STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:3487–3497
Published online 25 October 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2737

Biased odds ratios from dichotomization of age

Henian Chen[1, 2, 3, *, †], Patricia Cohen[1, 2, 3] and Sophie Chen[4]

[1]*Epidemiology of Mental Disorders, New York State Psychiatric Institute, New York, U.S.A.*
[2]*Department of Psychiatry, College of Physicians and Surgeons, Columbia University, New York, U.S.A.*
[3]*Department of Epidemiology, Joseph L. Mailman School of Public Health, Columbia University, New York, U.S.A.*
[4]*Calgary Health Region, Alberta, Canada*

SUMMARY

Dichotomizing a continuous variable is known to result in the loss of information, lower statistical power, and lower reliability. In many epidemiological studies, age is a scaled (continuous) variable prior to statistical analyses; however, despite pleas from methodologists, researchers frequently dichotomize age in their data analysis without an appropriate rationale. Using simulated case–control data, we show that dichotomizing age generally will lead to a biased odds ratio (OR). When age was a confounder (potentially representing common causes of risks and outcomes), including age as a scaled variable (whether the age effect was linear or non-linear in the logit), provided satisfactory control, whereas when age was categorized, the estimated risk factor effect was biased. We also demonstrate that the further the cutpoint is from the median age, the greater the increase in the OR; thus, in cases where age dichotomization is warranted, researchers are cautioned not to allow the size of the empirical OR to influence their choice of cutpoint. Recommendations are made for analysing age in epidemiological data and interpretation of empirical findings. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: odds ratio; age; confounding; logistic regression; dichotomizing; cutpoint

Dichotomizing a continuous or scaled variable is known to result in the loss of efficiency [1], lower reliability [2] and statistical power [3–5], and higher type I [6–8] and type II [9] errors. The disadvantages of dichotomization are related to the fact that it does not make use of

*Correspondence to: Henian Chen, Columbia University/NYSPI, 100 Haven Avenue, 31F, New York, NY 10032, U.S.A.
†E-mail: chenhen@pi.cpmc.columbia.edu

Contract/grant sponsor: National Institute of Child Health and Human Development; contract/grant number: HD-40685
Contract/grant sponsor: National Institute of Mental Health; contract/grant number: MH-60911

Received 15 February 2006
Accepted 21 September 2006

within-category information; nevertheless, researchers frequently categorize continuous variables and run the same risks. One common dichotomization splits the continuous variable at the sample median, thereby defining high and low groups. A median split with the resulting 50–50 distribution of the sample minimizes the reduction in the correlation in comparison to the scaled measure, assuming a linear relationship. In this case the expected r is reduced by approximately 20 per cent (i.e. a multiplicative factor of 0.80). Under the same assumption, if the split is more extreme, then the reduction is greater: for example, with a 90–10 split (or equivalently for a 10–90 split), the reduction is 41 per cent, a multiplicative factor of 0.59 [3, 10].

Nevertheless, even in research areas where these issues are a standard aspect of statistical instruction, substantive researchers often dichotomize continuous independent variables prior to conducting analyses [11]. MacCallum and colleagues [2] reviewed all articles published from January 1998 through December 2000 in three leading journals in the psychology field. They found that, out of a total of 958 studies, there was at least one instance of dichotomization of a quantitatively measured variable in 110 studies (11.5 per cent); moreover, only 22 of those studies (20 per cent) were accompanied by an explicit justification for dichotomization.

Researchers have argued that loss of power and efficiency is unimportant if statistically significant effects still are found with dichotomous variables [7, 8]; thus, why should research reports include a coefficient that may be less likely to be understood by typical research consumers? Indeed, if Type I and II errors were the only consequences and the results still met criteria for statistical significance, researchers might be justified in dichotomizing such measures so as to simplify conclusions. What is less appreciated, however, is that dichotomizing a continuous variable also yields biased measures of effect size, including the odds ratio (OR).

