Interaction and Its Solution in Individual Matching Case-control Study

XIAO-XIN HE AND SHUI-GAO JIN

Center for Public Health Surveillance and Information Services, Chinese Center for Disease Prevention and Control, 27 Nan Wei Road, Beijing 100050, China

Objective To indicate the deficiency of the classical method for analyzing data on individual matching case-control study in consideration of the interaction between the study factor (exposure) and the matching factor, and to find out a proper method for handling this deficiency. **Method** First, experimental data with 50 pairs of cases and controls were used for strata analysis according to the values of a matching factor to illustrate the possible interaction between a risk factor (exposure) and the matching factor. Second, a detailed procedure was proposed for analyzing such data. **Results** Interaction between the study factor and matching factor was demonstrated by using strata analysis and unconditional logistic regression analysis. Therefore the results from the classical analysis for such data might be incorrect. **Conclusion** Data from individual matching case-control study design should be dealt with strata analysis or multivariate analysis to explore and evaluate the possible interaction between the study factor and matching factor. The conclusion would be valid only after such analysis is conducted.

Key words: Interaction; Individual matching case-control study; Stratification analysis; Multivariate analysis

INTRODUCTION

The purpose of individual matching study design is to control confounders. Controls individually matched with cases, influence of the confounding factors on both the cases and controls would be the same[1, 2]. In addition to its confounding effect, other risk factors might exert effect modification on the study factor, which could not be dealt with by the classical analysis procedure for individual matching case-control study design. Since interaction between risk factors is interesting to all researchers, and it is necessary to find out timely and evaluate it accurately for disease prevention and control. Unfortunately, the classical analysis procedure for the data from individual matching case-control study design does not check the interaction between the study factor and matching factor at all. That is its defect that needs to be improved.

MATERIALS AND METHODS

Without losing generalization, supposing that a 1:1 individual matching case-control study was designed to explore the etiological relationship between disease D and risk factor

Biographical note of the first author: Xiao-Xin HE, male, born in 1968, Ph. D., majoring in epidemiology and health statistics.

0895-3988/2003
CN 11-2816

A1. For simplicity, it was supposed that there were only two risk factors for disease onset: A1 and A2, with each having only two levels (values, 0 for unexposed and 1 exposed). In order to control the impact of A2, every control was matched with a case by the value of A2. Fifty matching sets (100 individuals) were sampled, and information about A1 and A2 exposure history was collected and analyzed.

In this example, the classical method was to be used to illustrate the defect of the method for dealing with such data firstly, and then a new procedure would be proposed for the analysis of data from such design.

RESULTS

Classical Procedure for Analyzing Data From Individual Matching Case-control Design

Table 1-1 is the classical 2 by 2 tables, which summaries the results from a 1:1 matching design classified by risk factor A1. Totally 50 pairs were included in this study. And Table 1-2 summaries the crude odds-ratio (OR_c) and its confidential intervals estimated from Table 1-1.

TABLE 1-1

A Classical 1:1 Case-control Study

Case	Control		Total
	1	0	
1	5	30	35
0	12	3	15
Total	17	33	50

TABLE 1-2

OR_c Estimation

Parameter	Point Estimates	95% Confidence Interval	
		Lower	Upper
Adjusted *OR* (MH)	2.5000	1.2464	5.0145 (R)
Adjusted *OR* (MLE)	2.5000	1.2981	5.0624 (M)

Note. R: RGB; M: mid-P.

A conclusion could be made based on the results when the influence of A2 was ignored, A1 would be a risk factor of *D*. The relationship strength (crude odds ratio) was expressed as OR_c=2.5000 with a 95% CI of (1.2464, 5.0145).

Defect of the Above Results and Strata Analysis

As mentioned previously, both A1 and A2 had two levels (values). If the above table is separated according to the value of A2, we would see different results (Tables 2-1, 2-2, 3-1, and 3-2).

Tables 2-1 and 2-2 are the results when A2=0, and Tables 3-1 and 3-2 for A2=1.

TABLE 2-1

Association of A1 With D1 for A2=0

Exposed Cases	Exposed Controls		Total
	1	0	
1	2	10	12
0	9	1	10
Total	11	11	22

TABLE 2-2

Parameter Estimation[a] (A2=0)

Parameter	Point Estimation	95% Interval Confidence	
		Lower	Upper
Adjusted *OR* (MH)	1.1110	0.4013,	3.0765 (R)[b]
Adjusted *OR* (MLE)	1.1111	0.4418,	2.8274 (M)[b]

Note. [a]*P*=5000; [b] R: RGB; M: mid-P.

TABLE 3-1

Association of A1 With D1 for A2=1

Exposed Cases	Exposed Controls		Total
	1	0	
1	3	20	23
0	3	2	5
Total	6	22	28

TABLE 3-2

Parameter Estimation[a] (A2=1)

Parameter	Point Estimation	95% Interval Confidence	
		Lower	Upper
Adjusted *OR* (MH)	6.6667	2.0182	22.0216 (R)[b]
Adjusted *OR* (MLE)	6.6667	2.1720	28.1748 (M)[b]

Note. [a]*P*=0.0001; [b] R: RGB; M: mid-P.

