Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Takahashi et al., J Biomet Biostat 2013, 4:5
DOI: 10.4172/2155-6180.1000181

Research Article Open Access

Cubic Spline Regression of J-shaped Dose-Response Curves with Likelihood-based Assignments of Grouped Exposure Levels

Kunihiko Takahashi[1]*, Hiroyuki Nakao[2] and Satoshi Hattori[3]

[1]*Department of Biostatistics, Nagoya University Graduate School of Medicine, Japan*
[2]*Center for Public Health Informatics, National Institute of Public Health, Japan*
[3]*Biostatistics Center, Kurume University, Japan*

Abstract

In epidemiological studies that measure the risk at different levels of exposure, the data is often only available for the analyses that summarized the response data in grouped exposure intervals. In typical methods, the midpoints are used as the assigned exposure levels for each interval. Results of the analysis with grouped data may be sensitive to the assignment of the exposure levels. In this paper, we propose a procedure for assessing J-shaped associations based on the likelihood-based assignment of values to grouped intervals of exposure, and applying the cubic spline regression models. Numerical illustrations and comparisons based on simulations showed that the proposed procedure can yield better estimates for curves than those obtained using the typical assignment method based on the midpoints of each interval.

Keywords: Dose-response; Spline regression model; Grouped exposure interval; Dose assignment; Likelihood; J-shaped curve

Introduction

In epidemiological studies, it is often necessary to determine the relationship between exposure levels and the risk of disease. However, the data on exposure levels are often available in intervals because they are generally not recorded for each individual subject. For example in studies on the association between alcohol consumption and the risk of disease, researchers often treat the exposure levels as intervals when they interview participants about their consumption levels, but they are unable to obtain the exact values as continuous variables. Also in traditional meta-analysis based on aggregated data, it is not possible to obtain the original data, and the published articles do not include enough data. In such situations, meta-analysis of observational studies often has to rely on the summarized data where the exposure levels are grouped into intervals available from research reports.

Table 1 summarizes data from a study of the association between alcohol consumption and all-cause mortality, which was conducted by Lin et al. [1]. The alcohol intake of current drinkers was classified into five groups: non-drinkers, alcohol intake of 0.1-22.9 g/day, 23.0-45.9 g/day, 46.0-68.9 g/day, and ≥ 69.0 g/day. Table 2 summarizes the characteristics of two studies of coffee consumption and stroke, by Bidel et al. [2] and Grobbee et al. [3]. They were included into a meta-analysis of 11 studies by Larsson and Orsini [4], where the categories used for coffee consumption differed among the studies. Moreover the reference category was assigned to non-drinkers in Grobbee et al. [3], whereas 0-2 cups/day were used in Bidel et al. [2], and hence the meanings of the reported relative risks (RRs) were different and it was inappropriate to combine them directly.

Studies that measure the risk at different levels of exposure are usually analyzed based on a trend estimate by linear (or log-linear) regression analysis. When performing a regression analysis of summarized response data that are grouped into intervals, many researchers use the pre-assigned exposure levels from historical data or the midpoint values of each interval [5]. Results of regression analysis with grouped data may be sensitive to the assignment of the exposure levels. Recently, Takahashi and Tango [6] proposed a method for assigning values by applying the likelihood approach, and they showed the procedure can produce a more accurate linear regression coefficient than the typical procedure which uses the midpoint values.

On the other hand, some studies have reported that the risk of disease has a nonlinear relationship with the exposure level. For example, it is known that the association between alcohol and coronary heart disease [7] or total mortality [8] may be depicted as a J-shaped curve. Some evidence of a J-shaped association has also been reported recently between coffee consumption and the risk of stroke [4]. Curve-fitting methods based on a fractional polynomial model or a spline model has often been applied to these nonlinear dose-response associations for regression [9-11]. Di Catelnuovo et al. [8] applied fractional polynomials to fit the association between alcohol intake and the RR of total mortality in a meta-analysis of 34 prospective studies. On the other hand, for example, Larsson and Orsini [4] performed a dose-response meta-analysis and detected a potentially nonlinear association between coffee consumption and stroke using a cubic spline model, and cubic spline regression models may have many advantages over polynomials [12].

In this paper, we propose a procedure for assessing nonlinear associations between exposure levels and the risk of disease from a summarized grouped data, which is based on the assignment of levels to grouped exposure intervals by applying the likelihood-based assignment procedure proposed in Takahashi and Tango [6]. In particular, we focus on the restricted cubic spline model that was described in Orsini et al. [13] and Larsson and Orsini [4] for J-shaped dose-response curves. We demonstrate how to estimate a J-shaped curve from the grouped summarized data using only four or five class intervals. Also this procedure can provide the log relative risk on each

***Corresponding author:** Department of Biostatistics, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, Nagoya, 466-8550, Japan, Tel: +81-52-744-2489; Fax: +81-52-744-2488; E-mail: kunihiko@med.nagoya-u.ac.jp

Received November 18, 2013; **Accepted** November 28, 2013; **Published** November 30, 2013