In many circumstances in epidemiology, age is a scaled variable prior to statistical analyses. Most epidemiological studies examine the association of age with the dependent variable, or control age as a potential confounder of associations between the dependent variable and one or more independent variables of interest [12]. Without a suitable rationale, researchers frequently categorize age into two or more groups such as ‘younger’ and ‘older’ individuals, and use these groups to estimate appropriate parameters (e.g. OR). Here, we examined all the articles on studies testing age effects published in the *American Journal of Epidemiology* (AJE) for 5 years (2000–2004). We selected AJE because this popular journal is frequently cited. We found that 308 studies utilized age categories in the statistical analyses, and of these, 75 (24.4 per cent) dichotomized age and only 10 (3.2 per cent) offered any justification for doing so or for the particular cutpoint chosen.

The epidemiologist handles age in two ways: as a continuous variable, or as a categorized variable by combining a number of adjacent ages into a joint category. As we have seen above, age is often categorized into a dichotomous variable. The advantages of dichotomizing age include simplifying the statistical analysis, leading to easy interpretation and presentation of results. Aside from statistical convenience, however, there are three other classes of reasons for employing age as a categorized variable.

1. *Decision-oriented research*. Epidemiologists tend to use logistic regression and to dichotomize age. Part of the motivation in clinical and epidemiological studies, including those related to health outcomes, is that researchers need to frame the question in ways that will be obviously relevant to policy-makers or clinicians [13, 14].
2. *Study design considerations*. Study participants may be recruited as groups thought to differ on an outcome of interest. In some such studies age may be deliberately dichotomized prior

DOI: 10.1002/sim

to subject selection in order to maximize the statistical power to find a significant association between age and the outcome of interest [15]. Under this circumstance, even if there is a linear relationship between age and the outcome in the population, maximizing the difference between the means of two groups of participants on age will maximize the statistical power to find an associated difference on outcome for a fixed sample size.

3. *Secondary data analysis or meta-analysis*. Age may have been measured only as a dichotomy, or by means of an ordinal variable whose few levels are not well distributed, or published in five-year, ten-year age groups. Often in meta-analysis, for some studies the only available measure of age is by category. Here the issue is whether to use the available data to contribute information on a new or important issue or to abandon the effort.

In this study we use simulated data to show that (1) using dichotomized age as a risk factor in logistic regression models results in a biased OR; (2) using categories of age as a confounding variable is less satisfactory than using age as a continuous variable, whether the age effect is linear or non-linear in the logit; (3) the use of dichotomous age may considerably bias the effect size of the independent variable of interest.

AGE AS A RISK FACTOR

Linearity in the logit

We generated a simulated case data set ($n = 1000$, mean age $= 50$, SD $= 5$) and two control data sets ($n = 1000$, SD $= 5$, and mean ages 47 and 45 for studies A and B, respectively) by random sampling from normal distribution populations employing the SAS function RANNOR [16]. On using age as a scaled variable in a logistic regression model, we assume that a linear relationship exists between age and logit(Y). We begin by examining this assumption using the method of fractional polynomials [17].

Fractional polynomial regression (FP). FP is a simple parametric approach to modelling relationships between response and continuous risk factors. FP allows the variable to enter the model following transformation from a predefined class of functions. Royston and Altman [18] have emphasized that a great deal of more flexibility and stability can be obtained by examining fractional and inverse powers of the independent variable (x). Royston *et al.* [18, 19] point out that models containing as few as three different powers of x between x^{-2} and x^3 encompass a dramatic range of shapes. For a first order FP, $\beta_0 + \beta_1 * X^p$, the power is chosen from candidates $p \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. For a second order FP, the models are of the form $\beta_0 + \beta_1 * X^p + \beta_2 * X^q$ or, for the mathematical limit $p = q$, the models are $\beta_0 + \beta_1 * X^p + \beta_2 * X^q * \log X$. As before, p and q are chosen from $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. We employed SAS Macro MFP [20] to fit FP models to examine the relationship between age and outcome for both studies A and B. Consistent with the simulation, the straight line model with linear age was found as the best model for age and outcome for both study A and B.

Logistic regression. Because the results from FP models support linearity, we can fit two logistic regression models using age as a continuous variable for studies A and B. The logistic regression coefficients (SE) were 0.12 (0.01) for study A and 0.19 (0.01) for study B. OR (95 per cent CI) were 1.13 (1.10–1.15), and 1.21 (1.19–1.24) for studies A and B, respectively.

DOI: 10.1002/sim

Table I. Odds ratios (95 per cent CI) for varying cutpoints on age for four simulated data sets.