From the above tables, we could conclude that there is no statistically significant relationship between A1 (exposure) and *D* (disease) when A2=0 (OR_0=1.1110, with a 95%CI of (0.4013, 3.0765)). But we do find that the relatinship between the exposure of variable A1 and disease (D) was of statistical difference when A2=1. In the second case, the Odds ratio (OR_1) was equal to 6.6667 with a 95%CI between 2.0182 and 22.0216 (*P*=0.0001).

From an epidemiological view of point, A2 could be called a confounder of A1. Then the differences between Tables 1-1, 2-1, and 3-1 could be explained. As we know, there are two types of relationship between the two factors: independent from, or correlated with each other. If the two factors are independent from each other, the analysis like table 1-1 would be correct. Otherwise, the above procedure and its results would conceal the truth and produce a pseudo correct conclusion.

How should we deal with such data correctly? First, the data should be stratified according to the values of matching factor A2. Association between the exposure variable A1 and D should be analyzed in each stratum with stratum OR_is estimated. Second, homogeneity of stratum OR_is should be tested to judge the presence of interaction. If there was no interaction on the one hand, the classical analyzing procedure would be able to control confounding of A2, so the conclusion is correct. If there was interaction between A1 and A2 on the other hand, we can conclude that the above results are pseudo. At that moment, we should give the results in accordance with the criteria for interaction.

Test for the Homogeneity of Stratum OR_is

The stratum OR_is must be tested for homogeneity to check the presence of interaction between the two factors. We gave out the formulae for the test on the basis of that for group matching case-control study data[2].

The formulae for the summarized Odds Ratio (OR_w) and the Variance of log of stratum OR_i ($Var(\ln OR_i)$) were

$$OR_w = e^{\sum\left(W_i \times \ln OR_i \Big/ \sum W_i\right)},$$

$$Var(\ln OR_i) = 1 \Big/ W_i,$$

where $W_i = \dfrac{1}{\dfrac{1}{b_i} + \dfrac{1}{c_i}}$.

The formula for homogeneity test (χ^2_{k-1}) of the stratum OR_i s was

$$\chi^2_{k-1} = \sum \frac{(\ln OR_i - \ln OR_w)^2}{Var(\ln OR_i)},\ \nu = k-1.$$

If there was no interaction between A1 and A2, the stratified OR_is should be synthesized into one OR_w. The OR_w aslo should be tested for statistical significance with the following formulae:

$$OR_w = e^{\sum\left(W_i \times \ln OR_i \Big/ \sum W_i\right)}.$$

$$\begin{aligned}\chi^2_w &= \left(\sum W_i\right) \cdot (\ln OR_w - \ln OR_0)^2, \\ &= \left(\sum W_i\right) \cdot (\ln OR_w - \ln 1)^2, \\ &= \left(\sum W_i\right) \cdot (\ln OR_w)^2,\end{aligned}$$

there $W_i = \dfrac{1}{\dfrac{1}{b_i} + \dfrac{1}{c_i}}$.

Theoretically, OR_w and χ^2_w should be identical with the crude OR_c and χ^2_c.

Contrarily, if the interaction between A1 and A2 existed, the stratified specific conclusion, or the results from multivariate analysis results should be given out. In that case, the summary OR_w and χ_w^2 needed not to be calculated, as it was incorrect.

The above analyzing procedure for homogeneity test of stratum OR_is could be summarized in the following Table (Table 4).

TABLE 4

Test for Homogeneity of Stratum OR_is in 1:1 Individual Matching Case-control Study

Stratum	OR_i	$\ln OR_i$	W_i	$W_i \cdot \ln OR_i / \sum W_i$	OR_w	$Var(\ln OR_i)$	$\frac{(\ln OR_i - \ln OR_w)^2}{Var(\ln OR_i)}$
A2=0	1.1111	0.1054	4.7368	0.0679	2.0995	0.2111	1.9180
A2=1	6.6667	1.8971	2.6087	0.6737		0.3833	3.4827
Total	/	/	7.3455	0.7417	/	0.5944	5.4007

Because χ_{k-1}^2=5.4007, =k-1, P<0.05, we couclude that there is statistically significant difference between the strtified odds ratios (OR_1 and OR_2). In other words, they are drawn from different population, and the interaction between A1 and A2 exists. So the stratum OR_is can not be synthesized, and the crude OR_c and Chi-square are illogical.

As the interaction between A1 and A2 existed, there were two ways to report the results:

Reporting the stratum specific results, including the interaction between A1 and A2; when A2=0, there was no relation between A1 and *D*. In other words, *OR*=1; when A2=1, A1 was significantly associated with *D*, *OR*=6.6667, with a 95% confidence interval of (2.0182, 22.0216). That is to say, the interaction between A1 and A2 made A1 become a significant risk factor from non-risk one. (from *OR*=1.1111 to *OR*=6.6667).