Citation: Takahashi K, Nakao H, Hattori S (2013) Cubic Spline Regression of J-shaped Dose-Response Curves with Likelihood-based Assignments of Grouped Exposure Levels. J Biomet Biostat 4: 181. doi:10.4172/2155-6180.1000181

Alcohol intake (g/day)	No. of individuals	Person-years	No. of deaths	Adjusted RR (log value)	95% CI
Nondrinkers	7,839	75,352	1,281	1.00 (0.000)	Reference
0.1-22.9	6,140	59,708	616	0.80 (-0.223)	(0.72, 0.88)
23.0-45.9	8,072	77,757	998	0.90 (-0.105)	(0.82, 0.98)
46.0-68.9	7,085	68,622	765	0.95 (-0.051)	(0.86, 1.04)
≥ 69.0	3,337	32,076	440	1.32 (0.278)	(1.18, 1.48)

Table 1: Relative risks of death from all causes as well as alcohol intake among men in the JACC study in Japan (for details, see Lin et al. [1]).

	Coffee consumption (cups/day)	No. of individuals	No. of cases	Adjusted RR (log value)	95% CI
Bidel et al., [2]	0-2	644	35	1.00 (0.000)	Reference
	3-4	1,041	54	0.79 (-0.236)	(0.51, 1.22)
	5-6	1,356	69	0.66 (-0.416)	(0.43, 1.03)
	≥ 7	796	52	0.94 (-0.062)	(0.58, 1.52)
Grobbee et al. [3]	None	7,592	8	1.00 (0.000)	Reference
	≤ 1	13,048	23	0.58 (-0.545)	(0.25, 1.36)
	2-3	16,009	18	0.68 (-0.386)	(0.36, 1.31)
	≥ 4	8,940	5	0.48 (-0.734)	(0.18, 1.31)

Table 2: Characteristics of two studies of coffee consumption and stroke that were included in a meta-analysis [4].

Exposure x	No. of individuals	No. of cases	logRR	SE (logRR)
u_0-v_0	N_0	a_0	0.0	Reference
u_1-v_1	N_1	a_1	y_1	s_1
⋮	⋮	⋮	⋮	⋮
u_m-v_m	N_m	a_m	y_m	s_m
Total	N	A		

Table 3: Grouped exposure intervals and summarized response data in a cohort study.

exposure levels relative to that of 0, even if the level of reference category for reported RRs was not 0. We provide some numerical illustrations and comparisons based on simulations with typical assignments to determine the effects of exposure level assignments.

Methods

Likelihood based assignment of levels for grouped exposure intervals

In some cohort studies with N individuals, the data on exposure x_i of each individual i is summarized in a table of grouped exposure interval $x_i \in I_j = (u_j, v_j)$ and the corresponding summarized response data with the frequency N_j, the number of cases a_j, the log value of the relative risk y_j and its standard error s_j for each interval I_j, j=0,1,…,m (Table 3), where the lower endpoint u_0 is known but the upper endpoint v_m is unknown. N_j is given as the number of individuals, or the person-time, and y_j is sometimes shown as the adjusted values of some covariates. s_j can also be estimated from the confidence interval of the risks if the standard error of the logRR is not reported [14]. In case-control studies, however, the number of controls b_j is also given and y_j is the log value of the odds ratio (OR).

In typical methods, the midpoints are used as the assigned exposure levels for each interval I_j, (j=1,2,…,m) (here after midpoints assignment). On the other hand, we assign the exposure level based on the summarized data according to the procedure proposed by Takahashi and Tango [6] as follows. We assume that the exposure levels of all individuals in the study are a set of random variables $(x_1, x_2, \ldots, x_N)$ and that a power transformation of the exposure, x_i^{λ}, is obtained from a common normal distribution $N(\mu,\sigma^2)$ with the mean μ and variance σ^2, where $x_i^0 = \log x_i$ for λ=0. Given this assumption, the frequency N_j provides a log-likelihood based on a binomial distribution for the distribution of x=x_i. The unknown parameters λ, μ, σ^2 and v_m can be estimated by maximizing the log-likelihood. Based on the estimated distribution of x=x_i with the probability density function f (x) such as $x^{\lambda} \sim N(\mu,\sigma^2)$, the assigned exposure level d_j for the jth interval I_j=(u_j,v_j) is calculated as the mean of its truncated distribution:

$$d_j = \frac{\int_{I_j} x f(x)dx}{\int_{I_j} f(x)dx}$$

(j=0,1,…,m) (hereafter likelihood-based assignment).

