예 방 의 학 회 지 : 제 23 권 제 3 호
Korean J. of Preventive Medicine
Vol. 23, No. 3, September 1990

Comparison of Control Methods for Estimation Bias in Unmatched Analysis of Matched Data

Department of Preventive Medicine, Seoul National University
College of Medicine, Seoul 110-460, Korea

Keun-Young Yoo, M. D.

= 국문초록 =

짝을 이룬 자료분석시 야기되는 Estimation Bias의 Control Methods

유 근 영

서울대학교 의과대학 예방의학교실

짝짓기 방법은 교란변수를 통제하기 가장 좋은 방법으로 알려져 있으나, 모수추정시 그 계산방법이 복잡하고, 포함된 모든 정보를 이용할 수 없다는 단점을 갖고 있다. 그럼에도 불구하고, conditional 모델을 이용한 matched 분석법은 짝지은 자료 분석시 가장 좋은 방법으로 인정되고 있다. 그러나 명확한 confounding 현상을 통제할 목적이 아닌 상태에서 짝지워진 자료를 matched 분석법으로 모수추정하는 경우나, 올바로 짝지워진 자료를 분석법의 편이성 때문에 unmatched 분석을 시도하는 경우, 오히려 estimation bias가 야기될 수 있다. 이러한 estimation bias의 통제능력을 몇 가지 분석방법을 이용하여 비교하고자, 1 : 2로 대응된 한 환자-대조군 자료를 이용하여 Mantel-Haenszel 분석법, 두 가지의 unconditional model을 이용한 다변량분석법의 결과를 conditional model을 이용한 matched 분석법의 결과와 비교하였다.

1. Matched 분석법의 대용방법으로 사용된 세 가지 방법들은 모수추정면에서나 가설검정능력면에서 차이를 서로 보이지 않았다.
2. 짝짓기에 사용된 변수가 분석자료내에서 confounder나 effect modifier로 작용되지 않았음이 명백한 경우에는 이들 세 가지 통제 방법과 matched 분석법간에 차이가 없었다.
3. 짝짓기에 사용된 변수가 분석자료내에서 effect modifier로 작용하지는 않았으나, Confounder로 작용한 것으로 추정되는 경우, unmatched 분석법으로 인해 야기된 estimation bias의 통제능력이 이들 세 가지 대용방안 모두에서 인정되었다.
4. 짝짓기에 사용된 변수가 분석자료내에서 effect modifier로 작용하고 있음을 직접 확인할 수 있는 경우에는, overmatching에 의한 estimation bias를 의심할 수 있었으며, 이들 세 가지 통제방법은 오히려 unmatched 분석 방법에 가까운 모수를 추정하였다.

Key Words: *matching, unmatched analysis, confounder, effect modifier, stratified analysis, unconditional logistic model, conditional maximum likelihood estimate.*

Ⅰ. INTRODUCTION

Among the control methods for confoundings in epidemiologic studies, matching is preferred in the design or even in the analytic stage, because of its statistical advantage in better adjustment for confounding effect (Kleinbaum *et al.*, 1982). If matching has been done to control for apparent confounding effects, then, matched analysis with conditional likelihood method will be regarded as a method of choice for the most valid estimation of the odds ratio. Otherwise, unmatched analysis of a matched data may result in a biased estimator. In general, estimates from unmatched analysis of matched data is known to have a tendency to be biased towards the null value ; towards the unity for the odds ratio (Breslow and Day, 1980 ; Schlesselman, 1982 ; Rothman, 1986 ; Kelsey *et al.*, 1986). The bias between unmatched and matched analysis in matched pairs (Feinstein, 1987), and in matched triples (Yoo *et al.*, 1991) have been discussed with quantitative demonstration.

However, one may use unmatched analysis after pooling the matched data, if matching has been introduced merely for the convenience of sampling or if there has been no apparent reason for matching in the designing stage (Schlesselman, 1982 ; Feinstein, 1987). If the matching was unnecessary (overmatching) or when there was effect modification, instead of confounding, then, matched analysis, on the contrary, may bring about another bias.

