DISTRIBUTION : LIMITED
CDD/EDP/89.1
ENGLISH ONLY

DIARRHOEAL DISEASES CONTROL PROGRAMME

Case-Control Studies of Childhood Diarrhoea:

III. Matching

by
S.N. Cousens, R.G. Feachem, B. Kirkwood
T.E. Mertens, P.G. Smith

WORLD HEALTH ORGANIZATION

CASE-CONTROL STUDIES OF CHILDHOOD DIARRHOEA

III. MATCHING

S.N. Cousens (1)
R.G. Feachem (1)
B.R. Kirkwood (1)
T.E. Mertens (1)
P.G. Smith (1)

(1) Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT

Contents

PREFACE

This document is the third in a series prepared for the Diarrhoeal Diseases Control Programme of the World Health Organization. The series is a response to an upsurge of interest in the application of the case-control method to the study of childhood diarrhoea. That interest has been stimulated by the realization that, under certain circumstances, the case-control method can be a relatively quick, inexpensive, and reliable method for measuring the impact of diarrhoea control measures or for identifying and quantifying risk factors for diarrhoea.

Case-control studies can be complex in their design and analysis and it is not possible to prepare a manual that can be followed exactly in all circumstances. A considerable amount of epidemiological judgement and skill must be exercised. The aim of this series is to provide the investigator with a clear view of the most important problems in the design, analysis, and interpretation of case-control studies of childhood diarrhoea, and to provide practical suggestions for the resolution of those problems. For the trained and experienced epidemiologist, these documents provide specialized guidance on the application of case-control methods. For others, the series provides an awareness of the methodological issues involved, and a familiarity with the language and concepts of case-control studies. While it is hoped that the entire series will be of interest to and available to readers, each document has been prepared as an independent piece. For this reason the documents overlap each other in some areas.

Diarrhoeal diseases remain one of the leading causes of morbidity and mortality among children in poor communities in all parts of the world. Epidemiological studies have already contributed to an understanding of the risk factors involved and to the design and evaluation of appropriate interventions. Continued work in diarrhoea epidemiology is essential to further refine these interventions and to maximize their impact on severe illness and death. The Diarrhoeal Diseases Control (CDD) Programme of WHO is supporting a range of research projects in this field in many countries. Those seeking financial or technical support for their research, or wishing to contact others undertaking similar investigations, are invited to contact the CDD Programme.

ABSTRACT

In this paper we discuss the merits and demerits of matching in case-control studies of childhood diarrhoea. Two rather different reasons for matching are presented. First, matching on a confounding variable may improve the precision of the estimate of the odds ratio obtained from a study. Second, individual matching on, for example, neighbourhood may allow the investigator to control a spectrum of ill-defined and hard-to-quantify variables such as socioeconomic status. The fact that matching in the design of a case-control

study does not remove the need to account for the matched variable(s) in the analysis is emphasized. Disadvantages of matching are also discussed. We suggest that in studies that recruit clinic-based controls, approximate frequency matching of controls to cases with regard to age, date of recruitment, and clinic of recruitment will usually be sufficient and will be logistically simpler than recruiting individually matched controls. In studies that recruit community-based controls with the aim of controlling a range of socioeconomic and environmental variables, it is recommended that the investigator collect data to confirm that the matching has been successful in this aim.

RESUME

Cet article traite des avantages et des inconvénients de l'appariement lors d'études cas-témoins de la diarrhée infantile. Deux justifications assez différentes de l'appariement sont présentées. Premièrement, l'appariement sur une variable confondante peut améliorer la précision de l'estimation de "l'odds ratio" obtenue d'une étude. Deuxièmement, l'appariement individuel sur, par exemple, le voisinage peut permettre à l'enquêteur de contrôler un éventail de variables mal définies et difficiles à quantifier, comme le niveau socio-économique. On insiste cependant sur le fait qu'apparier dans le plan d'une étude cas-témoins ne soustrait pas à l'obligation de tenir compte de la ou des variables appariées dans l'analyse. Les inconvénients de l'appariement sont également évoqués. Il nous semble que, pour des études dans lesquelles les témoins sont recrutés dans les dispensaires, un appariement témoins-cas du point de vue de l'âge, de la date et du dispensaire de recrutement en fonction de la fréquence approximative de la diarrhée est généralement suffisant et logistiquement plus simple que le recrutement de témoins individuellement appariés. Pour les études utilisant des témoins recrutés dans la communauté et portant sur toute une gamme de variables socio-économiques et environnementales, il est recommandé de recueillir des données qui confirment que l'appariement a été un succès de ce point de vue.

1. INTRODUCTION

In 1985 the World Health Organization issued a document entitled "Measuring the impact of water supply and sanitation facilities on diarrhoea morbidity: prospects for case-control methods" (Briscoe et al., 1985). This document was one of the products of two scientific meetings held in Cox's Bazaar, Bangladesh, and Geneva, Switzerland, at which methodologies for measuring the impact of water supply and sanitation projects on health were discussed. In the document, case-control studies were put forward as an alternative to longitudinal studies, whose use in this field had been discouraged by a report of an expert panel to the World Bank (IBRD, 1976).