Nonlinear association modeling using cubic splines

Splines are smooth functions that can assume virtually any shape, and the most useful type of spline is generally a cubic spline function, which is restricted to be smooth at the junction of each cubic polynomial [12]. In epidemiological studies, a restricted cubic spline model has been often applied to nonlinear dose-response data. As noted by Larsson and Orsini [4], a restricted cubic spline with three knots was recently applied to a potential nonlinear association that was depicted as a J-shaped curve. However, in some studies, such as that in Table 2, the exposure level of the reference group is not assigned on x=0. Thus, in this paper, we consider a cubic spline model for the log relative risk on the exposure x, logRR(x), that satisfies logRR(d_0)=0 using assigned exposure levels d_0 for the reference interval I_0, as follows:

$$\log \mathrm{RR}(x) = y = (-\beta_1 d_0 - \beta_2 d_0^*) + \beta_1 x + \beta_2 x^* = \beta_1 (x - d_0) + \beta_2 (x - d_0^*) \qquad (1)$$

where $x^* = (x-k_2)_+^3 - \gamma(x-k_1)_+^3 - (1-\gamma)(x-k_3)_+^3$,

$d_0^* = (d_0-k_2)_+^3 - \gamma(d_0-k_1)_+^3 - (1-\gamma)(d_0-k_3)_+^3$, $\gamma = (k_3-k_2)/(k_3-k_1)$ with fixed knots $k_1 < k_2 < k_3$, and $(x-a)_+^3 = \max\{0,(x-a)^3\}$.

First, we construct an approximate covariance estimate for the adjusted log relative risks from a fitted table that conforms to the values proposed by Greenland and Longnecker [15], and we construct a variance-covariance matrix $\boldsymbol{\Sigma}$. In this step, we assume the assigned exposure levels d_j and standard errors s_j for each interval I_j to be fixed. The coefficients $\boldsymbol{b}$ of the restricted cubic spline model (1) are estimated using generalized least-squares regression with $\boldsymbol{\Sigma}$, i.e.,

$$\hat{\boldsymbol{b}} = (\hat{\beta}_1, \hat{\beta}_2)' = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$$

and the estimated variance-covariance matrix of $\hat{\boldsymbol{b}}$,

$$V(\hat{\boldsymbol{b}}) = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$$

where $\boldsymbol{A}$' and $\boldsymbol{A}^{-1}$ imply the transpose and inverse matrices of $\boldsymbol{A}$, and

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ m \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} d_1 - d_0 & d_1^* - d_0^* \\ d_2 - d_0 & d_2^* - d_0^* \\ \vdots & \vdots \\ d_m - d_0 & d_m^* - d_0^* \end{pmatrix}$$

respectively. Note that, if we need to estimate a spline model for the log relative risk on x relative to x=0, logRR$_0$(x), when the reported y implies that relative to the reference x=d_0, we can determine the spline model as

$$\log \mathrm{RR}_0(x) = \hat{\beta}_1 x + \hat{\beta}_2 x^* \qquad (2)$$

using the same $\hat{\boldsymbol{b}}$ estimated from (1), because $y = \log(p_x / p_{d_0}) = \log(p_x/p_0) + \log(p_x / p_{d_0})$ and $\log(p_x / p_{d_0}) = -\hat{\beta}_1 d_0 - \hat{\beta}_2 d_0^*$ where p_a is the probability of being a case with an exposure of $x=a$.

A crucial problem in spline regression is knot placement [10]. One simple approach is to have the observations x_i determine the positions of the knots [16]. Some studies such as those by Larsson and Orsini [4], Harrell et al. [12] and Orsini et al. [13], placed knots at fixed percentiles in the data. Therefore, we examine a procedure here for selecting the positions of three knots among the assigned exposure levels d_j (j=1,2,…,m) with the exception of d_0 of the reference group I_0. Thus, we assign the knots $k_1=d_1$, $k_2=d_2$, $k_3=d_3$ when m=3. If m>3, we require a procedure that is based on likelihood. Under the normal assumption of generalized linear regression, $\boldsymbol{y} \sim N_m(\boldsymbol{Xb}, \boldsymbol{\Sigma})$, we can derive the log-likelihood function of $\boldsymbol{b}$ and evaluate the likelihood of each resulting model for several candidate knot positions. Thus, the putatively better knots k_1, k_2, and k_3 are placed in a position that maximizes

$$l(\hat{\boldsymbol{b}} | k_1, k_2, k_3) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}})\Sigma^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}) \qquad (3)$$

Application

First for the data in Table 1 with person-years for N_j, the likelihood-based assignment procedure assigned the exposure levels as d_j=0, 14.26, 34.20, 56.09, and 86.41 g/day with the estimated parameters λ=0.5 and ν_m=139.90. When using the number of individuals for N_j, they are assigned as d_j=0, 14.27, 34.21, 56.09, and 86.41 g/day with the estimated parameters λ=0.5 and ν_m=140.00. Thus, they differed from the midpoints assignment of d_j=0, 11.5, 34.45, 57.45, and 82.8 (as 1.2 times the lower boundary for the open-ended upper category) or 80.45 (as assuming the open-ended upper category has the same amplitude as the adjacent category) g/day for each I_j. By using the exposure levels of d_j=0, 14.26, 34.20, 56.09, and 86.41 g/day, likelihood procedure (3) selected k_1=14.26, k_2=34.20, and k_3=86.41 as the knots, and the coefficients were estimated as β_1=-4.88×10^{-3} and β_2=-5.46×10^{-6}, respectively, while the midpoints assignment with d_4=80.45 estimated the coefficients as β_1=-2.51×10^{-3} and β_2=-1.36×10^{-5} with k_1=11.5, k_2=57.45 and k_3=80.45 (Figure 1).

