Cadmium Intake and Age in β_2 -Microglobulinuria: Categorical Data Analysis in Epidemiology

Key words: Cadmium— β_2 -Microglobulinuria—Data analysis.

While β_2 -Microglobulinuria (B-m-uria) is a sensitive index to cadmium (Cd) exposure, there is evidence that it is also related to the aging process.¹⁻³⁾ The literature further indicates that the effect of age on B-m-uria is seen only after age 70.⁴⁻⁶⁾ Indeed, Kjellström⁷⁾ declares that the increase of B-m-uria with age is so great that “its use in screening for cadmium poisoning is compromised.”

A recent report, using data from the Kakehasi River basin, Ishikawa Prefecture,⁸⁾ completely disregarded this aging effect and attributed the occurrence of B-m-uria solely to total lifetime Cd intake. It also presented a total intake of 2,000 mg as the threshold to cause a B-m-uria. Scrutiny of the data, however, strongly suggests the existence of an aging effect on the incidence of B-m-uria. Furthermore, I consider there are serious methodological problems in the analysis. The purpose of this letter is to discuss those problems and suggest an alternative method of analysis when epidemiological data are presented in a format of a contingency table.

1. Age and Prevalence of B-m-uria

According to the Kakehashi report⁸⁾ Table 2 and Table 3 which include the basic material for the analysis, at every level of Cd concentration in rice, a sudden jump in the prevalence of B-m-uria is observed after 70 years of exposure. For males in all but the most heavily polluted villages, the only groups showing higher prevalence than the control group are those exposed for longer than 70 years regardless of the length of exposure or the level of pollution. One must recognize that to be exposed to Cd contamination for more than 70 years, a subject must also be more than 70 years of age. According to the Kakehashi data, the average age of those who were exposed for more than 70 years are 76.6 years for males and 76.7 years for females. Note, however, that the average ages below 70 years of exposure are all around 60 years of age. Thus, the groups whose prevalence of B-m-uria are significantly higher than the control group are characterized not only by long exposure of more than 70 years but also by ages of 70 years and above.

An examination of Table 4 of the Kakehashi report makes me suspect that the age effect on B-m-uria may be even more important than that of total Cd intake. Based on that table they have presented a linear regression between the preva-

Table 1 Prevalence of B-m-uria among inhabitants with Cd intake 3,000–4,000 mg.

Cd in rice ($\mu\text{g/g}$)	Average age	Total Cd intake	N	B-m-uria (%)	
				/liter	/g cr.
Male					
0. 10–0. 29	76. 6	2998	83	20. 5	32. 5
0. 30–0. 49	59. 9	3016	34	0. 0	2. 9
	59. 4	3837	189	9. 5	11. 1
0. 50–0. 69	59. 9	3850	22	22. 7	27. 3
Female					
0. 10–0. 29	76. 7	2998	92	35. 9	51. 1
0. 30–0. 49	60. 3	2890	63	7. 9	12. 7
	61. 4	3926	162	12. 3	22. 2
0. 50–0. 69	60. 3	3666	48	6. 3	20. 8

Source: Table 4 of Nogawa et al.⁸⁾

lence of B-m-uria and the Cd intake in lifetime. For a similar level of intake, however, if the average age is more than 70 years, an abrupt rise takes place in the prevalence. For instance, Table 1 of this letter is an excerpt from Table 4 of the Kakehashi report showing the lifetime intake level of 3,000–4000 mg. It is clear that at around the level of 3,000 mg intake the B-m-uria prevalence jumps only if the subjects' average age exceeds 70 years. The scatter diagrams shown in the report⁸⁾ do show upward trends of B-m-uria prevalence with Cd intake. But that appears to be due to a large extent to the three points representing the groups of over 70 years. Take out these three points, and the regression coefficients are reduced by about one-half. For instance, B-m-uria by g creatinine is reduced from 0.0083 to 0.0046 in males and from 0.012 to 0.0065 in females. In fact, this reduced coefficient in males is not statistically different from 0.

2. Linear Regression and Categorical Data

The main technique to analyze the Kakehashi data was the simple regression model, a statistical method used for studying the relationship between measurable variables. But neither the response variable nor the explanatory variable in the data is a measurable variable. Let us first discuss the explanatory variable, Cd intake. The data set on which the analysis is based is in essence arranged in the form of a contingency table with the exposure time in four categories, the degree of pollution in three categories, and the presence of B-m-uria in two categories. Then what is observed is the frequency of persons falling in each cell of the $4 \times 3 \times 2$ table. Although never observed, the authors introduced total Cd intake as the explanatory variable, indirectly deriving from the average level of Cd concentration in rice, average length of residence, and average age of persons for each cell of the table. Besides, these represent exposure, not intake. Nevertheless, I will set aside the problem of heterogeneity in Cd intake for a given exposure. (Tsuchiya et al.³⁾ show a wide variation in the Cd concentration in either feces or urine among the individuals living in the same Cd-polluted area and among Cd

alloy workers.)