Age was cut at per cent	Study A	Study B	Study C	Study D
5	3.16 (1.99–5.02)	8.78 (4.67–16.54)	1.92 (1.26–2.92)	4.55 (2.74–7.56)
10	2.76 (2.00–3.80)	8.27 (5.39–12.71)	2.19 (1.61–2.98)	6.62 (4.44–9.87)
20	2.54 (2.01–3.20)	6.05 (4.62–7.94)	2.31 (1.84–2.91)	5.61 (4.30–7.33)
30	2.53 (2.02–3.13)	5.17 (4.16–6.42)	2.33 (1.91–2.84)	5.43 (4.36–6.76)
40	2.52 (2.10–2.99)	4.95 (4.07–6.01)	2.94 (2.44–3.54)	5.76 (4.72–7.02)
50	2.47 (2.06–2.95)	4.71 (3.87–5.72)	3.30 (2.75–3.97)	6.75 (5.55–8.20)
60	2.69 (2.23–3.23)	4.86 (4.02–5.88)	3.53 (2.92–4.27)	7.27 (5.92–8.93)
70	2.64 (2.16–3.22)	5.30 (4.26–6.59)	4.37 (3.54–5.40)	10.03 (7.84–12.84)
80	2.89 (2.29–3.66)	5.83 (4.45–7.62)	4.78 (3.70–6.18)	18.28 (12.50–26.73)
90	3.45 (2.47–4.83)	7.21 (4.79–10.87)	6.91 (4.61–10.35)	60.70 (22.46–164.06)
95	5.64 (3.28–9.70)	9.79 (5.06–18.93)	12.56 (6.07–26.02)	109.77 (15.28–788.60)

Note: Study A: case ($n = 1000$, mean age $= 50$, SD $= 5$) and control ($n = 1000$, mean age $= 47$, SD $= 5$); study B: case ($n = 1000$, mean age $= 50$, SD $= 5$) and control ($n = 1000$, mean age $= 45$, SD $= 5$); study C: case ($n = 1000$, mean age $= 50$, SD $= 5$) and control ($n = 1000$, mean age $= 47$, SD $= 4$); study D: case ($n = 1000$, mean age $= 50$, SD $= 5$) and control ($n = 1000$, mean age $= 45$, SD $= 4$).

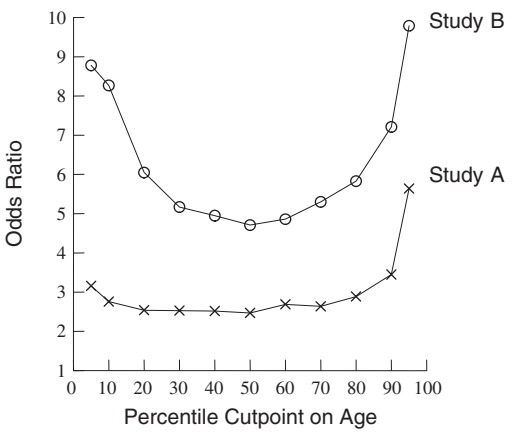

Figure 1. Odds ratio for varying cutpoints on age (study A and B).

ORs and different cutpoints of age. Suppose instead of using age as a scaled variable we dichotomize it, as is done so frequently. We choose some point at which to cut, calling those below the cutpoint 'Younger' and those above the cutpoint 'Older'. Table I and Figure 1 show the relationship between the different cutpoints on age and the ORs for both studies A and B. Age was cut at points ranging from the 5th percentile to 95th percentile. The minimum value of the OR appears when age was cut at the median, increasing with their distance away from the median age. Of course, the magnitude of OR is also dependent on the age difference between case and control. The OR curve for extreme cuts on age accelerates when there is a bigger age difference between case and control groups.