Next for the data in Table 2, the exposure levels were assigned by the likelihood-based assignment as d_j=1.65, 3.56, 5.44, and 7.94 cups/day with λ=2/3 and ν_m=14.4 in Bidel et al. [2] where the intervals were assumed as $0 \le x < 2.5$, $2.5 \le x < 4.5$, $4.5 \le x < 6.5$, and $6.5 \le x$, respectively.

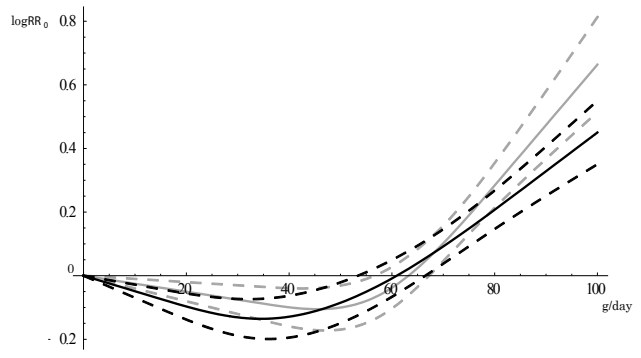

Figure 1: Fitted curve for the log of the adjusted relative risk logRR$_0$ of all-cause mortality associated with alcohol intake, as reported by Lin et al. [1] in Table 1 (black line: the likelihood-based assignment; gray line: the midpoints assignment). Dashed lines represent the 95% confidence intervals based on the asymptotic normal theory.

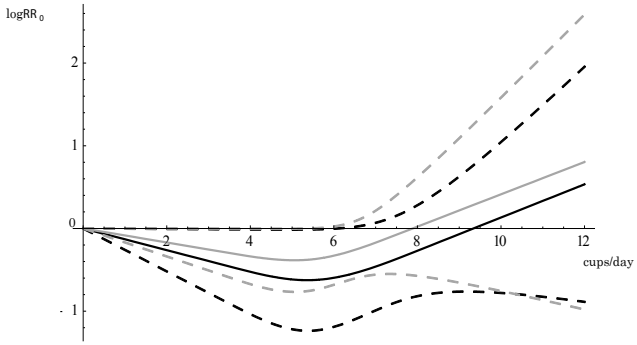

Figure 2: Fitted curve for the log of the adjusted relative risk logRR$_0$ of strokes associated with coffee consumption relative to the reference consumption level of zero in Bidel et al. [2] in Table 2 (black line: the likelihood-based assignment; gray line: the midpoints assignment with d_0=0). Dashed lines represent the 95% confidence intervals based on the asymptotic normal theory.

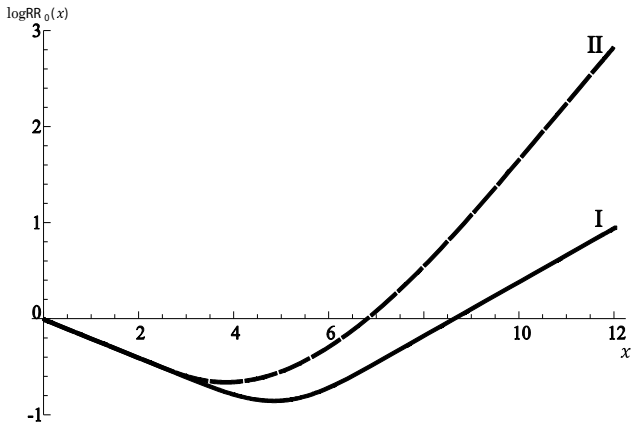

Figure 3: Cubic spline curves logRR$_0$(x)=-0.2x - 0.05x* for the true model with knots (k_1, k_2, k_3) of (I) (3, 5, 7) and (II) (2.3, 3.2, 10.5)).

k_1=3.56, k_2=5.44, and k_3=7.94 were set as the knots, the coefficients were estimated as β_1=-0.131 and β_1=-0.024, respectively, and we can determine the estimated curve of logRR$_0$ from the model (2) (Figure 2). On the other hand, the midpoints assignment with d_0=0.0 estimated β_1=-0.085 and β_1=-0.023.

Simulation

In this section, we discuss the simulation studies conducted to assess our proposed procedure, wherein we used cubic spline regression for the nonlinear association with the likelihood-based assignments in grouped exposure intervals. Two cubic spline curves

$$y = \log \mathrm{RR}_0(x) = \log \frac{p_x}{p_0} = -0.2x - 0.05x^* \qquad (4)$$

with three knots (k_1, k_2, k_3)=(I) (3, 5, 7) and (II) (2.3, 3.2, 10.5) were assumed to be the true model for the association between exposure x and the log relative risk y in a cohort study, respectively, where p_x and p_0 are the probabilities of being a case with the exposure x and x=0, respectively (Figure 3). We considered interval for grouping, $0 \le x < 2$, $2 \le x < 4$, $4 \le x < 6$, and $x \ge 6$, such as that the number of knots in each interval I_j was at most one for curve I, whereas the interval $2 \le x < 4$ had two knots, i.e., k_1=2.3 and k_2=3.2, for curve II.

