Analysis of matched case-control studies

Neil Pearce[1,2]

[1]Department of Medical Statistics and Centre for Global NCDs, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

[2]Centre for Public Health Research, Massey University, Wellington, New Zealand
neil.pearce@lshtm.ac.uk

Cite this as: *BMJ* 2016;352:i969
http://dx.doi.org/10.1136/bmj.i969

Accepted: 30 December 2015

There are two common misconceptions about case-control studies: that matching in itself eliminates (controls) confounding by the matching factors, and that if matching has been performed, then a "matched analysis" is required. However, matching in a case-control study does not control for confounding by the matching factors; in fact it can introduce confounding by the matching factors even when it did not exist in the source population. Thus, a matched design may require controlling for the matching factors in the analysis. However, it is not the case that a matched design requires a matched analysis. Provided that there are no problems of sparse data, control for the matching factors can be obtained, with no loss of validity and a possible increase in precision, using a "standard" (unconditional) analysis, and a "matched" (conditional) analysis may not be required or appropriate.

Matching on factors such as age and sex is commonly used in case-control studies.[1] This can be done for convenience (eg, choosing a control admitted to hospital on the same day as the case), to improve study efficiency by improving precision (under certain conditions) when controlling for the matching factors (eg, age, sex) in the analysis, or to enable control in the analysis of unquantifiable factors such as neighbourhood characteristics (eg, by choosing neighbours as controls and then controlling for neighbourhood in the analysis). The increase in efficiency occurs because it ensures similar numbers of cases and controls in confounder strata. For example, in a study of lung cancer, if controls are sampled at random from the source population, their age distribution will be much younger than that of the lung cancer cases. Thus, when age is controlled in the analysis, the young age stratum may contain mostly controls and few cases, whereas the old age stratum may contain mostly cases and fewer controls. Thus, statistical precision may be improved if controls are age matched to ensure roughly equal numbers of cases and controls in each age stratum.

There are two common misconceptions about case-control studies: that matching in itself eliminates confounding by the matching factors; and that if matching has been performed, then a "matched analysis" is required.

Matching in the design does not control for confounding by the matching factors. In fact, it can introduce confounding by the matching factors even when it did not exist in the source population.[1] The reasons for this are complex and will only be discussed briefly here. In essence, the matching process makes the controls more similar to the cases not only for the matching factor but also for the exposure itself. This introduces a bias that needs to be controlled in the analysis. For example, suppose we were conducting a case-control study of poverty and death (from any cause), and we chose siblings as controls (that is, for each person who died, we matched on family or residence by choosing a sibling who was still alive as a control). In this situation, since poverty runs in families we would tend to select a disadvantaged control for each disadvantaged person who had died and a wealthy control for each wealthy person who had died. We would find roughly equal percentages of disadvantaged people among the cases and controls, and we would find little association between poverty and mortality. The matching has introduced a bias, which fortunately (as we will illustrate) can be controlled by controlling for the matching factor in the analysis.

Thus, a matched design will (almost always) require controlling for the matching factors in the analysis. However, this does not necessarily mean that a matched analysis is required or appropriate, and it will often be sufficient to control for the matching factors using simpler methods. Although this is well recognised in both recent[2,3] and historical[4,5] texts, other texts[6-9] do not discuss this issue and present the matched analysis as the only option for analysing matched case-control studies. In fact, the more standard analysis may not only be valid but may be much easier in practice, and yield better statistical precision.

In this paper I explore and illustrate these problems using a hypothetical pair matched case-control study.