The present series of papers considers the wider application of the case-control method to the study of the epidemiology of diarrhoeal diseases and of interventions for their control. The previous papers in the series dealt with the minimization of bias (number I) and the choice of sample size (number II). This paper, the third of the series, focuses on studies in which controls are matched to cases for certain characteristics. Matching implies that controls are not selected entirely at random, but in such a way as to ensure that they are similar to the cases with regard to some factor or factors (e.g., age). The rationale for matching, its advantages and disadvantages, and its implications for the analysis of case-control studies are discussed.

We avoid the use of complex algebraic expressions and present instead simple numerical examples wherever possible. The statistical formulae used in the presentation of these examples are cited in the Annex. We begin by considering a hypothetical case-control study.

2. AN EXAMPLE OF A CASE-CONTROL STUDY

An unmatched case-control study, designed to assess the association between the presence of domestic animals in the home and the risk of diarrhoea morbidity in children aged less than 5 years, is conducted. The study is based on patients attending a single health facility. "Cases" are those children reporting to the clinic in whom diarrhoea caused by an enteric infection is diagnosed; "controls" are randomly selected from among children reporting to the clinic with conditions not thought to be related to the presence of domestic animals in the home and who are not suffering from diarrhoea. Information concerning the presence of animals in the households of both cases and controls is collected.

In their simplest form, the results of the study may be presented in the form of a 2 x 2 table:

	Cases	Controls	
Animals present	10	4	14
Animals not present	30	36	66
Total	40	40	80

The measure of association used in the analysis of case-control studies is the odds ratio. For the above table this is calculated as follows:

$$OR = \frac{10 \times 36}{4 \times 30} = 3.0$$

This result suggests that children who live in houses where domestic animals are present are approximately three times more likely to suffer an attack of diarrhoea than children in houses without animals.

To interpret the results correctly, we need to test the statistical significance of the association we have found in our sample. Is there really an underlying association between the presence of animals and risk of diarrhoea, or could our result have been obtained by chance? Even when studying a factor that is not associated with diarrhoea (i.e., true odds ratio = 1.0), we are unlikely to obtain an estimate exactly equal to 1.0 due to sampling variations. How likely is it that our estimate of 3.0 has arisen in this way? One method of testing the significance of an association in a 2 x 2 table is to perform a chi-squared (X^2) test with one degree of freedom (Annex). From the table,

$$X^2 = \frac{80 \, (\, |10x36 - 4x30| - 0.5x80)^2}{40x40 \times 66x14}$$

$$= 2.16$$

Comparing this value against a table of values for a chi-squared distribution with one degree of freedom, it may be seen that the probability of obtaining a similar or more extreme result purely by chance, in a situation in which the true odds ratio equals 1, is greater than 0.1. Thus our result is not statistically significant at the 10% level of significance. In this particular example we have not found strong evidence of an association. There are two possible reasons for this:

(1) no association exists between presence of animals and risk of diarrhoea,

(2) an association does exist, but our study was too small to detect it (i.e., to find a statistically significant association).

In our analysis and discussion of the example above, we implicitly assume that the cases and controls included in the study constitute unbiased samples of the children with and without diarrhoea. Our assessment of the statistical significance of the observed association is based on this assumption and may be invalid if bias has occurred to any great degree. One way in which bias may arise in case-control studies is through confounding (Schlesselman, 1982; Cousens et al., 1988a). Confounding is said to occur when a second, extraneous factor is associated with both the exposure of interest and the disease of interest (diarrhoea). Matching aims to ensure that the cases and controls are similar with regard to this second, extraneous factor. We illustrate the effect of confounding and the technique of matching with a second example.

3. AN EXAMPLE OF MATCHING

Consider a case-control study of the effect of breast-feeding on the incidence of diarrhoea, conducted in an area where breast-feeding is associated with socioeconomic status, and both non-breast-feeding and low socioeconomic status are associated independently with an increased risk of diarrhoea. The tables below describe such a population.

Low socioeconomic status

	Cases	Non-cases		
Non-breast-fed	15	1000	1015	Rate/1000 = 14.8
Breast-fed	24	4000	4024	Rate/1000 = 6.0
	39	5000	5039	

Incidence rate ratio = 2.5

High socioeconomic status

	Cases	Non-cases		
Non-breast-fed	10	4000	4010	Rate/1000 = 2.5
Breast-fed	1	1000	1001	Rate/1000 = 1.0
	11	5000	5011	

Incidence rate ratio = 2.5

Note that, within each socioeconomic group, children who are not breast-fed are about two and a half times as likely to suffer an episode of diarrhoea as children who are breast-fed. Notice also that children with lower socioeconomic status are much more likely to be breast-fed than children with higher socioeconomic status (80% versus 20%), and that diarrhoea is more common in the former group among both breast-fed and non-breast-fed children.