It should be also pointed out that a single intake value given for each cell actually represents a wide range of intakes, even under the assumption that a given exposure produces a fixed level of intake. For instance, consider persons belonging to the cell of 0.10–0.29 $\mu\text{g/g}$ Cd concentration in rice and 0–29 years of residence. There can be more than a twofold difference between the maximum and minimum intakes. (Using the authors' algorithm, the intake of a 50-year-old person living in a village of 0.10 $\mu\text{g/g}$ concentration for five years is 944 mg, and that of an 80-year-old man living in a village of 0.29 $\mu\text{g/g}$ concentration in rice for 28 years is 2,285 mg.) On the other hand, we can easily show persons with similar amounts of intake falling into different cells in the table, as seen in Table 1.

Let us now turn to the response variable, which is a proportion, its value ranging from 0 to 1. (The real response variable is dichotomous—positive or negative B-m-uria.) For a proportion, simple regression analysis is not generally recommended; the standard method currently practiced is logistic regression. A proportion violates the basic assumptions in simple regression—normality and homoscedasticity. Furthermore, the extrapolation of a regression line leads to the impossible values of a proportion—below zero and above unity. In the case of the Kakehashi report, the regression line cuts across the B-m-uria level of the control group at about 2,000 mg intake. Thus, the authors considered the point of intersection as the threshold intake. As the extrapolation of a straight regression line will eventually cut across the base line, there has to be a threshold dose in any linear regression. But how do we know linear regression is the appropriate model to describe the relation between the prevalence of B-m-uria and Cd intake? It has been shown that for a moderate level of response rate many different models, whether liner, logit, or probit regression, produce similar results. It is the low dose extrapolation that gives entirely different results depending on the statistical model.⁹⁾ Had the logit model been applied to the data, there would have been no threshold; only when the intake is zero, the response rate is zero, assuming no background response.

The regression equations for the Kakehashi data are obtained from unweighted grouped data. Each cell carried the same weight—whether the cell contained 10 subjects or 263 subjects did not matter. (The authors appeared to realize this point, as a cell with 4 persons was eliminated from the analysis.) They seem to possess the individual measurements on age, length of residence, etc. Then I do not understand why they have resorted to grouped data. Had individual data been used, the sample points would have been 878 instead of 11 in male and 972 instead of 12 in female. Also multiple regression, whether logistic or linear, could have been used. (If multiple linear regression was the method of choice, the response variable should be the urinary β_2 -Microglobulin concentration, a continuous variable, rather than positive or negative B-m-uria.)

Table 2 Estimates of logit effects for Table 2⁸⁾: Male (for odds of B-m-uria positive vs. negative)

Effect	Full table	Excluding 70-yr exposure
Constant	-2.048	-2.476
Exposure time (E)		
0-29	-0.754	-0.328
30-49	-0.520	-0.104
50-69	0.012	0.432
70-	1.262**	—
Pollution in rice (P)		
0.10-0.29	-0.570**	-0.562**
0.30-0.49	-1.94	-0.288
0.50-0.69	0.766**	0.850**

** p<0.01

3. Logit Analysis for the Contingency Table

The Kakehashi data are in cross-classified contingency tables. That is, the original data are not in measurable variables but in categorical variables. For analysis of categorical data, the log-linear model has proven to be a powerful technique for quantitatively assessing the relationship among the variables. By treating one of the variables as the response variable, the log-linear model may be easily converted to logit regression. For further discussion of the model, one should refer to a standard textbook, such as Fienberg.¹⁰⁾

The results of log-linear analysis to the Kakehashi data are presented in Table 2. Since the purpose of this analysis is in illustration, only one data set (male in Table 2⁸⁾) has been analyzed. Although the authors incorporated age in the calculation of Cd intake, the data were actually composed of three variables: level of pollution (P), exposure time (E), and presence of B-m-uria (M). The data were collapsed over age. Thus, there are only two possible factors affecting the presence of B-m-uria. The fit of the log-linear model [EP] [EM] [PM] is satisfactory ($p=0.27$). The estimated logit effects of the factors are shown in the column "full table" in Table 2.