SUMMARY POINTS

- Matching in a case-control study does not control for confounding by the matching factors
- A matched design may require controlling for the matching factors in the analysis
- However, it is not the case that a matched design requires a matched analysis
- A "standard" (unconditional) analysis may be most valid and appropriate, and a "matched" (conditional) analysis may not be required or appropriate

Options for analysing case-control studies

Unmatched case-control studies are typically analysed using the Mantel-Haenszel method[10] or unconditional

Table 1 | Hypothetical study population and case-control study with unmatched and matched standard analyses

	Young participants		Old participants		Total		Odds ratio (95% CI)	
	Exposed	Not exposed	Exposed	Not exposed	Exposed	Not exposed	Crude	Age adjusted
Total population:								
Cases	80	10	100	200	180	210	0.86 (0.70 to 1.05)	2.00 (1.59 to 2.51)*
Non-cases	80 000	20 000	20 000	80 000	100 000	100 000		
Unmatched case-control study:								
Cases	80	10	100	200	180	210	0.86 (0.65 to 1.14)	2.00 (1.38 to 2.89)
Controls	156	39	39	156	195	195		
Matched case-control study standard analysis:								
Cases	80	10	100	200	180	210	1.68 (1.25 to 2.24)	2.00 (1.42 to 2.81)*
Controls	72	18	60	240	132	258		

*"True" age adjusted.

logistic regression.[4] The former involves the familiar method of producing a 2×2 (exposure-disease) stratum for each level of the confounder (eg, if there are five age groups and two sex groups, then there will be 10 2×2 tables, each showing the association between exposure and disease within a particular stratum), and then producing a summary (average) effect across the strata. The Mantel-Haenszel estimates are robust and not affected by small numbers in specific strata (provided that the overall numbers of exposed or non-exposed cases or controls are adequate), although it can be difficult or impossible to control for factors other than the matching factors if some strata involve small numbers (eg, just one case and one control). Furthermore, the Mantel-Haenszel approach works well when there are only a few confounder strata, but will experience problems of small numbers (eg, strata with only cases and no controls) if there are too many confounders to adjust for. In this situation, logistic regression may be preferred, since this uses maximum likelihood methods, which enable the adjustment (given certain assumptions) of more confounders.

Suppose that for each case we have chosen a control who is in the same five year age group (eg, if the case is aged 47 years, then a control is chosen who is aged 45-49 years). We can then perform a standard analysis, which adjusts for the matching factor (age group) by grouping all cases and controls into five year age groups and using unconditional logistic regression[4] (or the Mantel-Haenszel method[10]); if there are eight age groups then this analysis will just have eight strata (represented by seven age group dummy variables), each with multiple cases and controls. Alternatively we can perform a matched analysis (that is, retaining the pair matching of one control for each case) using conditional logistic regression (or the matched data methods, which are equivalent to the Mantel-Haenszel method); if there are 100 case-control pairs, this analysis will then have 100 strata.

Table 2 | Hypothetical matched case-control study with matched analysis

	Control		
	Exposed	Not exposed	Pair matched odds ratio (95% CI)
Case exposed	84	96	2.00 (1.40 to 2.89)
Case not exposed	48	162	

The main reason for using conditional (rather than unconditional) logistic regression is that when the analysis strata are very small (eg, with just one case and one control for each stratum), problems of sparse data will occur with unconditional methods.[11] For example, if there are 100 strata, this requires 99 dummy variables to represent them, even though there are only 200 study participants. In this extreme situation, unconditional logistic regression is biased and produces an odds ratio estimate that is the square of the conditional (true) estimate of the odds ratio.[5 12]

Example of age matching

Table 1 gives an example of age matching in a population based case-control study, and shows the "true' findings for the total population, the findings for the corresponding unmatched case-control study, and the findings for an age matched case-control study using the standard analysis. Table 2 presents the findings for the same age matched case-control study using the matched analysis. All analyses were performed using the Mantel-Haenszel method, but this yields similar results to the corresponding (unconditional or conditional) logistic regression analyses.

Table 1 shows that the crude odds ratio in the total population is 0.86 (0.70 to 1.05), but this changes to 2.00 (1.59 to 2.51) when the analysis is adjusted for age (using the Mantel-Haenszel method). This occurs because there is strong confounding by age—the cases are mostly old, and old people have a lower exposure than young people. Overall, there are 390 cases, and when 390 controls are selected at random from the non-cases in the total population (which is half exposed and half not exposed), this yields the same crude (0.86) and adjusted (2.00) odds ratios, but with wider confidence intervals, reflecting the smaller numbers of non-cases (controls) in the case-control study.