This paper will demonstrate ignoring effect of matching in a data, in which two controls were matched for simple demographic reason. Then, various control methods for the bias arising from unmatched analysis of matched data will be compared. These will include Mantel-Haenszel's common odds ratio, adjusted odds ratios by two different types of unconditional linear logistic models, with a reference to matched odds ratio by conditional maximum likelihood method. From these results, appropriateness of various methods in matched data analysis will be discussed in terms of confounder or effect modifier in the data.

Ⅱ. MATERIALS AND METHODS

Data for the demonstration of ignoring effect of matching were drawn from a case-control study (Janerich *et al.*, 1991), which was designed to test the effect of birth characteristics on testicular cancer. They selected 413 males from a data file in the Connecticut Tumor Registry, U. S. A. All of them were Connecticut-born, and diagnosed as testicular cancer in 1935~1985. Two controls were matched per a case, among those who meet the following criteria ; male birth, the same race, and the birth year within the same year as the case. Reason for matching in this study was not to control for any known confounders, but just for the demographic convenience of control selection. Information on birth characteristics was abstracted from the birth certificate in the State.

Any triple with at least one missing observation among the case and two controls was deleted in the analysis of the variable, in order to avoid another source of bias. For the illustrative purpose, each variable has been dichotomized ; NFAGE (or NMAGE) for the age of father (or mother) at birth under versus over 35 (or 33), NFBYR (or NMBYR) for the birth year of father (or mother) before versus after 1930, BTHSTAT for fullterm versus prematurity, and BTHTYP for singleton versus plurality. Race (RACE) was a dichotomous matching variable (white / black). Another matching variable, birth year of the case, was divided into six strata of 10-year interval (QBTHYR).

For the demonstration of ignoring effect of matching, every case and control groups in the triples were separated into a pool of unmatched arrangement. Meanwhile, the original data set in a fully matched arrangement was kept as a reference.

The maximum likelihood estimate of the unmatched odds ratio (Cornfield, 1951) was calculated to observe ignoring effect of matching in the data. The logit method was used for 95% confidence bounds of the unmatched odds ratio (Woolf, 1955). Calculation of each parameter in the unmatched arrangement was carried out by the

PC-SAS (SAS Institute, 1987).

As a reference for comparison, the conditional maximum likelihood estimate of the matched odds ratio was measured by a linear logistic regression model that can be used to fit the fully matched data set (Holford *et al.*, 1978). The model was logit $P_i=B_i^*(X_{i0}-X_{i1})$, here X_{i0} for the case and X_{i1} for the control. The GLIM program was modified for analysis of matched triple data (The GLIM Working Party, 1987). The 95% confidence bounds of matched odds ratio were calculated by the test-based method (Miettinen, 1976), and the test of hypothesis was done by the likelihood ratio test (Breslow and Day, 1980). Every conditional procedure was done in a univariate setting by the GLIM system.

Among the various stratified analytic methods, the Mantel Haenszel procedure for a common odds ratio over strata was applied to control the estimation bias in unmatched analysis ; $\Omega_{mh}=[\Sigma\ a_id_i/n_i]/[\Sigma\ b_ic_i/n_i]$ (Mantel and Haenszel, 1959). Matching variables (RACE and QBTHYR) were adjusted for the estimation of a summary odds ratio. The Mantel-Haenszel odds ratio and its test-based confidence intervals (Miettinen, 1976) were measured by the the PC-SAS. In order to test the hypothesis that the odds ratios from each strata are all equal, the Breslow and Day's test for homogeneity (Breslow and Day, 1980) was done by the PC-SAS.

Multivariate analysis for unconditional maximum likelihood estimate was applied, using the usual linear logistic model. Two different types of models were built. First, 〈Multivariate Ⅰ〉, two matching variables (RACE and QBTHYR) were introduced separately in an unconditional linear logistic model (Breslow and Day, 1980) ; logit $P_i=\alpha_0+\alpha_1(\text{RACE})+\alpha_2(\text{QBTHYR})+\beta_iX_i$. Secondly, 〈Multivariate Ⅱ〉, an indicator variable for each of 12 subgroups of matching variables (2 for RACE × 6 for QBTHYR) was included in the model (Schlesselman, 1982) ; logit $P_i=\alpha_0+\alpha_i$ (RACE and QBTHYR)$+\beta_iX_i$. Since multivariate procedure was based on listwise deletion technique, every subjects who had missing information on any variable being considered in the model were deleted from that analysis. These modelling procedures were carried on by the GLIM system. Its 95% confidence interval was calculated by the logit methods (Woolf, 1955), and test of hypothesis was done by the likelihood ratio test (Breslow and Day, 1980).

