Acta Genet Med Gamellol 30:77-86 (1981)
The Mendel Institute/Alan R. Liss, Inc.

Review Article

Received 11 February 1981

Matching in Epidemiology as a Paradigm for Twin Research on the Etiology of Disease

Colin White

Yale University School of Medicine, New Haven, Connecticut

INTRODUCTION

This article is a review of the technique of matching as it applies to epidemiological investigations of the etiology of disease. It is written for researchers who wish to use twins as the matched pairs in studies of this type.

Key words: Biometry, Statistics, Twins, Epidemiologic methods, Follow-up studies, Sociometric technics

CASE-CONTROL AND COHORT STUDIES

To consider the role of matching in epidemiology, we must make certain distinctions about the design of the studies. Some of these begin with a group of persons who are exposed to a certain risk factor and a contrasting group of non-exposed persons, and the two groups are then followed to monitor the development of disease. This design is usually described as the cohort type, though the subjects may experience exposure at different periods and do not therefore often constitute a cohort in the strict sense of the term. Alternative names are prospective studies or follow-up studies.

A second type of study begins with a series of subjects who are recruited as they develop a particular disease. They are then contrasted, with respect to past exposures, with a comparison group that is free of the disease under study. This is usually called the case-control method; an alternative name is retrospective study, but this term can be confusing, since an exposed and non-exposed group on whom appropriate records are available might be defined as of some time in the past, and therefore studied in part retrospectively, even though the study method is cohort in its essential characteristics.

The distinction between the two kinds of investigation lies in the method of defining the study groups. It is this structure which dictates the distinctive type of estimates that can be made in these two instances – estimates of the incidence of disease in one case and of the percent who have suffered past exposure in the other. The common feature of the two methods is that, in either case, a test can be made of the hypothesis of interest, namely, whether there is an association between the exposure and the disease.

HISTORY OF ONE-TO-ONE MATCHING IN ETIOLOGICAL STUDIES

Methods of one-to-one matching were introduced into epidemiology in an informal way, so that there is no particular paper that can be said to be the first to propose matching as a useful methodological novelty. Matching was gradually incorporated into practice in the 1920s as one investigator followed another in taking what seemed to be commonsense precautions about selecting controls. All I shall venture to do about suggesting priority among investigators in the introduction of matching is to give a few instances of those who were among the early users of this technique.

It was natural that the first attempts at matching should take the form of selecting controls in a purposeful way so that the distribution of a matching variable in the control group was constrained to approximate the corresponding distribution for the cases; in particular, the means of the distributions, or the proportions with a certain characteristic, were made more or less identical. Lane-Claypon [5] did this in her study of breast cancer in 1926. Pearl [8], a few years earlier, had matched patients and controls on age with the objective of studying the longevity of the parents of the patients and controls. These parents were divided into eight groups according to sex, according to whether the parent was alive or dead, and according to whether the index case or control was alive or dead. The number in each of these eight groups of parents of patients agreed exactly with the number in the corresponding group of parents of controls. The principle of matching at this stage was apparently to select in each stratum as many control parents as there were parents of patients. It can therefore be described as matching by subcategories and can be thought of as a method that would eventually lead to one-to-one matching. The matching was elaborate enough to create confusion, since it was inappropriate to match the parents on living or dead status when the outcome variable was the length of life of the parents.

Individual, or one-to-one, matching of the type used when twins are involved, was clearly described a few years later. In 1928, Lombard and Doering [7] matched 217 cancer patients by having the investigator who collected the records of the patients fill out a similar record for an individual without cancer, "of the same sex and approximately the same age."

Sociologists became interested in matching about the same time as the epidemiologists, as is well documented by Chapin [1] and Greenwood [4]. An example discussed by Greenwood is the study of Sletto in 1934 of juvenile delinquency. The cases were 1,046 children who had been found delinquent by a county Juvenile Court. The controls were drawn from 12,168 school children from the same area and were matched to the cases one-to-one on age, sex, and sibship size.

MATCHING VERSUS BLOCKING

During the 1930s and subsequently, the use of matching spread rapidly. This movement may have been aided by the parallel development of experimental design under the influence of R. A. Fisher, a famous twin and, incidentally, an occasional researcher of twinning. The use of randomized blocks, which includes pairing as a special case, became common in agricultural experiments and began to interest workers in other fields. Pairing, as used in such a context, resembles the method of matching employed by epidemiologists; this is, however, a similarity that has caused confusion, since the difference between pairing and matching is important.