Initially, assume that controls are selected at random from children without diarrhoea (non-cases); i.e., an unmatched study is conducted. Then we would expect to obtain results similar to those shown below. The expected numbers of controls in each category have been calculated as follows: there are 50 controls altogether (to go with 50 cases), to be selected from a total population of 5000 + 5000 = 10 000 potential controls; there are 1000 potential controls from the group with low socioeconomic status who are not breast-fed; therefore, the number of controls we would expect to recruit from this category is 50 x 1000/10000 = 5.

	Socioeconomic status					
	Low			High		
	Cases	Controls		Cases	Controls	
Non-breast-fed	15	5	20	10	20	30
Breast-fed	24	20	44	1	5	6
	39	25	64	11	25	36

From these data we obtain the following results (Annex):

Mantel-Haenszel odds ratio = 2.5, 95% confidence interval (0.75,8.30)
Mantel-Haenszel chi-squared = 2.24, p=0.13

The Mantel-Haenszel estimator of the odds ratio (Mantel and Haenszel, 1959) provides an unbiased summary estimate of the population odds ratio, and the 95% confidence interval around the estimate is from 0.75 to 8.30. Failure to stratify on socioeconomic status (i.e., adding together the two tables above) would have produced the following table:

	Cases	Controls	
Non-breast-fed	25	25	50
Breast-fed	25	25	50
Total	50	50	100

From this table the estimated odds ratio is

OR = 1.0

This crude estimate of the odds ratio is biased, suggesting that there is no association between non-breast-feeding and diarrhoea. This bias occurs because socioeconomic status is a confounder of the association between breast-feeding and diarrhoea in this population.

Now suppose that, in order to reduce the confounding effect of socio-economic status, a matched study is conducted. Instead of being selected at random, controls are selected to be similar to cases with regard to socio-economic status (but without regard to the controls' breast-feeding status); i.e., for each case recruited from the group with low socioeconomic status a control from that group is recruited. Then we would expect to obtain results similar to those shown below. This time the numbers of controls in each category have been calculated as follows: in the group with low socioeconomic status there are 39 controls altogether (to go with 39 cases), to be selected from a total population of 5000 potential controls; there are 1000 potential controls from the group with low socioeconomic status who are not breast-fed; therefore the number of controls we would expect to recruit from this category is 39 x 1000/5000 = 7.8; this figure is rounded off to 8 in the table below.

	Socioeconomic status					
	Low			High		
	Cases	Controls		Cases	Controls	
Non-breast-fed	15	8	23	10	9	19
Breastfed	24	31	55	1	2	3
	39	39	78	11	11	22

From these data we obtain the following results (Annex):

Mantel-Haenszel odds ratio = 2.4, 95% confidence interval (0.82,6.97)
Mantel-Haenszel chi-squared = 2.56, p=0.11

(The slight discrepancy between the estimate of the odds ratio and the population odds ratio is explained by the fact that the expected number of controls in each cell has been rounded to a whole number.)

Comparing these results with those from the unmatched study, it can be seen that, while both studies have provided an unbiased estimate of the odds ratio, matching has narrowed the 95% confidence interval around the estimate, from (0.75,8.30) to (0.82,6.97); i.e., matching has increased the precision of the estimate of the odds ratio.

Matching on a confounding variable in a follow-up study removes the need to control (stratify on) that variable during the analysis (see, for example, Rothman, 1986). This is not, however, the situation in a case-control study. To illustrate this point we present an analysis of the matched study described above in which the matching is ignored. The summary 2 x 2 table is shown below.

	Cases	Controls	
Non-breast-fed	25	17	42
Breast-fed	25	33	58
Total	50	50	100

OR = 1.94

It can be seen that, in contrast to follow-up studies, matching on a confounding variable in the design alone, without control of that variable in the analysis, has resulted in a biased estimate of the odds ratio. It may also be seen that, in this example, the effect of the bias is to underestimate the strength of the association (bias towards the null value). Failure to control in the analysis the confounding variables matched for in the design of a case-control study will always result in this type of bias (Rothman, 1986).

The point illustrated above is extremely important. Matching in the design of a case-control study may help to improve the precision of the estimated odds ratio obtained from the study. Matching on confounding variables in the design alone will not, however, eliminate the confounding effect of those variables unless the variables used for matching are controlled in the analysis.

4. INDIVIDUAL MATCHING

In the preceding section we considered a study in which the number of controls sampled from each socioeconomic stratum is chosen to be equal to the number of cases recruited from that stratum, i.e., matching is applied to groups of subjects. This type of matching is often called frequency matching (Rothman, 1986) or stratum matching. A more extreme form of matching occurs when each control is matched to a single case, i.e., the matching is applied to individual subjects. Such a strategy is commonly employed and is referred to as individual matching (Rothman, 1986) or pairwise matching.