DOI: 10.1002/sim

Non-linearity in the logit

We next produced two simulated data sets in which age was associated with the outcome (Y), but the relationship between age and logit(Y) was non-linear. The simulated data sets for study C and D differed from study A and B in that the SD of age was 4 rather than 5 for the control groups (case: $n = 1000$, mean age $= 50$, SD $= 5$; control C: $n = 1000$, mean age $= 47$, SD $= 4$; control D: $n = 1000$, mean age $= 45$, SD $= 4$). We examined the linear assumption by employing SAS Macro MFP to fit FP models. Consistent with the simulation, $\text{logit}(Y) = -130.03 - 2416573.34 * \text{age}^{-2} +$

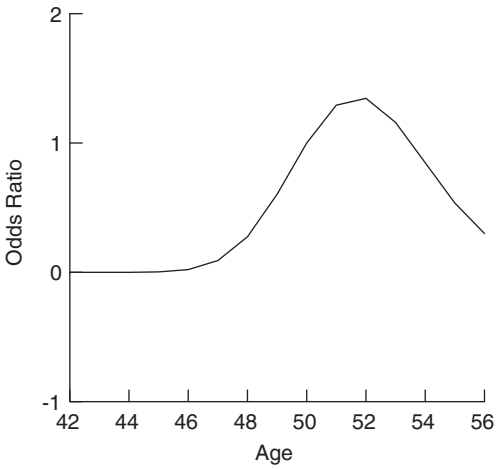

Figure 2. Study C, odds ratios for age (OR $= 1$ for age $= 50$).

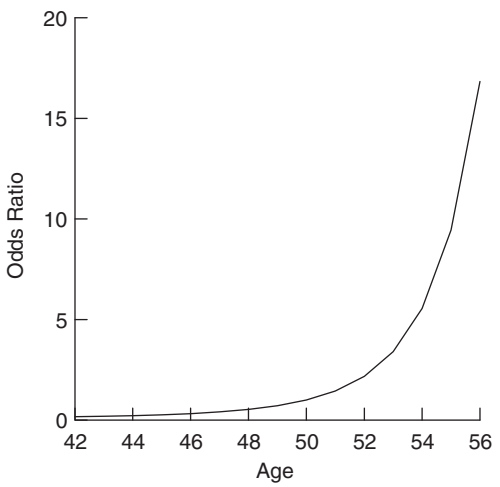

Figure 3. Study D, odds ratios for age (OR $= 1$ for age $= 50$).

DOI: 10.1002/sim

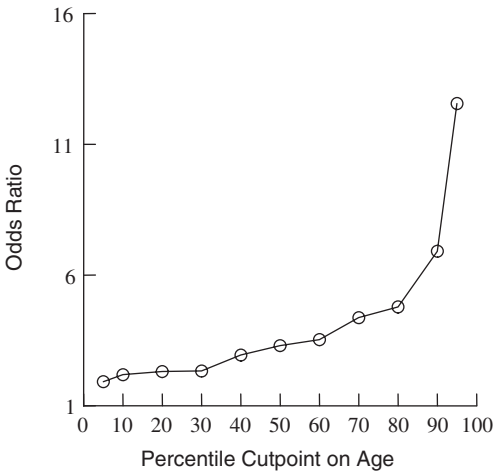

Figure 4. Odds ratio for varying cutpoints on age (study C).

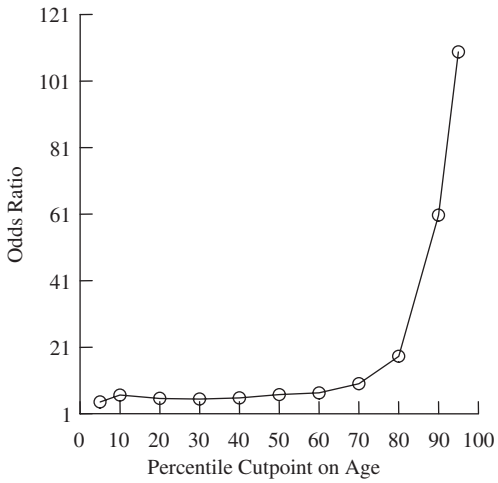

Figure 5. Odds ratio for varying cutpoints on age (study D).

$701397.63 * \text{age}^{-2} * \log(\text{age})$ was found as the best model for study C. $\text{Logit}(Y) = 0.85482 - 0.00059 * \text{age}^3 + 0.00015 * \text{age}^3 * \log(\text{age})$ was found as the best model for study D. Thus, non-linearity was created by the discrepancy in variance between case and controls. Figures 2 and 3 demonstrate that the resulting relationship between outcome and age for study C and D were not linear.