In experimental design, blocking comes before the treatment is imposed; this is not true in case-control studies, where the treatment has run its course prior to the matching; with the important exception of twins, it is not true in the usual type of cohort study, where the exposed and non-exposed are also matched some time after the exposure has begun. The relevant measurements on the matching variables are those that obtained at the time of exposure, and such measurements can be determined reliably only for variables that do not change over time; place of birth and date of birth, for example, meet this condition, but immunity to various risk factors may not do so.

A more serious point of distinction between pairing as used in experimental design and matching in an observational study is the difference in objectives. In experimental design, pairing has no role in ensuring that the estimate of treatment differences is valid. That objective is attained, by and large, by randomization, and the purpose of pairing is to reduce the variance of the estimate of treatment differences – that is, to increase the efficiency. In contrast to this, the primary objective of matching in observational studies is to attempt to make the treatment comparisons valid by removing from consideration certain alternative explanations of a treatment difference; for example, an effect cannot be due to differences in age if the members of the groups that are being compared have been closely matched one-to-one on age. Matching in the observational study, takes over, in a sense, the role of randomization rather than the role of blocking. Matching becomes the method by which the control group is selected, just as randomization is the basis for selection in the randomized block design.

In recent times, there has been a partial deviation from this point of view; and many papers have been written on the efficiency, or lack of efficiency, of matching. Efficiency is important in cases in which validity is not an issue, but validity has priority. The history of disputes over etiological factors is interesting in this regard. In a 1979 report of the United States Surgeon General [10], reference was made to eight major prospective epidemiological studies of the relationship between cigarette smoking and lung cancer, a topic that has also been studied in monozygous twins. Altogether, there were 1.2 million males and 0.7 million females included in these studies, and all of these were additional to more than 15,000 subjects investigated by the case-control method. Substantial differences between the exposed and non-exposed have been claimed, but there are still a few responsible investigators who do not accept the findings. The fact that studies of the same hypothesis have been repeated many times with a collective total of more than two million subjects implies that the validity rather than the efficiency of the studies has been questioned.

Since an important function of matching is to attempt to attain validity by controlling confounding, this latter topic will now be considered.

CONFOUNDING

Confounding literally means pouring together. When two factors are invariably associated, it is impossible to allocate any resulting effects to one rather than the other, and the factors are said to be confounded. If we have data showing that women aged 65 years have higher blood cholesterol than men of 45 years, the finding may be secure, but in attempting to explain it from those data alone, we cannot separate effects associated with differences in sex from those associated with differences in age. In the treatment of hypertension, hexamethonium was originally used in the form of a bromide, and some observers attributed its beneficial effect to the sedative action of the bromide rather than to the ganglionic-blocking action of the hexamethonium.

This phenomenon, in the modified form of partial rather than complete confounding, is so frequent in observational data of the type encountered in many epidemiological studies, and in everyday life, that common sense prompts us to look for alternative explanations of the observations we make.

To illustrate confounding in an introductory way, we can consider a study made of the admissions in 1973 to the Graduate Division of the University of California, Berkeley [3]. About 44% of the men but only 35% of the women were accepted, thereby providing prima facie evidence of a sex bias in the admissions policy. The six largest majors, however, show the following percentages of applicants admitted.

Major	A	B	C	D	E	F	All
Number of male applicants	825	560	325	417	191	373	2,691
Percentage of men admitted	62	63	37	33	28	6	44
Number of female applicants	108	25	593	375	393	341	1,835
Percentage of women admitted	82	68	34	35	24	7	30

When the data are considered separately for the various majors, it is clear that the percentage admitted is comparable for men and women, except for major A, where it is perhaps higher for women. The explanation of the misleading overall percentages is that over 50% of the men applied to majors A and B, which admitted a high proportion of applicants, and over 90% of the women applied to C, D, E, or F, in which there was a much smaller percentage of successful applications.

In simple terms, a confounding variable is one, the omission of which distorts the relation between two factors of interest. The relation between two binary variables may be conveniently described by means of the odds ratio. In a cohort type of etiologic study, the numerator of this ratio is the odds in favor of disease among those exposed to the risk factor, and the denominator is the odds in favor of disease among those not exposed to the risk factor. When the study is based on a comparison of cases and controls, the odds refer to presence of the risk factor rather than occurrence of the disease, but one of the valuable properties of the ratio is that it is the same for both types of design.