Multivariate analysis. Unconditional logistic regression analysis could be used with *D* as an outcome variable, A1, A2 and interactive item A1∗A2 as the explanatory variables, to explore the impact of A1, A2 and A1∗A2 on *D*.

Unconditional Logistic Regression Analysis

Judging the presence of interaction. Different models would be compared with each other to estimate the impact of interactive items A1∗A2 on the relationship between A1, A2 and *D*.

Model 1: D as an outcome variable and A1 as an explanatory variable:

$$LogitP1=-0.7885+1.5106A1,$$
$$-2\ln L_1=125.3500.$$

Model 2: D as an outcome variable, A1 and A1∗A2 as explanatory variables

Based on the above model, an interaction term would be added and the second model to be established:

$$LogitP2=-0.7885+0.8755A1+1.2567(A1*A2),$$
$$-2\ln L_2=121.0349,$$
$$Q=-2\ln(L_1/L_2)=125.3500-121.0349=4.3151,\ df=1,\ P<0.05.$$

Therefore the interactive item A1∗A2 had an significant impact on *D*.

$$OR\ (A1=1:\ A1=0)=\exp.\ (0.8755+1.2567A2).$$

When A2=0, *OR* (A1=1:A1=0)=exp. 0.8755=2.4001. When A2=1, *OR* (A1=1:A1=0)= exp. (0.8755+1.2567)=8.4334. A2 imposed a significant influence on the relation between A1 and *D*.

Model 3: D as an outcome variable, A2 as an explanatory variable:

$$LogitP_3=0,$$
$$-2\ln L_3=138.6294.$$

Model 4: D as an outcome variable, A2 and A1∗A2 as explanatory variables:

$$LogitP_4=-1.4856A2+2.8253\ (A1*A2),$$
$$-2\ln L_4=116.4412,$$
$$Q=-2\ln\ (L_3/L_4)=138.6294-116.4412=22.1882,\ df=1,\ P<0.05.$$

Therefore, the interactive item A1∗A2 had an significant impact on *D*.

$$\text{OR (A2=1:A2=0)=exp. (-1.4856+A1).}$$

When A1=0, *OR* (A2=1:A2=0)= exp. (-1.4856)=0.2264. When A2=1, *OR* (A1=1:A1=0) = exp. (-1.4856+2.8253)=3.8179

A1 imposed a significant influence on the relation between A2 and *D*.

From the above analysis, it was obvious that there did exist interaction between A1 and A2. And the interactive term A1∗A2 exerted impact both on the association of *D* with A1, and on that of *D* with A2.

Unconditional Logistic Regression Analysis

Unconditional logistic regression analysis was to be carried out with *D* as outcome variable, A1, A2 and interactive item A1∗A2 as the explanatory variables, to explore the impact of A1, A2 and interactive item A1∗A2 on *D*.

$$LogitP=-0.0953+0.1823A1-1.3863A2+2.6430(A1*A2).$$

TABLE 5

Significant Test for Explanatory Variables

Term	Coefficient	$\bar{x}\pm s$	Z statistics	*P* value
A1	0.1823	0.6043	0.3017	0.7629
A1∗A2	2.6430	0.9060	2.9174	0.0035
A2	-1.3863	0.6606	-2.0986	0.0359
Constant	-0.0953	0.4369	-0.2181	0.8273

Based on Table 5, we can conclude that, the interaction does exist between A1 and A2.

$$OR\ (A1=1:\ A1=0\mid A2=0)=\exp.\ (0.1823)=1.2000,$$
$$OR\ (A1=1:\ A1=0\mid A2=1)=\exp.\ (0.1823+2.6430)=16.8660,$$
$$OR\ (A2=1:\ A2=0\mid A1=0)=\exp.\ (-1.3863)=0.2500,$$
$$OR\ (A2=1:\ A2=0\mid A1=1)=\exp.\ (-1.3863+2.6430)=3.5138.$$

CONCLUSION

As to individual matching case-control design, data analysis should not be carried out with the classical methods, which might ignore the interaction between the matching factor and factors interested, and reach a seemingly correct but really biased result. The proper procedure for analyzing such data should be stratified according to the value of matching factor first, and followed by calculating stratified OR_is, and testing the homogeneity of stratifieds OR_is to make the adjustiment on the presence of interaction between the matching factor and studying factors. If the interaction does not exist, the crude association or synthesized one could be reported. Otherwise, the association of the studying factor with the disease should be reported accoding to the level of the matching factor, or the results from multivariate analysis.

REFERENCES

1. Wang, Tiangen (1994). Case-control study. In *Epidemiology* (3rd ed., Lian Zhihao Eds.), pp.66-88. Beijing: People' s Hygiene Press.
2. Wu, Zhenglai (1994). Case-control study. In *Methods and Application of Modern Epidemiology*, (1st ed., Zeng Guang Eds.), pp.83-95. Beijing: Beijing Medical University and Chinese Union Medical University Union Press.

(*Received January* 27, 2002 *Accepted September* 15, 2002)