Ⅲ. RESULTS

1. Estimation bias between unmatched and matched odds ratio ; ignoring effect of matching variable

As shown in Table 1, there were absolutely no bias due to ignoring effect of matching in NFAGE and NMAGE. It implies that it was unnecessary to use matching procedure in the design stage to control for the potential confounding in at least these two variables. However, unmatched odds ratios of NFBYR, NMBYR and BTHSTAT showed apparent underestimation of the matched odds ratios towards the unity, which seems to be due to inappropriate use of analytic method. Unmatched odds ratio of BTHTYP was slightly overestimated, on the contrary to the general concept of estimation bias towards the null. In spite of under- or overestimation, change in statistical significance or 95% confidence limits were not noticeable.
not noticeable.

2. Demonstration of the source of estimation bias ; potential confounder or effect modifier

In order to demonstrate the source of estimation bias arising from unmatched analysis of matched data, each unmatched odds ratio was partitioned into stratum-specific odds ratios in unmatched arrangement (Table 2). This procedure is a useful way to find out effect modification in a data. During the procedure, another matching variable, RACE, was not used, because of rarity of the black in the data.

The stratum-specific odds ratios of NFAGE and NMAGE were neither varied within the strata of the matching variable (QBTHYR), nor different from the unmatched odds ratio. By definition (Kleinbaum *et al.*, 1982 ; Kelsey *et al.*, 1986), the matching variable did not modify the

Table 1. Estimation bias arising from unmatched analysis of fully matched data showing ignoring effect of matching in a matched case-control study

Methods		NFAGE	NMAGE	NFBYR	NMBYR	BTHSTAT	BTHTYP
Unmatched	$_{U}$OR	1.17	1.18	0.80	0.68*	2.66*	6.14*
analysis	l-CI	0.90	0.88	0.57	0.50	1.53	1.65
	u-CI	1.53	1.57	1.11	0.92	4.63	22.84
	No.	1,221	1,236	1,221	1,236	954	978
Matched	$_{M}$OR	1.17	1.18	0.64	0.41*	3.12*	6.00*
analysis	l-CI	0.90	0.88	0.39	0.29	1.67	1.63
	u-CI	1.52	1.56	1.02	0.70	5.63	22.16
	No.	407	412	407	412	318	326

$_{U}$OR : unmatched odds ratio estimated by unconditional maximum likelihood method
$_{M}$OR : matched odds ratio estimated by a linear logistic regression model to fit triple-matched data for a conditional maximum likelihood estimte ; logit $P_i=\beta_1^*(X_{i0}-X_{i1})$
l-CI : 95% lower confidence limit calculated by logit methods for $_{U}$OR, and by test-based method for $_{M}$OR
u-CI : 95% upper confidence limit calculated by logit methods for $_{U}$OR, and by test-based method for $_{M}$OR

Table 2. Demonstration of effect modification of QBTHYR on the risk of testicular cancer associated with BTHSTAT

Unmatched	Strata of birth year of cases					
OR	-1919	1920-29	1930-39	1940-49	1950-59	1960-76
NFAGE						
1.17	1.00	1.16	1.44	1.19	0.96	1.64
(1221)	(84)	(135)	(132)	(315)	(384)	(171)
NMAGE						
1.18	0.90	1.06	0.65	1.07	1.34	2.11
(1236)	(84)	(135)	(132)	(324)	(384)	(177)
NFBYR						
0.80	–	–	–	–	0.80	0.56
(1221)	(84)	(135)	(132)	(315)	(384)	(171)
NMBYR						
0.68	–	–	–	0.39	0.40	0.60
(1236)	(84)	(135)	(132)	(324)	(384)	(177)
BTHSTAT						
2.66	–	–	2.90	1.12	3.92	3.57
(954)	(0)	(0)	(99)	(309)	(381)	(165)
BTHTYP						
6.14	–	–	4.19	–	–	6.27
(978)	(0)	(0)	(99)	(318)	(384)	(177)