Once more, we dichotomize age using various cutpoints, creating 'Younger' and 'Older' groups. Table I and Figure 4 show the relationship between the different cutpoints on age and the ORs for study C. We see that the minimum value of the OR appears when age is cut at 5 per cent. The

DOI: 10.1002/sim

ORs are increased from 1.92 to 12.56 when age is cut from 5 to 95 per cent. Table I and Figure 5 show the relationship between the different cutpoints on age and the ORs for study D. We see that the minimum value of the OR also appears when age is cut at 5 per cent. The ORs are increased from 4.55 to 109.77 when age is cut from 5 to 95 per cent.

AGE AS A CONFOUNDER

Linearity in the logit

Age is often treated as a confounder rather than a risk variable of primary interest. We generated a simulated data set of study E based on study A by adding one binary exposure (X) to the data set and age was treated as a confounder. The correlation coefficients are 0.37 (Y and X), 0.28 (Y and age), and 0.36 (X and age). Table II shows the ORs of X and deviances (-2 log likelihood) from different logistic regression models produced by continuous age and various categorizations of age (two to nine groups).

The logistic regression model controlling three groups of age was significantly improved ($X^2 = 2487.60 - 2475.79 = 11.81$, $\mathrm{df} = 1$, $p < 0.01$) compared with the model including age as a dichotomous variable. The four age group model was not significantly improved, but the model controlling five groups of age is significantly improved ($X^2 = 2475.75 - 2469.22 = 6.53$, $\mathrm{df} = 1$, $p < 0.05$) compared to the model controlling four groups of age. No model was significantly improved by increasing the number of age groups from five to nine. However, the model controlling age as a continuous variable was significantly improved (X^2 from 9.4 to 28.69, $\mathrm{df} = 1$, all $p < 0.01$) compared to all models including two to nine groups of age (Figure 6).

The OR of X, the presumed variable of interest, is 4.73 without controlling age. Its OR is 3.62 after controlling age as a continuous variable. However, its ORs range from 4.21 to 3.71 when age is categorized into two to nine groups (Figure 7). As we can see, when age is categorized, the test of the risk factor (X) can be overestimated by 10 to 59 per cent.

Table II. Odds ratios (95 per cent CI) of X and deviances from different logistic regression models by adjusting continuous and categorized age.

	Study E		Study F	
Predictor	Deviance	OR (95 per cent CI)	Deviance	OR (95 per cent CI)
X		4.73 (3.92–5.72)		4.73 (3.92–5.72)
X + continuous age	**2458.91**	**3.62** (2.94–4.45)	**2434.46**	**3.20** (2.59–3.95)
X + 2 groups of age	***2487.60***	***4.21*** (3.40–5.21)	***2471.54***	***3.57*** (2.86–4.45)
X + 3 groups of age	2475.79	3.90 (3.17–4.80)	2449.88	3.39 (2.74–4.18)
X + 4 groups of age	2475.75	3.87 (3.14–4.77)	2453.72	3.34 (2.69–4.15)
X + 5 groups of age	2469.22	3.74 (3.03–4.61)	2444.81	3.26 (2.63–4.04)
X + 6 groups of age	2470.98	3.77 (3.06–4.65)	2446.07	3.23 (2.61–4.01)
X + 7 groups of age	2471.58	3.77 (3.05–4.65)	2447.56	3.29 (2.65–4.08)
X + 8 groups of age	2468.31	3.72 (3.02–4.59)	2445.57	3.23 (2.60–4.01)
X + 9 groups of age	2468.60	3.71 (3.00–4.58)	2444.91	3.23 (2.60–4.01)

DOI: 10.1002/sim

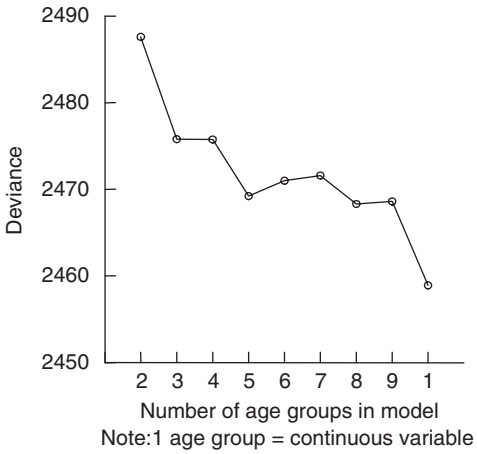

Figure 6. Deviances from different logistic regression models by adjusting continuous and categorized age (study E).