One objective of individual matching may be to ensure that the distribution of cases and controls is similar with regard to a number of clearly defined and measurable variables. For example, a study of the association between water supply/sanitation facilities and diarrhoea, conducted in six health facilities in Nicaragua, recruited clinic controls individually matched to cases on age, sex, clinic of recruitment, and date of recruitment (Sandiford, 1988). A case-control study of risk factors for diarrhoeal death, conducted in a diarrhoeal diseases hospital in Bangladesh (Islam and Khan, 1986), selected as controls the closest surviving patient in the hospital register, matched to the case for sex, age, and etiological agent.

Breslow and Day (1980) have pointed out that, in addition to improving the precision of a study, individual matching of controls to cases on place of residence (neighbourhood) may be equivalent to matching for a complex of underlying factors, some of which may be only vaguely defined and difficult to quantify (e.g., access to health services, socioeconomic status). Such factors might be difficult, if not impossible, to control without the use of individual matching. Victora et al. (1987) used community controls matched to cases for neighbourhood in a case-control study of infant mortality due to diarrhoea performed in Brazil. In a study of the effectiveness of BCG vaccination conducted in Colombia (Shapiro et al., 1985), controls were selected from among the members of the household of each case (i.e., matched on household). Another study of the effectiveness of BCG, conducted in Sri Lanka (Smith, 1987), chose controls from among the neighbours of tuberculosis cases (neighbourhood controls). This strategy should, however, be used with care, since it may not always achieve its desired ends (see section 5).

The use of individual matching in a study has implications for the analysis of that study. In section 3 we have shown that matching in the design alone does not eliminate bias due to confounding, and that it is necessary to retain the stratification used in the matching in the analysis. In a study in which an individual control is matched to an individual case (e.g., on neighbourhood, age, and sex), each stratum consists of one case and its control. In a study such as that conducted in Nicaragua (Sandiford, 1988), in which some 1500 case-control pairs were recruited, it would be very tiresome to calculate the Mantel-Haenszel estimate of the odds ratio across 1500 2 x 2 tables. Fortunately, there is a simpler method of calculating the odds ratio in this situation, which we illustrate below using some data presented by Islam and Khan (1986).

Each case-control pair may be classified into one of four categories:

a. case "exposed", control "exposed",
b. case "exposed", control "unexposed",
c. case "unexposed", control "exposed",
d. case "unexposed", control "unexposed".

We may thus construct a 2 x 2 table in which each of the four cells corresponds to one of the categories listed above. Application this procedure to the data collected by Islam and Khan (1986) concerning the association between hyponatraemia and risk of death leads to the following table:

		Controls		
		+	-	Total
Cases	+	19	73	92
	-	37	150	187
		56	223	279

where + indicates the presence of hyponatraemia, and - indicates the absence of hyponatraemia.

Note that each entry in the table represents a number of case-control pairs, not a number of individual cases or controls as was the case in previous tables. Thus there were 19 instances when the case and its matched control both had hyponatraemia. In situations such as this, the odds ratio is estimated by:

$$OR = \frac{73}{37} = 1.97, \quad 95\% \text{ confidence interval } (1.32, 2.93)$$

and the statistical significance is tested by calculating

$$X^2 = \frac{(|73-37| - 1)^2}{73+37} = 11.14, \quad p=0.0008$$

The statistical significance of the association may be assessed by comparing X^2 with the chi-squared distribution with one degree of freedom. Notice that the odds ratio is calculated using the numbers of discordant pairs. Pairs in which both case and control are exposed or unexposed do not contribute to the estimate of the odds ratio.

The analysis illustrated above is known as a matched-pairs analysis, and the chi-squared test used is called McNemar's test. For details of the formulae used refer to the Annex. This type of analysis may be extended to deal with studies in which individual matching has been applied, but in which more than one control has been recruited per case. Details of these extensions may be found in Breslow and Day (1980) and Schlesselman (1982).

It is also interesting to examine the effect, in this particular example, of performing an unmatched analysis. Retabulation of the data presented above, ignoring the pairing of controls with individual cases, leads to the following table:

	Hyponatraemic	Non-hyponatraemic	Total
Cases	92	187	279
Controls	56	223	279
	148	410	558

$$OR = 1.96, \quad 95\% \text{ confidence interval } (1.32, 2.90)$$

$$X^2 = 11.26, \quad p=0.0006$$

There were 92 instances in which a case was hyponatraemic. In 19 of these instances the corresponding matched control was also hyponatraemic, and in 73 instances the corresponding control was not (see previous table). Comparing the results of the unmatched analysis with those obtained from the matched-pairs analysis, it can be seen that, in this study, failure to stratify (retain the matching) in the analysis has not led to any substantial bias in the estimate of the odds ratio. This is in contrast to the first example, in which failure to stratify on socioeconomic status led to substantial bias in the estimate of the odds ratio. This difference between the two examples can probably be explained in the following way. In the first example, the variable used for matching - socioeconomic status - is a strong confounder of the association between breast-feeding and diarrhoea. Failure to control socioeconomic status in the analysis of the matched or unmatched study leads, therefore, to a biased estimate of the odds ratio. In the second example, the variables used for matching (including age, sex, and etiological agent) do not confound the association between hyponatraemia and risk of death. Failure to control (stratify on) these variables does not, therefore, lead to any important bias in the estimate of the odds ratio. It may further be noted that, in this second example, the estimate obtained from performing an unmatched analysis is slightly more precise than that obtained from a matched-

pairs analysis. In this instance no benefit has been derived from matching. In the following section we consider under what circumstances matching may be advantageous.

