. Am J Epidemiol 1992;136:464–474.
5. Lagakos SW. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. Stat Med 1988;7:257–274.
6. Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology 1995;6:450–454.
7. Wartenberg D, Northridge M. Defining exposure in case-control studies: a new approach. Am J Epidemiol 1991;133:1058–1071.
8. Altman DG. Categorising continuous variables (Letter). Br J Cancer 1991; 64:975.
9. Hsieh C-c, Maisonneuve P, Boyle P, Macfarlane GJ, Robertson C. Analysis of quantitative data by quantiles in epidemiologic studies: classification according to cases, noncases, or all subjects? Epidemiology 1991;2:137–140.
10. Altman DG. Problems in dichotomizing continuous variables (Letter). Am J Epidemiol 1994;139:442.
11. Schulgen G, Lausen B, Olsen JH, Schumacher M. Outcome-oriented cutpoints in analysis of quantitative exposures. Am J Epidemiol 1994;140:172–184.
12. Maclure M, Greenland S. Tests for trend and dose response: misinterpretations and alternatives. Am J Epidemiol 1992;135:96–104.
13. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology 1995;6:356–365.
14. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 1968;24:295–313.
15. Becher H. The concept of residual confounding in regression models and some applications. Stat Med 1992;11:1747–1758.
16. Cederlöf R, Doll R, Fowler B, Friberg L, Nelson N, Vouk V. Air pollution and cancer: risk assessment methodology and epidemiological evidence. Environ Health Perspect 1978;22:1–12.
17. Weinberg CR. How bad is categorization? (Editorial). Epidemiology 1995; 6:345–347.
18. Lee I-M, Manson JE, Hennekens CH, Paffenbarger RS. Body weight and mortality: a 27-year follow-up of middle-aged men. JAMA 1993;270:2823–2828.
19. Poikolainen K. Alcohol and mortality: a review. J Clin Epidemiol 1995;48: 455–465.
20. Ulm K. A statistical method for assessing a threshold in epidemiologic studies. Stat Med 1991;10:341–349.
21. SAS Institute. SAS Language: Reference. version 6. 1st ed. Cary, NC: SAS Institute, 1990.
22. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). Appl Stat 1994;43:425–467.

Appendix 1

Spline regression is a simple alternative to basic categorical analysis that can easily be performed with conventional regression programs. As with conventional categorical analysis, continuous covariates are categorized in the first place. In contrast to conventional categorical analysis, however, spline regression allows for (1) continuity in risk estimates across categories (that is, no sudden jumps at the cutpoints of categorization) and (2) variation of risk within categories.

In this paper, we used the following simple example of inclusion of linear splines into the logistic regression model to control the association of the exposure variable X_1 with disease risk for confounding by the continuous covariate X_2:

$$\text{logit}\,(R|X_1,X_2) = \alpha + \beta X_1 + \gamma_1 X_2 + \gamma_2 s_2 + \ldots + \gamma_k s_k,$$

where α denotes the intercept and β and γ_i denote the regression coefficients, X_2 is assumed to be categorized into k categories, c_i $(i=2,\ldots,k)$ denotes the cutpoint between the $(i-1)$th and the ith category of the categorized covariate, and $s_i = 0$ if $X_2 \leq c_i$, $X_2 - c_i$ if $X_2 > c_i$.