(Number of observation)

effect of the disease-risk association nor confounding effect in the data. In contrast, odds ratios of NFBYR and NMBYR were observed to vary slightly within the strata. Although most of the cells were missed due to the high collinearity of QBTHYR with NFBYR (or NMBYR), it can be inferred that there might be little effect modification in NFBYR and NMBYR. However, BTHSTAT and BTHTYP showed the most marked variation in the stratum-specific odds ratios, suggesting that modifying role of the matching variable on the risk of testicular cancer associated with prematurity and plurality seems to be apparent. The source of bias in BTHTYP was not so clear.

3. Comparison of various control methods for the estimation bias

Table 3 shows results of comparison among various control methods for the estimation bias in matched data analysis. As a whole, these three control methods did not show any difference in parameter estimation, nor in statistical significance. In terms of control for the estimation bias, however, interesting findings were observed.

As expected, adjusted estimators of NFAGE and NMAGE were almost identical with the value of matched odds ratio. However, adjusted odds ratios of NFBYR and NMBYR have been shifted much closer to the matched

Table 3. Comparison of various control methods for estimation bias arising from unmatched analysis of matched triples

Methods		NFAGE	NMAGE	NFBYR	NMBYR	BTHSTAT	BTHTYP
Stratified	$_{C}OR$	1.17	1.18	0.70	0.50	2.65	6.20
	l-CI	0.90	0.88	0.46	0.33	1.55	1.94
	u-CI	1.53	1.57	1.06	0.75	4.54	19.76
	No.	407	412	407	412	318	326
Multivariate	$_{A}OR_1$	1.17	1.18	0.69	0.50	2.66	6.22
(Ⅰ)	l-CI	0.90	0.88	0.45	0.33	1.53	1.67
	u-CI	1.53	1.57	1.06	0.75	4.64	23.19
	No.	407	412	407	412	318	326
Multivariate	$_{A}OR_2$	1.17	1.18	0.69	0.50	2.66	6.22
(Ⅱ)	l-CI	0.90	0.88	0.45	0.33	1.53	1.67
	u-CI	1.53	1.57	1.06	0.75	4.64	23.21
	No.	407	412	407	412	318	326
Matched	$_{M}OR_2$	1.17	1.18	0.64	0.41	3.12	6.00
	l-CI	0.90	0.88	0.39	0.29	1.67	1.63
	u-CI	1.52	1.57	1.02	0.70	5.63	22.16
	No.	407	412	407	412	318	326

$_{C}OR$: Mantel-Haenszel common odds ratio adjusting for matching variables ; $\Omega_{mh}=[\Sigma a_i d_i / N_i] / [\Sigma b_i c_i / N_i]$

$_{A}OR_1$: Adjusted odds ratio for matching variables from an unconditional linear logistic model ; logit $P_i=\alpha_0+\alpha_1(RACE)+\alpha_2(QBTHYR)+\beta_1(each\ variable)$

$_{A}OR_2$: Adjusted odds ratio for matching variables from an unconditional linear logistic model ; logit $P_i=\alpha_0+\alpha_1(RACE\ and\ QBTHYR)+\beta_1(each\ variable)$

$_{M}OR$: matched odds ratio estimated by a linear logistic regression model to fit triple-matched data for a conditional maximum likelihood estimate ; logit $P_i=\beta_1^*(X_{i0}-X_{i1})$

l-CI : 95% lower confidence limit calculated by logit methods for $_{A}OR_1$ and $_{A}OR_2$, and by test-based method for $_{C}OR$ and $_{M}OR$

u-CI : 95% upper cpnfidence limit calculated by logit methods for $_{A}OR_1$ and $_{A}OR_2$, and by test-based method for $_{C}OR$ and $_{M}OR$

odds ratios. It means that the estimation bias may be alleviated after adjustment by these methods, when the risk factor variable was thought to be affected by confounding effect. Therefore, effectiveness of the three control methods for estimation bias can be recognized, when matching was done with apparent reason for confounding. Adjusted values of BTHSTAT were approaching to the almost identical value with the unmatched odds ratio, not towards the matched one. It can be referred that unconditional adjustment, as well as Mantel-Haenszel procedure, may alter the adjusted value against the matched odds ratio ; into the opposite direction to common concept, when the matching variable was a effect modifier.