5. ADVANTAGES OF MATCHING

When deciding whether or not to match in a case-control study of diarrhoea, the investigator must bear in mind what matching can or may achieve and what it cannot: matching on a confounding variable can, in certain circumstances, improve the precision of the odds ratio estimated by the study; matching on a confounding variable may enable the investigator to control a range of vaguely specified or unquantifiable variables; matching cannot remove the need to take account of confounding variables in the analysis (although pairwise matching may simplify the analysis).

Thompson (1980) has assessed the effect of matching on a confounding variable over a range of situations. Schlesselman (1982) has reviewed this work and presented what he considers to be the major findings. Of these, the following are of most interest and importance to investigators working in the field of diarrhoea.

(a) Matching on a confounding variable will, in most commonly occurring situations, lead to a reduction in the variance of the log odds ratio of between about 5 and 15%. This is consistent with the findings reported by Smith and Day (1984) that, in general, it is not necessary to increase the size of an unmatched study by more than about 15% to allow for the effect of a single confounding variable (see Cousens et al., 1988b).

(b) Matching on a confounding variable will be more effective in improving the precision of the odds ratio estimate than suggested above when either the matching variable is a strong risk factor for diarrhoea, or the exposure of interest is rare in the population. These findings are also consistent with those of Smith and Day (1984). Schlesselman (1982) suggests that a matching variable is a strong risk factor for diarrhoea if the odds ratio is 20 or more and that exposure is rare if less than 10% of the population are exposed.

(c) The strength of the association between the risk factor of interest and the matching variable does not greatly influence the effectiveness of the matching.

Returning to the first example - of breast-feeding, socioeconomic status, and diarrhoea, we have already seen that the odds ratio of the association between non-breast-feeding and diarrhoea is 2.5. By retabulating the population data we may also look at the strength of the associations between breast-feeding and socioeconomic status, and between socioeconomic status and diarrhoea. The tables below reveal the association between breast-feeding and socioeconomic status.

	Cases			Controls		
	Socioeconomic status					
	Low	High		Low	High	
Non-breast-fed	15	10	25	1000	4000	5000
Breast-fed	24	1	25	4000	1000	5000
	39	11	50	5000	5000	10000
	OR = 0.0625			OR = 0.0625		

These tables show that children in the group with low socioeconomic status are about 16 times more likely to be breast-fed than children in the group with high socioeconomic status; i.e., there is quite a strong association between breastfeeding and socioeconomic status. The next pair of tables shows the association between socioeconomic status and diarrhoea.

	Breast-fed			Non-breast-fed		
Socioecon. status	Cases	Controls		Cases	Controls	
Low	24	4000	4024	15	1000	1015
High	1	1000	1001	10	4000	4010
	25	5000	5025	25	5000	5025
	OR = 6.0			OR = 6.0		

These tables show that children with low socioeconomic status are about six times more likely to suffer an attack of diarrhoea than children with higher socioeconomic status. Thus, in this population, the risk factor of interest (non-breast-feeding) is a relatively weak but common risk factor, while the confounder (socioeconomic status) is a somewhat stronger risk factor; and there is a strong association between the risk factor of interest and the confounder.

We can also calculate the effect of matching in this population. The variance of the estimate of the log odds ratio obtained from the unmatched study is approximately 0.409 (see Annex); for the matched study it is 0.340. In this population, the effect of matching has been to reduce this variance by about 17%. This reduction is marginally better than the 5-15% quoted by Schlesselman because socioeconomic status is a moderately strong risk factor for diarrhoea, and there is a strong association between breast-feeding and socioeconomic status.

It has also been suggested (Breslow and Day, 1980) that individual matching, together with an appropriate analysis, may allow the investigator to control a number of underlying variables that may be difficult to define or measure. In a case-control study of diarrhoea, the investigator might hope that controls chosen in the same neighbourhood (or household) as cases will match the cases with regard to socioeconomic status, access to health services, and a variety of environmental risk factors not under investigation. For example, at some future date an investigator may wish to evaluate the effectiveness of a rotavirus immunization programme. In such a study it will be important to control access to health services and immunization, since this will be related to risk of exposure (non-immunization) and may also be related to risk of (rotavirus) diarrhoea. It is difficult to measure or quantify an individual's access to something like immunization. Choosing neighbourhood controls <u>may</u> be the best way of dealing with this problem if neighbours share similar levels of access to health services. Such individual matching may not, however, always achieve its desired ends. Smith (1987) reports that, in the Sri Lankan study to assess the effectiveness of BCG vaccination, neighbourhood matching was used because it was thought that this would be an effective way of matching controls to cases with regard to socioeconomic status (a strong risk factor for tuberculosis). This did not turn out to be the case, for analyses of the data revealed a strong tendency for cases to come from poorer households than controls (Smith, 1987).