Let

p' = *proportion that develop disease among the exposed*

p = proportion that develop disease among the non-exposed

and

ψ = the odds ratio;

then

$$\psi = p'(1 - p)/p(1 - p').$$

The odds ratio will be used in the following account of confounding as the measure of the relation between disease and exposure. It is similar to the relative risk when the disease is rare. The latter, however, may change radically if one uses survival for example, rather than death, as an outcome, whereas the log odds ratio merely changes its sign.

The manner in which the relation between exposure and the occurrence of disease is altered by a confounding variable can be made more precise by considering a simplified model in which a population consisting of exposed and non-exposed persons is kept under surveillance for an appropriate period of follow-up. During this period, some develop the disease of interest and some do not. Each person is characterized by the presence or ab-

sence of a third factor that may influence the association between exposure and disease, and that will be denoted by a binary covariable Z, which may assume the value $z_1 = 0$, or $z_2 = 1$.

The data from such a study can be presented as counts falling into one of the eight cells of a 2 X 2 X 2 table. We may suppose initially that we are dealing with a complete population of interest, so that the fluctuations of sampling are not present. The population may be described in logical sequence by considering three steps: (a) the division into two strata, (b) the exposure to the risk factor in each of the strata, and (c) the occurrence of disease in each of the four groups thereby created. Thus, seven parameters are needed.

(a) r_1 = proportion in the stratum in which $Z = z_1$ (first stratum). The proportion in the second stratum in which $Z = z_2$ would be $(1 - r_1)$, and it will sometimes be convenient to give this proportion the symbol r_2.
(b) q_1 = proportion exposed in the first stratum.
q_2 = proportion exposed in the second stratum.
(c) p_1' = the proportion that develops the disease among the exposed in the first stratum. Expressed in terms of conditional probability $p_1' = P(D \mid E, z_1)$.

$p_1 = P(D \mid \bar{E}, z_1)$
$p_2' = P(D \mid E, z_2)$
$p_2 = P(D \mid \bar{E}, z_2)$

The proportions falling into each of the eight classes are shown in the Table. The odds ratio for the first stratum is given as $\psi(z_1)$ and for the second stratum as $\psi(z_2)$. If the covariable Z is ignored, the table is reduced to a 2 X 2 format; when this is done, the odds ratio for the resulting table is given by ψ.

Instances of confounding will first be illustrated. Suppose that exposure has no effect in either stratum; that is, $p_1' = p_1$ and $p_2' = p_2$, or expressed in another way, $\psi(z_1) = \psi(z_2) = 1$. We shall assume that p_1 differs from p_2; otherwise, we would have the situation

TABLE. Proportion of Subjects by Exposure Status, Disease Status, and Value of Covariable Z; and Odds Ratio When $Z = z_1$, $Z = z_2$, and $Z = z_1$ *or* z_2

	$Z = z_1$			
	Disease present	Disease absent	Total	
Exposed	$p_1'q_1r_1$	$(1 - p_1')q_1r_1$	q_1r_1	$\psi(z_1) = p_1'(1 - p_1)/(1 - p_1')p_1$
Not exposed	$p_1(1 - q_1)r_1$	$(1 - p_1)(1 - q_1)r_1$	$(1 - q_1)r_1$	
	$Z = z_2$			
	Disease present	Disease absent	Total	
Exposed	$p_2'q_2r_2$	$(1 - p_2')q_2r_2$	q_2r_2	$\psi(z_2) = p_2'(1 - p_2)/(1 - p_2')p_2$
Not exposed	$p_2(1 - q_2)r_2$	$(1 - p_2)(1 - q_2)r_2$	$(1 - q_2)r_2$	

The odds ratio, ψ, when Z is ignored is as follows:

$$\psi = \frac{[p_1'q_1r_1 + p_2'q_2r_2][(1 - p_1)(1 - q_1)r_1 + (1 - p_2)(1 - q_2)r_2]}{[(1 - p_1')q_1r_1 + (1 - p_2')q_2r_2][p_1(1 - q_1)r_1 + p_2(1 - q_2)r_2]}.$$

https://doi.org/10.1017/S0001566000006644 Published online by Cambridge University Press

in which neither stratum nor exposure affected the incidence of disease, and in such a case there would be no possibility of confounding. When the incidence proportions, p_1 and p_2, do differ in the two strata, however, it is possible for ψ to differ from 1.

The condition for $\psi > 1$ is that the stratum with the higher probability of disease is the one with the higher proportion of subjects exposed. This may be established by taking the expression for ψ in Table 1, substituting $p_1' = p_1$ and $p_2' = p_2$, and finding the conditions for $\psi > 1$. After algebraic simplification of the expression for ψ, one finds that the condition is that $(p_1 - p_2)(q_1 - q_2) > 0$. This implies that $(p_1 - p_2)$ and $(q_1 - q_2)$ have the same sign; that is, the stratum with the higher baseline incidence of disease must also have the higher proportion exposed.