We set the population size N=2,000 and the probability p_0=0.05 for the reference x=0. We generated a set of exposures $\boldsymbol{x}=\{x_1, x_2,\ldots,x_{2,000}\}$ from a truncated normal distribution of N(3.5, 8.0) with the interval $0 \le x \le 12$. By calculating p_{x_i} using (4), we generated a set of 1 or 0 Bernoulli random numbers $\boldsymbol{w}=\{w_1, w_2,\ldots,w_{2,000}\}$ using $\Pr\{W_i=1\}=p_{x_i}$ for each x_i, where the sample was counted as a case when w_i=1, and we summarize the generated data$\{\boldsymbol{x},\boldsymbol{w}\}$ in Table 4. Note that each relative risk was calculated relative to the reference group $0 \le x<2$.

We compared the following four procedures:

(i) The likelihood-based assignment and the three knots were set at them: (d_0, d_1, d_2, d_3)=(1.22, 3.0, 4.91, 7.96) .

(ii) The midpoints assignment with d_0=1.0 and the three knots were set at them: (d_0, d_1, d_2, d_3)=(1.0, 3.0, 5.0, 7.0).

(iii) The midpoints assignment with d_0=0.0 and the three knots were set at them: (d_0, d_1, d_2, d_3)=(0.0, 3.0, 5.0, 7.0).

(iv) The likelihood-based assignment (d_0, d_1, d_2, d_3)=(1.22, 3.0, 4.91, 7.96), but the three knots were fixed on the "true knots" (k_1, k_2, k_3)=(3.0, 5.0, 7.0) for true curve I and (k_1, k_2, k_3)=(2.3, 3.2, 10.5) for true curve II.

Procedure (i) is our proposed method, and procedures (ii) and (iii) are typical methods that use midpoints, where the highest interval was assigned by assuming that the boundary had the same amplitude as the adjacent category [4]. In addition, we compared the proposed method with procedure (iv), which has true knots. Note that for true curve I, the positions of the three knots in curves according to procedures (ii) and (iii), (k_1, k_2, k_3)=(3.0, 5.0, 7.0), were in the same positions as the true knots.

We generated B=1,000 sets of $\boldsymbol{w}_1, \boldsymbol{w}_2,\ldots,\boldsymbol{w}_B$ for the fixed $\boldsymbol{x}$. The curves were estimated using procedures (i-iv) for each set, and we estimated $\log RR_0(x)$ as $\hat{y}_b(x)$ (b=1,2,…,B) for each point of x=1,2,…,12. Comparing with the values derived from the true model (4), Tables 5 and 6 show the bias $\mathrm{Bias}\left(\hat{y}(x)\right)=\left(\frac{1}{B}\right)\sum_{b=1}^{B}\hat{y}_b(x)-y(x)$, the mean squared error (MSE) $\mathrm{MSE}\left(\hat{y}(x)\right)=\left(\frac{1}{B}\right)\sum_{b=1}^{B}\left(\hat{y}_b(x)-y(x)\right)^2$, and the coverage probability CP(x) for a 95% confidence interval of B=1,000 sets for each x=1,2,…,12 based on the fitted curves produced by each procedure using the summarized data for true curves I and II, respectively. The means and standard deviations of estimated coefficients $\hat{\boldsymbol{b}}=(\hat{\beta}_1,\hat{\beta}_2)$ are also shown as E($\hat{\boldsymbol{b}}$) and SD($\hat{\boldsymbol{b}}$) in the tables. Note that the accuracy of the fitted curves cannot be measured based only on $\hat{\beta}$, because $\hat{\beta}$ and the accuracy must be affected directly by the positions of the knots. Figure 4 shows the curves joining the mean values $\left(\frac{1}{B}\right)\sum_{b=1}^{B}\hat{y}_b(x)$.

First, procedure (iv), which used the likelihood-based assignment and the true knots, had a small bias and small MSE values in each scenario. It also gave closer coverage probabilities to 0.95, and E($\hat{b}$) was close to true $\boldsymbol{b}$. Overall the curve produced using this procedure had a good fit to the true curve.

x	No. of individuals	No. of cases	log RR	SE (log RR)
0-2	426	a_0	0.000	Reference
2-4	606	a_1	y_1	s_1
4-6	558	a_2	y_2	s_2
≥ 6	410	a_3	y_3	s_3
Total	2,000	A		

Table 4: Summarized data for $\{\boldsymbol{x},\boldsymbol{w}\}$.