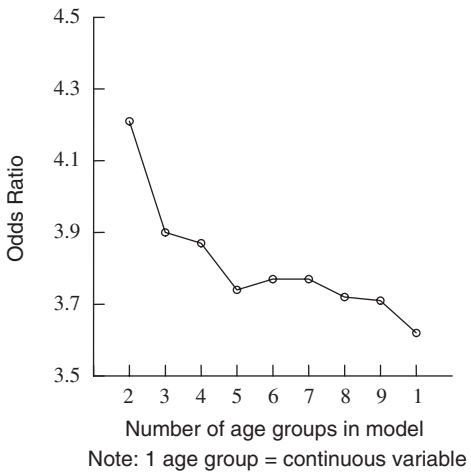

Figure 7. Odds ratios from different logistic regression models by adjusting continuous and categorized age (study E).

Non-linearity in the logit

A study F simulated data set was generated based on study C with one binary risk factor (X) and, again, age was treated as a (non-linear) confounder. The correlation coefficients are 0.37 (Y and X), 0.32 (Y and age), and 0.40 (X and age). Table II shows the ORs of X and deviances from different logistic regression models including scaled age and various categorizations of age (two to nine groups). The model that includes scaled age is a significant improvement (X^2 from 10.35 to 37.08, df = 1, all $p<0.01$) compared with all models with age represented as two to

DOI: 10.1002/sim

nine categories. The OR of X was 4.73 in the model without age, 3.20 after controlling age as a scaled variable. The ORs are from 3.57 to 3.23 after controlling categorized age with groups from two to nine. We can see that inclusion of age as a continuous variable provides better control for confounding even in situations in which the model assumption of linearity is clearly violated.

SUMMARY

It has long been common to categorize age into two or more groups when estimating the odds of being in one of two dependent variable classifications. Dichotomizing age is known to result in the loss of efficiency, lower reliability and statistical power, higher Type I and Type II errors, and a biased OR [21]. Forcing individuals into two groups, 'younger' and 'older,' is widely perceived to simplify analyses and facilitate presentation and interpretation of findings. In fact, age is so frequently dichotomized that some may believe this to be a recommended practice. Indeed, categorization of age was required by classical methods used to analyze epidemiological data, for example, the Mantel–Haenszel estimate, that were based primarily on contingency table analysis. However, with the advent of increased computer power and associated programs, such methods have been replaced by regression models which do not require categorization. It is understandable that researchers interested in substantive issues reach for tried and familiar methods of analysis, especially when they appear to produce satisfactory results. Use of 'younger' and 'older' age groups in logistic regression models also yields results that are readily understood by colleagues, policy makers and the interested public. Nevertheless, as we have demonstrated in this paper, such results often distort the reported relationship no matter at what point cut on age is made.

Age often is treated as a confounder in epidemiological data analyses and some researchers will assume that biased estimates of age effects will have no bearing on the effects of the independent variables of interest. Becher [22], however, demonstrated that residual confounding arises when a continuous confounder is divided into a categorical variable for use in multivariate logistic regression. Brenner and Blettner [23, 24] reported that categorization often is inadequate when controlling for continuous confounders, and that crudely categorized covariates can result in misleading estimates of the association between exposure to another risk factor and an outcome of interest. Austin and Brunner [6] found that the inflation of type I error rate is induced by the residual confounding that occurs when a continuous confounder such as age is categorized. In our simulated data, we showed that inclusion of age as a continuous variable provides satisfactory control as a confounder, but that when age is categorized the test of the risk factor may be significantly biased. This is true even in situations in which the model assumption of linearity clearly is violated.

Many researchers favour using logistic regression models due to the ease of interpreting estimated effects; however, even if the aim is only to adjust for age rather than to obtain its estimated association, we see high costs without compensatory benefits to categorizing age in such models.

RECOMMENDATIONS

1. Never choose an 'optimal' cutpoint on age based on maximizing the OR. This simply exploits the distortion introduced by dichotomization. If age is to be dichotomized, the choice of

DOI: 10.1002/sim

cutpoint should be made prior to analysis and with some theoretical justification. A choice of cutpoint far away from the median should be avoided.