Ⅳ. DISCUSSIONS

Epidemiologic studies, in which matching had been introduced to control for known confounder, are generally not so common, except several well-designed studies to test a specific hypothesis. Instead, it was a common fashion to use 'age' or 'sex' as a matching variable in a case-control study. However, if the confounding effect of a matching variable was uncertain, or if the matching was apparently unnecessary, then, the matching will inevitably incur overmatching. Matched analysis of such an overmatched data may lead to estimation bias, even if it were a well-organized analytic method. Therefore, it must be emphasized that conditional estimation method in matched data analysis is not always a method of choice for matched data.

Nevertheless, if the matching variable were either conditionally independent of the disease given the risk factors or conditionally independent of the risk factor given the disease status, unmatched analysis may have a rationale for the matched data (Breslow and Day, 1980). Undoubtedly, unmatched analysis of unmatched data will never raise estimation bias.

On the other hand, in spite of apparent reason for matching (confounding effect), if one do not use matched analytic method, serious estimation bias will be induced. Often, matching variable used to be ignored, based on a priori experience of the author. However, there are a lot of evidence that estimation bias should be unavoidable, if one tries to do unmatched analysis with matched data set, when confounding is apparent (Schlesselman, 1982).

In this matched case-control data, matching variables were chosen only for the convenience of control selection, like in many other epidemiologic studies. There was no evidence of confounding by QBTHYR on NFAGE (or NMAGE) and the risk of the disease. As a result, the estimator of NFAGE (and NMAGE) showed no bias. It seems to be reasonable interpretation that each of the variable was neither confounder nor effect modifier in the data. Stratum-specific values in Table 2 will adhere to this interpretation ($X^2_{homogeneity}=1.9(5)$, $p=0.86$ for NFAGE ; $X^2_{homogeneity}=4.8(5)$, $p=0.44$ for NMAGE). In this occasion, neither matched analysis nor adjustment by stratification nor multivariate modelling will be necessary.

On the other hand, NFBYR and NMBYR seems to be affected by a confounding effect of QBTHYR. It is because unmatched values were markedly deviated from the the matched odds ratios. Stratified analysis in Table 2 showed that each stratum-specific values was not significantly different from the other value in the strata ($X^2_{homogeneity}=0.7(1)$, $p=0.41$ for NFBYR ; $X^2_{homogeneity}=0.3(2)$, $p=0.85$ for NMBYR), in spite of relatively small number of strata. These findinds are highly compatible with the definition of confounder, rather than the effect modification (Kleinbaum *et al.*, 1982 ; Kelsey *et al.*, 1986). Moreover, high collinearity of QBTHYR with NFBYR (or NMBYR), as well as the fact that adjusted values have been changed towards the matched odds ratio will favor the role of QBTHYR as a confounder. If it were not a confounder, QBTHYR may be a proxy variable of NFBYR (or NMBYR) in the chain of the association. Unfortunately, comfirmation of confounding effect of QBTHYR is not feasible in this data, because case and controls were matched by the very variable of QBTHYR. Anyhow, elimination of the estimation bias by the three unmatched control procedures seems to be effective, if the bias were arisen from unmatched analysis of properly matched data (due to confounder).

The QBTHYR's role of effect modification in BTHSTAT is becoming clear, when one see the marked variation of stratum specific odds ratios in Table 2. Even though test results for homogeneity were not statistically significant, the chi-square value with degree of freedom was relatively higher than the other variables. ($X^2_{homogeneity}$=3.42(3), p=0.33 for BTHSTAT). This finding will hold the role of effect modification. In addition, the fact that effect modification of a variable will not be altered by its use for case-control matching will support it, too (Breslow and Day, 1980). Therefore, it can be drawn from the results that matched analysis of matched data could be inversely biased, if matching were done improperly ; not for confounder, but for effect modifier. Practically, if we could assess the effect modification in an unmatched arrangement of matched data, then, the bias may be prevented. If it were effect modifier, adjustment by stratification and by multivariate modelling wil give a good result. If so, unmatched and three adjusted odds ratios of BTHSTAT in Table 3 might be a true value, rather than the matched odds ratio, 3. 12.