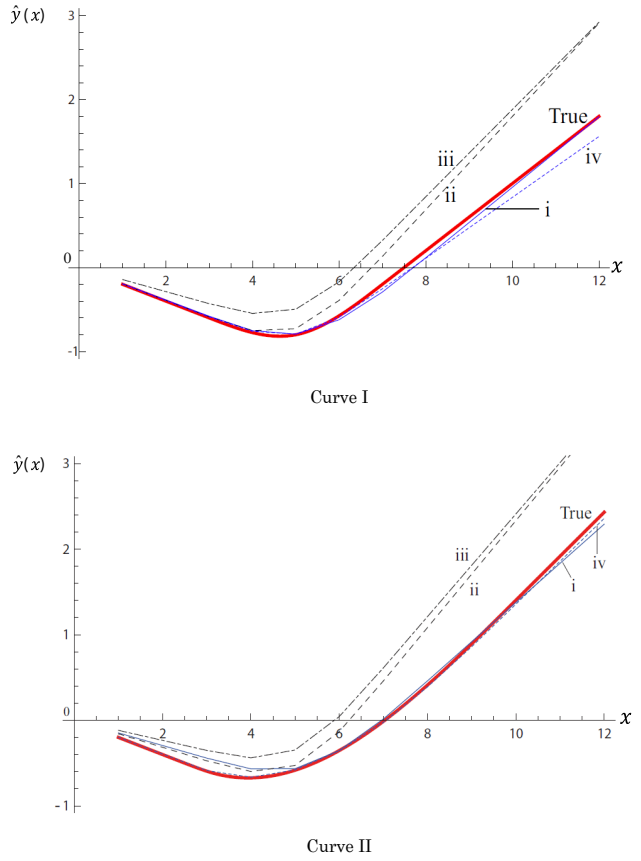

Figure 4: Comparisons of curves joining the mean values of $\hat{y}(x)$ with x=1,2,…,12 for true curves I and II.

Next, the proposed procedure (i), which used the likelihood-based assignment and the selected knots, also had a small bias, small MSE values for each x. In particular, when $x \ge 9$ for true curve I, it was shown that the curve produced using this procedure had a better fit to the true curve than that produced using procedure (iv). When x<5 for true curve II, this procedure had a slightly higher bias than procedure (iv). However, the coverage probabilities remained higher than 90%.

Procedure (ii), which used the midpoints assignment, was very similar to procedure (i) when x<5 for curves I and II. When x>5, the bias and MSE increased gradually, so this procedure could not deliver stable estimates with a large x.

Procedure (iii), which used the midpoints assignment and d_0=0, estimated the value slightly higher with a low x compared with the other procedures. With a high x, it showed similar behavior to procedure (ii).

Discussion

In this paper, we proposed a procedure for assessing the nonlinear association between exposure levels and the risk of disease by assigning exposure levels to grouped exposure intervals. In particular, we focused on a restricted cubic spline model for J-shaped dose-response curves. The procedure can be applied to the log relative risks when they are given relative to the reference point x=0 and also to the interval $x \in I_0$.

x	1	2	3	4	5	6	7	8	9	10	11	12
True	-0.20	-0.40	-0.60	-0.775	-0.80	-0.575	-0.20	0.20	0.60	1.00	1.40	1.80
Procedure (i) $E(\hat{\boldsymbol{b}}) = (-0.193, -0.035)$, $SD(\hat{\boldsymbol{b}}) = (0.114, 0.014)$												
Bias($\hat{y}$ (x))	0.007	0.014	0.021	0.024	0.007	-0.047	-0.092	-0.084	-0.064	-0.045	-0.025	-0.005
MSE($\hat{y}$ (x))	0.013	0.052	0.117	0.201	0.256	0.244	0.200	0.164	0.168	0.219	0.314	0.456
CP($\hat{y}$ (x))	0.941	0.941	0.941	0.940	0.941	0.924	0.905	0.921	0.969	0.980	0.986	0.988
Procedure (ii) $E(\hat{\boldsymbol{b}}) = (-0.196, -0.063)$, $SD(\hat{\boldsymbol{b}}) = (0.109, 0.024)$												
Bias($\hat{y}$ (x))	0.004	0.009	0.013	0.024	0.072	0.184	0.333	0.488	0.644	0.799	0.954	1.110
MSE($\hat{y}$ (x))	0.012	0.048	0.108	0.183	0.221	0.207	0.248	0.409	0.693	1.102	1.635	2.291
CP($\hat{y}$ (x))	0.949	0.949	0.949	0.948	0.935	0.894	0.848	0.860	0.894	0.922	0.930	0.937
Procedure (iii) $E(\hat{\boldsymbol{b}}) = (-0.143, -0.055)$, $SD(\hat{\boldsymbol{b}}) = (0.081, 0.020)$												
Bias($\hat{y}$ (x))	0.057	0.113	0.170	0.229	0.304	0.405	0.523	0.642	0.762	0.882	1.002	1.121
MSE($\hat{y}$ (x))	0.010	0.039	0.088	0.153	0.209	0.258	0.367	0.570	0.867	1.258	1.744	2.324
CP($\hat{y}$ (x))	0.882	0.882	0.882	0.868	0.818	0.650	0.623	0.784	0.852	0.902	0.920	0.935
Procedure (iv) $E(\hat{\boldsymbol{b}}) = (-0.196, -0.047)$, $SD(\hat{\boldsymbol{b}}) = (0.115, 0.019)$												
Bias($\hat{y}$ (x))	0.004	0.009	0.013	0.016	0.008	-0.017	-0.053	-0.090	-0.127	-0.165	-0.202	-0.239
MSE($\hat{y}$ (x))	0.013	0.053	0.119	0.203	0.256	0.232	0.185	0.166	0.182	0.234	0.322	0.445
CP($\hat{y}$ (x))	0.941	0.941	0.941	0.941	0.941	0.925	0.908	0.920	0.958	0.969	0.981	0.983