6. DISADVANTAGES OF MATCHING

In the preceding sections we have concentrated on the potential benefits of matching. We now consider the disadvantages of matching, of which there are several.

(a) The ability to investigate the association between the variable(s) used for matching and diarrhoea is lost. This can be shown using our first example - the study of breast-feeding. The results obtained from the unmatched study can be retabulated to examine the association between socioeconomic status and diarrhoea as shown below.

	Breast-fed			Non-breast-fed		
	Cases	Controls		Cases	Controls	
Low SE status	24	20	44	15	5	20
High SE status	1	5	6	10	20	30
	25	25	50	25	25	50

Mantel-Haenszel OR = 6.0 (1.93,18.66)

$X^2 = 9.58$, p=0.002

As we have shown in section 5, the population odds ratio of the association between socioeconomic group and diarrhoea is 6.0. Thus the unmatched study of breast-feeding and diarrhoea also provides us with an unbiased estimate of the association between socioeconomic status and diarrhoea. If we perform a similar retabulation of the results from the matched study we obtain the following results.

	Breast-fed			Non-breast-fed		
	Cases	Controls		Cases	Controls	
Low SE status	24	31	55	15	8	23
High SE status	1	2	3	10	9	19
	25	33	58	25	17	42

Mantel-Haenszel OR = 1.66 (0.33,8.42)

$X^2 = 0.37$, p=0.54

The estimate of the odds ratio obtained from the matched study is biased towards 1.0 because controls have been deliberately chosen to be similar to cases with regard to socioeconomic status. This bias cannot be corrected by any analytical technique.

(b) Potential controls will need to be screened, and some will be excluded because they do not fulfill the matching criteria; this will increase the work involved in performing the study, and perhaps also the cost. As the matching criteria are made more stringent, so the number of potential controls excluded from the study will increase.

(c) In some situations, matching may reduce the precision of the study (Thompson, 1980; Schlesselman, 1982). This will occur if the variable used for matching is associated with the risk factor/ exposure of interest but is not associated with diarrhoea (and is not therefore a confounder). We illustrate this effect with a hypothetical example. Suppose that it is wished to evaluate the effectiveness of a new rotavirus vaccine, which has been shown to have high protective efficacy in field trials, in the context of a routine immunization programme. A case-control study is performed to evaluate the vaccine's effectiveness and the investigator decides to match controls to cases with regard to sex. The pair of tables below describes the study population.

	Girls			Boys		
	Cases	Non-cases		Cases	Non-cases	
Not immunized	100	4000	4100	25	1000	1025
Immunized	5	1000	1005	20	4000	4020
	105	5000	5105	45	5000	5045
	OR = 5.0			OR = 5.0		

Children who have not been immunized are about five times more likely to suffer an attack of rotavirus diarrhoea than children who have been immunized (vaccine effectiveness = (1 - 1/OR) x 100% = 80%). Boys are more likely than girls to have been vaccinated, i.e., the matching variable is associated with the exposure of interest. Boys and girls are otherwise at equal risk of rotavirus diarrhoea, i.e., the matching variable is not associated with disease. The results of unmatched and matched studies conducted in this population are shown below.

Unmatched study

	Girls			Boys		
	Cases	Controls		Cases	Controls	
Not immunized	100	60	160	25	15	40
Immunized	5	15	20	20	60	80
	105	75	180	45	75	120

Mantel-Haenszel OR = 5.0 (2.65,9.45)

$X^2 = 24.55$, $p < 0.0001$

Matched study

	Girls			Boys		
	Cases	Controls		Cases	Controls	
Not immunized	100	84	184	25	9	34
Immunized	5	21	26	20	36	56
	105	105	210	45	45	90

Mantel-Haenszel OR = 5.0 (2.54,9.84)

$X^2 = 21.70$, $p < 0.0001$

Both studies provide an unbiased estimate of the odds ratio. However, instead of narrowing the confidence interval around the estimated odds ratio, matching has, in this example, increased the width of the interval.

(d) Matching may lead to bias in the estimate of the odds ratio. This effect is commonly known as overmatching and occurs when the matching variable lies on the causal pathway between the risk factor/ exposure of interest and the disease of interest (Breslow and Day, 1980). For example, keeping domestic animals in the yard may lead to increased diarrhoea in the children of the household by increasing their exposure to animal faeces. A case-control study of the role of domestic animals in the transmission of childhood diarrhoea that matched controls to cases with regard to the presence of animal faeces in the yard would underestimate any association that did exist between domestic animals and diarrhoea.