Why matching factors need to be controlled in the analysis

Now suppose that we reconduct the case-control study, matching for age, using two very broad age groups: old and young (table 1). The number of cases and controls in each age group are now equal. However, the crude odds ratio (1.68, 1.25 to 2.24) is different from both the crude (0.86) and the adjusted (2.00) odds ratios in

the total population. In contrast, the adjusted odds ratio (2.00) is the same as that in the total population and in the unmatched case-control study (both of these adjusted odds ratios were estimated using the standard approach). Thus, matching has not removed age confounding and it is still necessary to control for age (this occurs because the matching process in a case-control study changes the association between the matching factor and the outcome and can create an association even if there were none before the matching was conducted). However, there is a small increase in precision in the matched case-control study compared with the unmatched case-control studies (95% confidence intervals of 1.42 to 2.81 compared with 1.38 to 2.89) because there are now equal numbers of cases and controls in each age group (table 1).

A pair matched study does not necessarily require a pair matched analysis

However, control for simple matching factors such as age does not require a pair matched analysis. Table 2 gives the findings that would have been obtained from a pair matched analysis (this is created by assuming that in each age group, and for each case, the control was selected at random from all non-cases in the same age group). The standard adjusted (Mantel-Haenszel) analysis (table 1) yields an odds ratio of 2.00 (95% confidence interval 1.42 to 2.81); the matched analysis (table 2) yields the same odds ratio (2.00) but with a slightly wider confidence interval (1.40 to 2.89).

Advantages of the standard analysis

So for many matched case-control studies, we have a choice of doing a standard analysis or a matched analysis. In this situation, there are several possible advantages of using the standard approach.

The standard analysis can actually yield slightly better statistical precision.[13] This may apply, for example, if two or more cases and their matched controls all have identical values for their matching factors; then combining them into a single stratum produces an estimator with lower variance and no less validity[14] (as indicated by the slightly narrower confidence interval for the standard adjusted analysis (table 1) compared with the pair matched analysis (table 2). This particularly occurs because combining strata with identical values for the matching factors (eg, if two case-control pairs all concern women aged 55-59 years) may mean that fewer data are discarded (that is, do not contribute to the analysis) because of strata where the case and control have the same exposure status. Further gains in precision may be obtained if combining strata means that cases with no corresponding control (or controls without a corresponding case) can be included in the analysis. When such strata are combined, a conditional analysis may still be required if the resulting strata are still "small,"[13] but an unconditional analysis will be valid and yield similar findings if the resulting strata are sufficiently large. This may often be the case when matching has only been performed on standard factors such as sex and age group.

The standard analysis may also enhance the clarity of the presentation, particularly when analysing subgroups of cases and controls selected for variables on which they were not matched, since it involves standard 2×2 tables for each subgroup.[15]

A further advantage of the standard analysis is that it makes it easier to combine different datasets that have involved matching on different factors (eg, if some have matched for age, some for age and sex, and some for nothing, then all can be combined in an analysis adjusting for age, sex, and study centre). In contrast, one multicentre study[16] (of which I happened to be a coauthor) attempted to (unnecessarily) perform a matched analysis across centres. Because not all centres had used pair matching, this involved retrospective pair matching in those centres that had not matched as part of the study design. This resulted in the unnecessary discarding of the unmatched controls, thus resulting in a likely loss of precision.

Conclusions

If matching is carried out on a particular factor such as age in a case-control study, then controlling for it in the analysis must be considered. This control should involve just as much precision as was used in the original matching[14] (eg, if exact age in years was used in the matching, then exact age in years should be controlled for in the analysis), although in practice such rigorous precision may not always be required (eg, five year age groups may suffice to control confounding by age, even if age matching was done more precisely than this). In some circumstances, this control may make no difference to the main exposure effect estimate—eg, if the matching factor is unrelated to exposure. However, if there is an association between the matching factor and the exposure, then matching will introduce confounding that needs to be controlled for in the analysis.