Table 5: Comparisons of the estimated values for $\log RR_0(x)$ on x=1,2,…,12, which were derived from the curves using procedure (i-iv) for Curve I, y=-0.2x–0.05x* with knots (3, 5, 7).

x	1	2	3	4	5	6	7	8	9	10	11	12
True	-0.20	-0.40	-0.594	-0.673	-0.583	-0.354	-0.013	0.413	0.894	1.403	1.917	2.431
Procedure (i) $E(\hat{\boldsymbol{b}}) = (-0.147, -0.035)$, $SD(\hat{\boldsymbol{b}}) = (0.115, 0.019)$												
Bias($\hat{y}$ (x))	0.053	0.106	0.153	0.105	0.018	0.001	0.028	0.049	0.025	-0.027	-0.083	-0.139
MSE($\hat{y}$ (x))	0.015	0.059	0.131	0.195	0.237	0.224	0.178	0.143	0.141	0.178	0.259	0.385
CP($\hat{y}$ (x))	0.916	0.916	0.916	0.936	0.939	0.927	0.915	0.935	0.975	0.987	0.990	0.989
Procedure (ii) $E(\hat{\boldsymbol{b}}) = (-0.196, -0.063)$, $SD(\hat{\boldsymbol{b}}) = (0.109, 0.024)$												
Bias($\hat{y}$ (x))	0.042	0.085	0.121	0.074	0.055	0.222	0.473	0.672	0.815	0.930	1.040	1.150
MSE($\hat{y}$ (x))	0.013	0.051	0.114	0.175	0.204	0.211	0.347	0.594	0.889	1.245	1.662	2.177
CP($\hat{y}$ (x))	0.925	0.925	0.926	0.946	0.940	0.880	0.716	0.716	0.812	0.883	0.914	0.931
Procedure (iii) $E(\hat{\boldsymbol{b}}) = (-0.143, -0.055)$, $SD(\hat{\boldsymbol{b}}) = (0.081, 0.020)$												
Bias($\hat{y}$ (x))	0.083	0.166	0.243	0.233	0.236	0.397	0.625	0.800	0.918	1.008	1.094	1.179
MSE($\hat{y}$ (x))	0.013	0.052	0.115	0.148	0.165	0.245	0.473	0.767	1.071	1.400	1.791	2.251
CP($\hat{y}$ (x))	0.795	0.795	0.797	0.865	0.859	0.650	0.436	0.611	0.765	0.864	0.906	0.927
Procedure (iv) $E(\hat{\boldsymbol{b}}) = (-0.196, -0.047)$, $SD(\hat{\boldsymbol{b}}) = (0.115, 0.019)$												
Bias($\hat{y}$ (x))	0.003	0.005	0.008	0.008	0.004	-0.002	-0.011	-0.022	-0.034	-0.047	-0.059	-0.072
MSE($\hat{y}$ (x))	0.016	0.065	0.145	0.216	0.242	0.226	0.187	0.152	0.142	0.173	0.250	0.374
CP($\hat{y}$ (x))	0.950	0.950	0.950	0.947	0.937	0.923	0.912	0.931	0.970	0.985	0.992	0.994

Table 6: Comparisons of the estimated values for $\log RR_0(x)$ on x=1,2,…,12, which were derived from the curves using procedure (i-iv) for Curve II, y=-0.2x–0.05x* with knots (2.3, 3.2, 10.5).

Our simulation results showed that the estimated curve was sensitive to the assignment, and the likelihood-based assignment could estimate the nonlinear association accurately. It showed that the proposed procedure using the likelihood-based assignment had a lower bias when used for estimation compared with other procedures that used the midpoints assignment.