(e) Matching may complicate the analysis of a case-control study. In section 4 we have seen how to analyse a study of matched pairs. Such an analysis, however, controls only the variables that were matched in the recruitment process. If we wish to control another confounding variable that was not matched upon, we are faced with a problem. Imagine a study of water supply and diarrhoea which has recruited neighbourhood controls matched for age. Suppose we now wish to control for the effect of maternal education. If we wish to retain the original pairing in the analysis, we can only include those pairs in which the case and control have mothers with a similar level of education. We may be forced to drop a large number of pairs from the analysis, reducing the precision of the study. Alternatively, we can forget the original pairing and perform a stratified analysis of all the data, but this may lead to a biased estimate of the odds ratio. Finally, we can use a more advanced statistical procedure, conditional regression analysis, which allows us to retain the pairing and to take account of other confounding variables (Breslow and Day, 1980). However, such an analysis requires appropriate computer hardware and software, in contrast to the other analyses illustrated here, all of which can be performed on a hand-held calculator.

7. SUMMARY

In the preceding sections we present arguments for and against matching. Two distinct reasons for matching may be distinguished and we discuss these in turn.

Matching may be used to reduce the variance of the odds ratio estimated by the study. This reduction will, in most situations, be modest, and matching for this purpose is probably only worth considering when the matching variable(s) is/are strongly associated with diarrhoea. Any variables used for matching should not themselves be of interest as exposures since the matching will preclude any assessment of their association with diarrhoea. Matching will also complicate the conduct of a study. In many health facilities the performance of a case-control study will place an increased burden on an already hard-pressed staff: it will be unfair to these staff, and perhaps counterproductive, to further complicate matters by using complex matching criteria. In the light of these observations we recommend that individual matching of clinic controls should not be undertaken, and that frequency matching should be used only sparingly, typically on age, period of recruitment, and health facility.

Age is known to be strongly associated with risk of diarrhoea, and is not itself of interest as a risk factor. It should be possible, in most settings, to ask staff to achieve a rough balance of cases and controls in broad age groups, say 0-6 months, 6-12 months, 12-24 months, etc.

Diarrhoea is highly seasonal in many settings. Approximate frequency matching on week or fortnight of recruitment should be relatively easy to achieve, may help to reduce the variance of the odds ratio, and will ensure that the odds ratio estimates the incidence rate ratio (Rodrigues and Kirkwood, in preparation).

Finally, it is recommended that studies conducted in more than one health facility should ensure a balance of cases and controls in each facility, since exposure to a range of risk factors may vary from clinic to clinic. Again, this is a relatively easy variable on which to achieve a rough match.

The second possible reason for matching is in order to control a range of variables that are ill-defined and/or difficult to quantify. In case-control studies of diarrhoea, this type of matching is most likely to involve individual matching of neighbourhood controls to cases. It might be hoped that this procedure will match controls to cases for socioeconomic status, access to health services and a range of environmental risk factors. However, as we have mentioned, a study of BCG in Sri Lanka (Smith, 1987) that used neighbourhood controls, hoping to match controls to cases with regard to socioeconomic status, failed to do so. We would therefore recommend that, if this type of individual matching is used, the investigator should be clear about what s/he is trying to achieve by it, and try to collect data that will permit the confirmation or otherwise of the effectiveness of the matching.

ACKNOWLEDGEMENTS

The Diarrhoeal Diseases Control Programme of the World Health Organization provided financial support for the preparation of this document. The authors would like to thank the following people for their constructive comments on earlier drafts of this document: I. de Zoysa, J. Martines, M.H. Merson, N.F. Pierce, and P. Sandiford.

REFERENCES

Breslow, N.E. and Day, N.E. Statistical methods in cancer research: Volume 2. The analysis of case-control studies. IARC Scientific Publication No. 32, Lyon, 1980.

Briscoe, J., Feachem, R.G. and Rahaman, M.M. Measuring the impact of water supply and sanitation facilities: prospects for case-control methods. Unpublished document WHO/CWS/85.3, World Health Organization, Geneva, 1985.

Cousens, S.N., Feachem, R.G., Kirkwood, B.R., Mertens, T.E. and Smith, P.G. Case-control studies of childhood diarrhoea: I. Minimizing bias. Unpublished document WHO/EDP/88.2, World Health Organization, Geneva, 1988a.

Cousens, S.N., Feachem, R.G., Kirkwood, B.R., Mertens, T.E. and Smith, P.G. Case-control studies of childhood diarrhoea: II. Sample size. Unpublished document WHO/EDP/88.3, World Health Organization, Geneva, 1988b.

International Bank for Reconstruction and Development. Measurement of the health benefits of investments in water supply - Report of an Expert Panel, Public Utilities Department Report No. PUN 20, World Bank, Washington D.C., 1976.

Islam, S.S. and Khan, M.U. Risk factors for diarrhoeal deaths:a case-control study at a diarrhoeal disease hospital in Bangladesh. International Journal of Epidemiology, 15: 116-121, 1986.

Mantel, N. and Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22: 719-748, 1959.

Rodrigues, L. and Kirkwood, B.R. Case-control designs in the study of common diseases: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls (in preparation).

Rothman, K.J. Modern epidemiology. Little, Brown and Company, Boston, 1986.