So when is a pair matched analysis required? The answer is, when the matching was genuinely at (or close to) the individual level. For example, if siblings have been chosen as controls, then each stratum would have just one case and the sibling control; in this situation, an unconditional logistic regression analysis would suffer from problems of sparse data, and conditional logistic regression would be required. Similar situations might arise if controls were neighbours or from the same general practice (if each general practice only had one or a few cases), or if matching was performed on many factors simultaneously so that most strata (in the standard analysis) had just one case and one control.

Provided, however, that there are no problems of sparse data, such control for the matching factors can be obtained using an unconditional analysis, with no loss of validity and a possible increase in precision.

Thus, a matched design will (nearly always) require controlling for the matching factors in the analysis. It is not the case, however, that a matched design requires a matched analysis.

I thank Simon Cousens, Deborah Lawlor, Lorenzo Richiardi, and Jan Vandenbroucke for their comments on the draft manuscript. The Centre for Global NCDs is supported by the Wellcome Trust Institutional Strategic Support Fund, 097834/Z/11/B.

Competing interests: I have read and understood the BMJ policy on declaration of interests and declare the following: none.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Rothman KJ, Greenland S, Lash TL, eds Design strategies to improve study accuracy.*Modern epidemiology.*3rd ed. Lippincott Williams & Wilkins, 2008.
2 Rothman KJ. *Epidemiology: an introduction.*Oxford University Press, 2012.
3 Rothman KJ, Greenland S, Lash TL, eds. *Modern epidemiology.*3rd ed. Lippincott Williams & Wilkins, 2008.
4 Breslow NE, Day NE. *Statistical methods in cancer research. Vol I: the analysis of case-control studies.*IARC, 1980.
5 Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic research: principles and quantitative methods.*Lifetime Learning Publications, 1982.
6 Dos Santos Silva I. *Cancer epidemiology: principles and methods.*IARC, 1999.
7 Keogh RH, Cox DR. *Case-control studies.*Cambridge University Press, 2014doi:10.1017/CBO9781139094757
8 Lilienfeld DE, Stolley PD. *Foundations of epidemiology.*3rd ed. Oxford University Press, 1994.
9 MacMahon B, Trichopolous D. *Epidemiology: principles and methods.*2nd ed. Little Brown, 1996.
10 Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;22:719-48.
11 Robins J, Greenland S, Breslow NE. A general estimator for the variance of the Mantel-Haenszel odds ratio. *Am J Epidemiol* 1986;124:719-23.
12 Pike MC, Hill AP, Smith PG. Bias and efficiency in logistic analyses of stratified case-control studies. *Int J Epidemiol* 1980;9:89-95. doi:10.1093/ije/9.1.89.
13 Brookmeyer R, Liang KY, Linet M. Matched case-control designs and overmatched analyses. *Am J Epidemiol* 1986;124:693-701.
14 Greenland S. Applications of stratified analysis methods. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern epidemiology.*3rd ed. Lippincott Williams & Wilkins, 2008.
15 Vandenbroucke JP, Koster T, Briët E, Reitsma PH, Bertina RM, Rosendaal FR. Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. *Lancet* 1994;344:1453-7. doi:10.1016/S0140-6736(94)90286-0.
16 Cardis E, Richardson L, Deltour I, et al. The INTERPHONE study: design, epidemiological methods, and description of the study population. *Eur J Epidemiol* 2007;22:647-64. doi:10.1007/s10654-007-9152-z.
17 Mansournia MA, Hernán MA, Greenland S. Matched designs and causal diagrams. *Int J Epidemiol* 2013;42:860-9. doi:10.1093/ije/dyt083.
18 Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15:615-25. doi:10.1097/01.ede.0000135174.63482.43.