For a group (frequency) matching, unmatched analysis can be applied, if the stratum size should be kept relatively large (Breslow and Day, 1980). That is the reason why stratified analysis can be used as an alternative method to the matched method using conditional likelihood. The Mantel-Haenszel method for a common odds ratio over strata is the most common method in stratified analysis. Its calculation procedure is so easy to understand, and it is widely accessible to various statistical softwares in computer system. In addition, the Mantel-Haenszel formular is not affected by zero cell entries and will give a consistent estimate of the common odds ratio even with large numbers of small strata (Breslow and Day, 1980). Its efficacy to control bias in matched data analysis was revealed to be similar to the multivariate modelling techniques.

The unconditional regression model for matched data provides estimation of stratum parameter, which is included in the model to adjust the ignoring effect of matching variable. The results from the two different multivariate models in this comparison were nearly identical, implicating no theoretical difference in the adjustment of matching variable in such a relatively small number of stratum parameter.

Conditional analysis for matched data has a disadvantage of loss of information in the data. Its computational complexity looks sometimes too far to access, and too hard to understand. Such problems may lead analysts to unmatched analysis, i. e. unconditional logistic regression method. Fortunately, there is a theoretical basis on the estimation of conditional maximum likelihood estimate, using a usual linear logistic regression model (Holford *et al.*, 1978). It enables epidemiologists to obtain matched odds ratio in a pair-matched case-control study. It can be done with usual statistical packages capable of linear modelling, i. e. the GLIM, the EGRET, the EPILOG. For the SAS system (Version 6.03), another program is available for pair-matched (Yoo, 1990).

ACKNOWLEDGEMENT

I would like to give my sincere thanks to Dr. Dwight T Janerich and Dr. Robert Dubrow for their thoughtful advise and provision of the data for this article, while my stay in the Cancer Prevention Research Unit for the Connecticut in Yale University School of Medicine.

REFERENCES

Breslow NE, Day NE. *Statistical methods in cancer research ; the analysis of case control studies. Lyon, IARC Sci Publ, 1980*

Cornfield J. *A method of estimating comparative rates from clinical data. Application to cancer of the lung, breast and cervix. JNCI 1951 ; 11 : 1269*

Feinstein AR. *Quantitative ambiguities in matched versus unmatched analyses of the 2 x 2 table for a case-control study. Int J Epidemiol 1987 ; 16 : 128-134*

Holford TR, White C, Kelsey JL. *Multivariate analysis for matched case-control studies. Am J Epidemiol 1978 ; 107 : 245-256*

Janerich DT, Yoo KY, Dubrow R. *Investigation of birth characteristics*

of males diagnosed with testicular cancer. 1991 (in preparation)

Kelsey JL, Thompson WD, Evans AS. *Methods in observational epidemiology. New York : Oxford University Press, 1986.*

Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic research ; principles and quantitative methods. New York, Lifetime Learning Publications, Van Nostrand Reinhold Company, 1982*

Mantel N, Haenszel W. *Statistical aspects of the analysis of data from retrospective studies of disease. JNCI 1959 ; 22 : 719~48*

Miettinen OS. *Estimability and estimation in case-referent studies. Am J Epidemiol 1976 ; 103 : 226-35*

Rothman KJ. *Modern epidemiology. Boston : Little Brown, 1986*

Schlesselman JJ. *Case-control studies. Chapter 4. New York : Oxford Univ Press, 1982*

SAS Institute, Inc. *SAS/STAT guide for personal computers, Version 6 edition. Cary, NC : SAS Institute, Inc, 1987*

The GLIM Working Party. *The generalised linear interactive modelling system, release 3. 77. Oxford : Numerical Algorithms Group, Ltd, 1987*

Woolf B. *On estimating the relation between blood group and disease. Ann Human Genetics 1955 ; 19 : 251-53*

Yoo KY. *Use of a SAS program for conditional maximum likelihood estimate in linear logistic model to fit pair matched data. Korean J Epidemiol 1990 ; 12(1) : 93-99*

Yoo KY, Dubrow R, Janerich DT. *Estimation bias arising from unmatched analysis of 1-to-2 matched triples. 1991 (in preparation)*