One of the applications to estimate a dose-response association from summarized data is that it should be applicable to a meta-analysis. In general, the exposure categories were different in the

studies so they should not be combined. Furthermore, in some of the meta-analysis studies described in Table 2, the reference category was also different, so it was inappropriate to combine them directly. Some methods have been discussed for meta-analysis to obtain the pooled estimate without estimating the association within individual studies. For example, Greenland and Longnecker [15] described the pool-first method for meta-analysis of trend involving data pooling before trend analysis. On the other hand, Rota et al. [11] proposed a random-effects meta-regression model for nonlinear dose-response relationship fitting second-order fractional polynomial models, where the two-step procedures requires initially fitting second-order fractional polynomial models within each study, and then pooling the study-specific two trend components. They tried to fit a pool-first method on the data of a small number of studies, and they obtained the identical results achieved by using their random-effects approach. It may also be possible to estimate pooled curves using our proposed procedure by a multivariate random effect meta-analysis [17], as discussed in Larsson and Orsini [4]. However, it has to be noted that both pool-first method and two-step procedure use pre-assigned exposure levels for grouped exposure intervals, and they assigned values by the midpoints assignment. Thus, the likelihood-based assignment could give different results in the pooled estimates. Moreover, in situations such as those for assessment of publication or other availability bias by the funnel plot, it is important to accurately estimate individually.

In the procedure reported herein, we fixed three knots in the J-shaped curve. The choice of the location of the knots is a crucial problem and the estimate of the curve was sensitive to their positions. Although we showed simulation results only for a situation of m=3 here, we also examined a procedure where the choice was based on likelihood for $m>3$, and the results of simulation studies could show that it produced well-fitted curves. In a similar manner to the procedure proposed herein, we can apply restricted cubic spline regression using other numbers of fixed knots to produce a more flexible curve shape. However, the choice of the number of knots is generally a crucial problem in spline regression. In this situation, the model with the best fit can be selected using a similar procedure by evaluating Akaike's Information Criterion (AIC), which is a penalized likelihood that takes into account the number of parameters estimated in the model based on likelihood (3). Also non-cubic spline models have been discussed in epidemiological studies. Further discussions of such models, including evaluations of different methods, are required in the future.

Our work, moreover, could be located in the errors-in-variable field, which aims to correct for bias that arises if measurement error in x is ignored. The statistical approaches developed in those fields might be applied in the situation discussed in this paper. We would also like to leave such a study including comparisons with the proposed procedure here, in our future work.

References

1. Lin Y, Kikuchi S, Tamakoshi A, Wakai K, Kawamura T, et al. (2005) Alcohol consumption and mortality among middle-aged and elderly Japanese men and women. Ann Epidemiol 15: 590-597.
2. Bidel S, Hu G, Qiao Q, Jousilahti P, Antikainen R, et al. (2006) Coffee consumption and risk of total and cardiovascular mortality among patients with type 2 diabetes. Diabetologia 49: 2618-2626.
3. Grobbee DE, Rimm EB, Giovannucci E, Colditz G, Stampfer M, et al. (1990) Coffee, caffeine, and cardiovascular disease in men. N Engl J Med 323: 1026-1032.
4. Larsson SC, Orsini N (2011) Coffee consumption and risk of stroke: a dose-response meta-analysis of prospective studies. Am J Epidemiol 174: 993-1001.
5. Berlin JA, Longnecker MP, Greenland S (1993) Meta-analysis of epidemiologic dose-response data. Epidemiology 4: 218-228.
6. Takahashi K, Tango T (2010) Assignment of grouped exposure levels for trend estimation in a regression analysis of summarized data. Stat Med 29: 2605-2616.
7. Corrao G, Rubbiati L, Bagnardi V, Zambon A, Poikolainen K (2000) Alcohol and coronary heart disease: a meta-analysis. Addiction 95: 1505-1523.
8. Di Castelnuovo A, Costanzo S, Bagnardi V, Donati MB, Iacoviello L, et al. (2006) Alcohol dosing and total mortality in men and women: an updated meta-analysis of 34 prospective studies. Arch Intern Med 499: 2437-2445.
9. Greenland S (1995) Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology 6: 356-365.
10. Bagnardi V, Zambon A, Quatto P, Corrao G (2004) Flexible meta-regression functions for modeling aggregate dose-response data, with an application to alcohol and mortality. Am J Epidemiol 159: 1077-1086.
11. Rota M, Bellocco R, Scotti L, Tramacere I, Jenab M, et al. (2010) Random-effects meta-regression models for studying nonlinear dose-response relationship, with an application to alcohol and esophageal squamous cell carcinoma. Stat Med 29: 2679-2687.
12. Harrell FE Jr, Lee KL, Pollock BG (1988) Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst 80: 1198-1202.
13. Orsini N, Li R, Wolk A, Khudyakov P, Spiegelman D (2012) Meta-analysis for linear and nonlinear dose-response relations: examples, an evaluation of approximations, and software. Am J Epidemiol 175: 66-73.
14. Greenland S (1987) Quantitative methods in the review of epidemiologic literature. Epidemiol Rev 9: 1-30.
15. Greenland S, Longnecker MP (1992) Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. Am J Epidemiol 135: 1301-1309.
16. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. (2ndedn), Springer, NewYork, USA.
17. Jackson D, White IR, Thompson SG (2010) Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. Stat Med 29: 1282-1297.