Sandiford, P. A case-control study of environmental sanitation and childhood diarrhoea morbidity in rural Nicaragua. Report submitted in partial fulfillment of MSc degree, London University, 1988.

Schlesselman, J.J. Case-control studies, Oxford University Press, New York, 1982.

Shapiro, C., Cook, N., Evans, D., Willet, W., Fajardo, I., Koch-Weser, D., Bergonzoli, G., Bolanos, O., Guerrero, R., and Hennekens, C.H. A case-control study of BCG and childhood tuberculosis in Cali, Colombia. International Journal of Epidemiology, 14: 441-446, 1985.

Smith, P.G. and Day N.E. The design of case-control studies: the influence of confounding and interaction effects. International Journal of Epidemiology, 13: 356-365, 1984.

Smith, P.G. Evaluating interventions against tropical diseases. International Journal of Epidemiology, 16: 159-166, 1987.

Thompson, W.D. A study of matched and unmatched designs for case-control investigations of disease aetiology. Doctoral dissertation, Yale University, 1980.

Victora, C.G., Smith, P.G., Vaughan, J.P., Nobre, L.C., Lombardi, C., Teixeira, A.M.B., Fuchs, S.M.C., Moreira, L.B., Gigante, L.P., and Barros, F.C. Evidence for protection by breastfeeding against infant deaths from infectious diseases in Brazil. Lancet, ii: 319-323, 1987.

ANNEX

STATISTICAL FORMULAE

1. Unmatched analysis of a single 2 x 2 table

	Case	Control	
Exposed	a	b	r1
Unexposed	c	d	r2
	m1	m2	n

$$\text{Odds ratio} = \frac{a \times d}{b \times c}$$

$$X^2 = \frac{n \times [\,|axd - bxc| - 0.5xn\,]^2}{m1 \times m2 \times r1 \times r2}$$

The statistical significance of the observed association is found by comparing the value of X^2 with the percentage points of the chi-squared distribution with one degree of freedom. If X^2 is greater than 3.84 then the association is significant at the 5% level; if X^2 is greater than 6.63 then the association is significant at the 1% level.

2. Stratified analysis

The data have been divided into several strata, each of which may be represented in the form of a 2 x 2 table. The (i) indicates that this table represents the ith strata.

	Case	Control	
Exposed	a(i)	b(i)	r1(i)
Unexposed	c(i)	d(i)	r2(i)
	m1(i)	m2(i)	n(i)

$$\text{Mantel-Haenszel OR} = \frac{\frac{a(1)xd(1)}{n(1)} + \frac{a(2)xd(2)}{n(2)} + \ldots\ldots}{\frac{b(1)xc(1)}{n(1)} + \frac{b(2)xc(2)}{n(2)} + \ldots\ldots}\ldots$$

$$\text{Mantel-Haenszel } X^2 = \frac{N^2}{D}$$

where

$$N = \left|\frac{a(1)xd(1) - b(1)xc(1)}{n(1)} + \frac{a(2)xd(2) - b(2)xc(2)}{n(2)} + \ldots\ldots\right| - 0.5$$

and

$$D = \frac{m1(1) \times m2(1) \times r1(1) \times r2(1)}{n(1) \times n(1) \times [n(1)-1]} + \frac{m1(2) \times m2(2) \times r1(2) \times r2(2)}{n(2) \times n(2) \times [n(2)-1]} + \ldots\ldots$$

Annex

The statistical significance of the observed overall association, as estimated by the Mantel-Haenszel odds ratio, is found by comparing the value of the Mantel-Haenszel X^2 statistic with the percentage points of the chi-squared distribution with one degree of freedom. If X^2 is greater than 3.84, then the association is significant at the 5% level; if X^2 is greater than 6.63, then the association is significant at the 1% level.

3. Test-based confidence limits (Miettinen)

Test-based confidence limits for the odds ratio may be calculated from the odds ratio and the chi-squared statistic alone (see, for example, Schlesselman, 1982).

The variance of the natural log of the odds ratio is calculated as

$$VAR(\ln(OR)) = \ln(OR) / X^2$$

Then 95% confidence limits for the estimate of the odds ratio may be computed using the following formula:

$$\text{Upper and lower limits} = \exp([1 \pm 1.96 / X] \times \ln(OR))$$

where X is the square root of the chi-squared statistic.

4. Matched analysis of a 2 x 2 table

		Controls		
		Exposed	Unexposed	Total
Cases	Exposed	a	b	a+b
	Unexposed	c	d	c+d
	Total	a+c	b+d	n

$$\text{Odds ratio} = \frac{b}{d}$$

$$X^2 = \frac{(|b-d|-1)^2}{b+d} \quad \text{(McNemar's test)}$$

The statistical significance of the observed association is found by comparing the value of X^2 with the percentage points of the chi-squared distribution with one degree of freedom. If X^2 is greater than 3.84 then the association is significant at the 5% level; if X^2 is greater than 6.63, then the association is significant at the 1% level.

- - -