The overall odds of positive B-m-uria are given by the exponential of the constant term, $e^{-2.048}=0.129$. The odds are modified by the logit effects as specified in the model. For instance, the expected odds of having B-m-uria for a person who lived in the village of Cd in rice 0.10-0.29 $\mu\text{g/g}$ for more than 70 years are $\exp^{-2.048-0.570+1.262}=0.258$, that is, the expected probability of B-m-uria is $0.258/(1+0.258)=0.025$, which happens to be a perfect fit. However, it should be noted that among the individual effects of categories of E, only one of them, that for 70-years of exposure, is significant. As indicated earlier, this category is confounded with age. If the group having longer than 70 years of exposure is eliminated from the data, the logit effects of E are no longer significant, as seen in the last column of Table 2. The difference in the chi-square values between models [EP] [PM] [EM] and [EP] [PM] is barely 2.86 for the degrees of

freedom 2, warranting the elimination of the effect of E on M. As the last category of E is confounded with age and other levels of E do not affect the odds of M, the significant effect of the longest exposure period may, in fact, originate from age rather than exposure time.

4. Hypothetical Data Showing the Variable Age

To determine whether or not the variable age (A) is a significant factor, the data must also be cross-classified by categories of A, in addition to E, P, and M. Unfortunately, the age distribution within each cell of the table is not given. For the purpose of illustration, I have created a hypothetical table as shown in Table 3. The original data (Table 1),⁸⁾ show four categories of ages, but I have lumped together the last two age groups in my hypothetical data. Although hypothetical, the table collapsed over variable A exactly matches Table 2 of the Kakehashi report. Each cell frequency of Table 3 is actually created by prorating, using the age and sex distribution of subjects shown in the original data. For instance, in the cell of the lowest pollution level and shortest length of exposure, there are 32 males. In my hypothetical data I have allotted 17 to the age group of 50-59, 13 to age group of 60-69, and 2 to the oldest group. The algorithm is simple. Since the last group of E consists of 162 persons aged 70 and above only, there are $878-162=716$ persons in the first three categories of E. Of these, 32 are in the first category of E. Table 1 of the Kakehashi report shows that 382 men are in the 50-59 age group and 285 men are of ages 60-69. Thus our imputed number of men falling to the cell of E=1, P=1, and A=1 is $382 \times (32/716)=17$, and the number to the cell of E=1, P=1, and A=2 is $285 \times (32/716)=13$. I have applied arbitrary proportions to these imputed numbers in allotting the positive cases of B-m-uria. But there were two conditions: older groups generally have higher positive rates than younger groups within a given level of P and E, and the marginals are consistent with the original data.

With this hypothetically created data, three marginal tables can be formed.

Table 3. Hypothetical data to include the age of subjects for Table 2⁸⁾ male.

Rice Cd conc.		0, 10-0, 29			0, 30-0, 49			0, 50-0, 69		
Age		50-59	60-69	70-	50-59	60-69	70-	50-59	60-69	70-
Length of residence										
0-29	-	16	13	2	5	3	1	3	1	0
	+	1	0	0	0	1	0	0	0	0
30-49	-	23	17	3	18	15	1	9	6	2
	+	1	1	0	0	0	0	2	2	1
50-69	-	133	96	16	95	66	10	52	34	5
	+	7	8	3	6	9	3	11	13	2
70-	-	0	0	66	0	0	28	0	0	21
	+	0	0	17	0	0	13	0	0	17
Total		181	135	107	124	94	56	77	56	48

- : B-m-uria negative; + : B-m-uria positive

Table 4 Prevalence of B-m-uria by collapsing over one variable in Table 3.

I. Prevalence of B-m-uria by age and average Cd concentration in rice, as obtained by collapsing over length of residence for the hypothetical data of Table 3.

Age	50-59		60-69		70-		Total	
Av. Age ^a	54.0		64.0		76.6		62.7	
Length of residence	52.7		52.7		70.9		57.0	
	N	%	N	%	N	%	N	%
Cd in rice								
0.10-0.29	181	5.0	135	6.7	107	18.7	423	9.0
0.30-0.49	124	4.8	94	10.6	56	28.6	274	11.7
0.50-0.69	77	16.9	56	26.8	48	41.7	181	26.5
Total	382	7.3	285	11.9	211	26.5	878	13.4

II. Prevalence of B-m-uria by age and residence time in Cd-polluted area, as obtained by collapsing over average Cd concentration for the hypothetical data of Table 3.

Length of residence	0-29		30-49		50-69		70-		Total	
Cd conc.	0.33		0.34		0.33		0.61			
	N	%	N	%	N	%	N	%	N	%
Age										
50-59	25	4.0	53	5.7	304	7.9	0	0	382	7.3
60-69	18	5.6	41	7.3	226	13.3	0	0	285	11.9
70-	3	0.0	7	14.3	39	20.5	162	29.0	211	26.5
Total	46	4.3	101	6.9	569	10.9	162	29.0	878	13.4

^a In computing averages it is assumed that the average age of 50-59 is 54 years and 60-69 years is 64 years.