2. Comparisons of findings across studies using OR as effect size measures need to take differential cut-off points of age into account, including those reflected in or produced by study design and participant recruitment. For example, the constraints imposed by the statistical power in small samples may lead to variable 'cuts' that are close to the median, at a point where the age OR may be near its minimum. Such investigations will not generally have effects comparable to reported effects of more extreme cuts from large studies.
3. Always check the assumption of linearity in the logit of age. If this assumption is supported, age can best be used as a continuous variable in a logistic regression model. Careful interpretation of the resulting OR as representing the effect for each additional year can be readily understood by readers. This is clearly better than introducing bias or distortion into the results.
4. When age is highly skewed or its relation with the outcome is non-linear, the best strategy is to use other regression models such as fractional polynomials regression [17–20] instead of a logistic regression model. In this case, exponentials of continuous age will provide an adequate estimate of the effect of age and also an adequate control in the consideration of some other age-correlated variable of interest. It may result in incredible results if age will present estimates for only age without its exponentials.
5. When age is a potential confounder, it should be treated as a continuous variable in the logistic regression model. Inclusion of age as a continuous variable generally provides more satisfactory control than any categorizations of age even in situations in which the model assumption of linearity is clearly violated.

ACKNOWLEDGEMENTS

This study was supported by the National Institute of Child Health and Human Development (grant HD-40685) and National Institute of Mental Health (grant MH-60911).

REFERENCES

1. Zhao LP, Kolonel LN. Efficiency loss from categorizing quantitative exposures into qualitative exposures in case–control studies. *American Journal of Epidemiology* 1992; **136**:464–474.
2. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. *Psychological Methods* 2002; **7**:19–40.
3. Cohen J. The cost of dichotomization. *Applied Psychological Measurement* 1983; **7**:249–253.
4. Ragland DR. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology* 1992; **3**:434–440.
5. Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose–response and tread analysis. *Epidemiology* 1995; **6**:450–454.
6. 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* **23**:1159–1178.
7. Vargha A, Rudas T, Delaney HD, Maxwell SE. Dichotomization, partial correlation, and conditional independence. *Journal of Educational and Behavioral Statistics* 1996; **21**:264–282.
8. Maxwell SE, Delaney HD. Bivariate median splits and spurious statistical significance. *Psychological Bulletin* 1993; **113**:181–190.
9. Streiner DL. Breaking up is hard to do: the heartbreak of dichotomizing continuous data. *Canadian Journal of Psychiatry* 2002; **47**:262–266.

DOI: 10.1002/sim

10. Cohen J, Cohen P, West SG, Aiken LS. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd edn). Lawerence Erlbaum Associates: Mahwah, NJ, 2003.
11. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* 2006; **25**:127–141.
12. Reijneveld SA. Age in epidemiological analysis. *Journal of Epidemiology and Community Health* 2003; **57**:397.
13. Farrington DP, Loeber R. Some benefits of dichotomization in psychiatric and criminological research. *Criminal Behaviour and Mental Health* 2000; **10**:100–122.
14. Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. Measuring the potency of a risk factor for clinical or policy significance. *Psychological Methods* 1999; **4**:257–271.
15. Allison DB, Allison RL, Faith MS, Paultre F, Pi-Sunyer FX. Power and money: designing statistically powerful studies while minimizing financial costs. *Psychological Methods* 1997; **2**:20–33.
16. SAS Institute, Inc. *Statistical Analysis System*, *Version 9.0*, SAS Institute, Cary, NC, 2005.
17. Hosmer DW, Lemeshow S. *Applied Logistic Regression* (2nd edn). Wiley: New York, 2000.
18. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling. *Applied Statistics* 1994; **43**:429–467.
19. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* 1999; **28**:964–974.
20. 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* 2006; **50**:3464–3485.
21. Altman DG, Royston P. The cost of dichotomizing continuous variables. *Biometrical Journal* 2006; **332**:1080.
22. Becher H. The concept of residual confounding in regression models and some applications. *Statistics in Medicine* 1992; **11**:1747–1758.
23. Brenner H, Blettner M. Controlling for continuous confounders in epidemiological research. *Epidemiology* 1997; **8**:429–434.
24. Brenner H. A potential pitfall in control of covariates in epidemiologic studies. *Epidemiology* 1998; **9**:68–71.

DOI: 10.1002/sim