One of them, obtained by collapsing over the variable A perfectly agrees with Table 2 of the Kakehashi report. In Table 4, I have presented the other two marginal tables, by collapsing over E and P, respectively. Each table provides its own feature. When collapsed over A, the conclusion could be that the occurrence of B-m-uria "increased with increasing Cd concentration in rice and time of residence in a Cd-polluted area."⁸⁾ when collapsed over E, however, we can say that the incidence increased with age and Cd concentration in rice, and when collapsed over P, we can say that the incidence increased with age and time of residence. In statistical jargon these reflect marginal associations, as data are collapsed over a third variable. It has been shown that the marginal association can be misleading (Simpson's paradox).^{11,12)}

For illustration I have applied log-linear analysis to the hypothetical data. Some possible log-linear models for the data are shown in Table 5. By theory, the interaction among the explanatory variables [APE] must be included in all models. The importance of an effect may be determined by the difference of chi-square values between two competing models. Then Model #3, [APE] [AM] [PM], is the one best fit to the "data". The conclusion may be that age

Table 5. Goodness-of-fit of some log-linear models for Table 3.

Model #	Defining set	Degrees of freedom	G ²	P	Parameter under test	ΔG^2
1	[APE] [AM] [PM] [EM]	22	6.82	0.99		
2	[APE] [AM] [EM]	24	37.42	0.04	[PM]	31.40**
3	[APE] [AM] [PM]	25	7.30	0.99	[EM]	1.28
4	[APE] [EM] [PM]	24	17.44	0.83	[AM]	11.42**
5	[APE] [AM]	27	38.41	0.07	[PM]	31.11**

A: age; P: pollution in rice; E: length of exposure; M: B-m-uria

** p<0.01

and level of pollution significantly affect the odds of positive B-m-uria but length of residence does not. I repeat that Table 3 is not the real data. Consequently, the analysis does not reflect the real situation. But I would like to emphasize that if the Kakehashi data were rearranged to include age categories and proper analytic techniques were applied, different conclusions might have been reached.

REFERENCES

- 1) Kitamura M, Koizumi N. Results of urine tests for bedridden old persons in Hyogo Prefecture (Report 2). Kankyo Hoken Rep (Japanese Public Health Association) 1975; 36: 135-7 (In Japanese).
- 2) Nomiyama K, Nomiyama H, Yotoryama M, Taguchi T. Some recent studies on the renal effects of cadmium. In: Cadmium 77. Proc 1st Int Cadmium Conference, San Francisco. London: Metal Bull Ltd, 1978; 186-94.
- 3) Tsuchiya K, Iwao S, Sugita M, Sakurai H. Increased urinary β_2 -Microglobulin in cadmium exposure: dose-effect relationship and biological significance of β_2 -Microglobulin. Environ Health Perspect 1979; 28: 147-53.
- 4) Kjellström T, Piscator M. Quantitative analysis of β_2 -Microglobulin in urine as an indicator of renal tubular damage induced by cadmium. Phadedoc No. 1 Sweden: Uppsala, 1977; 3-21.
- 5) Kowal NE, Kraemer DF. Urinary cadmium and beta-2-microglobulin levels of persons aged 20-74 years from a subsample of the national HANES II survey, 1976-1980. in: Cadmium 81, Proc 3rd Int Cadmium Conf, Miami. London: Cadmium Assoc, 1982; 119-22.
- 6) Nomiyama K, Yotoryama M, Nomiyama H. Urinary beta-2-microglobulin and renal function in elderly people in an area with no known cadmium pollution (Japan). Arch Environ Contam Toxicol. 1983; 12: 143-6.
- 7) Kjellström T. Renal effects. In: Friberg L, Elinder C-G, Kjellström T, Nordberg GF, eds. Cadmium and health: a toxicological and epidemiological appraisal. Vol II. Effects and response. Boca Raton: CRC Press, 1986; 70.
- 8) Nogawa K, Honda R, Kido T, Tsuritani I, Yamada Y, Ishizaki M, Yamada H. A dose-response analysis of cadmium in the general environment with special reference to total cadmium intake limit. Environ Res 1989; 48: 7-16.
- 9) Ashton WD. The logit transformation. New York: Hafner Pub Co. 1972.
- 10) Fienberg SE. The analysis of cross-classified categorical data. 2nd ed. Cambridge, MA: MIT Press. 1980.
- 11) Simpson EH. The interpretation of interaction in contingency tables. J Royal Statist Soc B. 1951; 13: 238-41.
- 12) Blyth CR. On Simpson's paradox and the sure-thing principle. J Am Statist Assoc 1972; 67: 364-6.

*University of Hawaii, School of Public Health
1960 East-West Rd, Honolulu, HI 97822, USA*

Chai Bin PARK

(Received February 12, 1991 and in revised form March 4, 1991)