This is the kind of confounding that encourages the view that a certain exposure leads to the disease when in fact it does not. When the data from the individual strata show that exposure has no effect, but an analysis that ignores the stratification appears to show otherwise, the confounding is referred to as Simpson's paradox. The data on admission to Berkeley Graduate School provide an example.

The name positive confounding has been used in the general case in which $\psi > \psi(z_1) = \psi(z_2)$; that is, ignoring Z leads to a higher estimate of the odds ratio. A typical case occurs when one is trying to identify a relatively weak risk factor in the presence of a factor that is both strong and common. The following provides a numerical illustration.

	$Z = z_1$			$Z = z_2$			Z is ignored	
	D	$\bar{D}$		D	$\bar{D}$		D	$\bar{D}$
E	20	110	E	185	44	E	205	154
$\bar{E}$	8	111	$\bar{E}$	20	12	$\bar{E}$	28	123
	$\psi(z_1) = 2.52$			$\psi(z_2)$ 2.52			$\psi = 5.85$	

A possible illustration would be given by the testing of occupational exposure to a chemical as a cause of lung cancer. The variable Z would identify cigarette smokers or non-smokers.

A modification of the result given above deals with negative confounding; that is, the case in which $\psi(z_1) = \psi(z_2) > \psi$. This type of confounding will occur when the stratum with the higher risk of disease has the lower exposure. A numerical illustration follows.

	$Z = z_1$			$Z = z_2$			Z is ignored	
	D	$\bar{D}$		D	$\bar{D}$		D	$\bar{D}$
E	20	110	E	12	20	E	32	130
$\bar{E}$	2	111	$\bar{E}$	11	185	$\bar{E}$	13	296
	$\psi(z_1) = 10.09$			$\psi(z_2) = 10.09$			$\psi = 5.60$	

To emphasize the role played by difference in the relative size of the strata and difference in the proportion exposed in producing confounding, one may examine the formula for ψ when such differences do not exist; that is, when $r_1 = r_2 = 0.5$ and $q_1 = q_2$. In such a case

$$\psi = [2/(p_1 + p_2) - 1]/[2/(p_1' + p_2') - 1]$$

When, as is often true, p_1', p_2', p_1, and p_2 are all small, ψ approximates $(p_1' + p_2')/(p_1 + p_2)$, which implies that ψ lies between p_1'/p_1 and p_2'/p_2; since these relative risks for the two strata are approximately equal to the respective odds ratios for the strata, ψ, under the

given assumptions, would lie between the two odds ratios. This would also ensure that when the two strata had the same odds ratio, this common value would agree with ψ within the limits imposed by the approximation described; in other words, there would be virtually no confounding.

One special case of negative confounding that is worthy of comment is shown in the following table:

	$Z = z_1$			$Z = z_2$		
	D	$\bar{D}$		D	$\bar{D}$	
E	a	2a–k	E	a	k	$a > k > 1$
$\bar{E}$	k	a	$\bar{E}$	2a–k	a	

The odds ratio is the same in the two strata and is greater than 1. If z is ignored, the odds ratio becomes exactly 1. The negative confounding fits into the pattern described above in that the first stratum shows the higher proportion exposed, but the lower proportions developing the disease in the non-exposed group.

RELATIONSHIP BETWEEN EXPOSURE, DISEASE, AND CONFOUNDER

If the data from an example of confounding are studied, it will be noted that the covariable is related to the exposure and to the disease. Moreover, these associations are not necessarily with the marginal distributions of exposure and disease status. Instead, we can assert the following: (a) Association between the covariable and disease is conditional on exposure status; that is, the covariable is dependent on disease status within the exposed group and also within the non-exposed group. (b) Association between the covariable and exposure status is conditional on disease status; that is, the covariable is dependent on exposure status within the diseased group and also within the nondiseased group.

Consider, for example, the following data previously presented:

	z_1			z_2	
	D	$\bar{D}$		D	$\bar{D}$
E	20	110	E	185	44
$\bar{E}$	8	111	$\bar{E}$	20	12

A simple illustration of the association between covariable and disease is obtained by noting that the odds ratio in the following tables differs from 1:

(a) Association conditional on exposure

	$D\|E$	$\bar{D}\|E$
z_1	20	110
z_2	185	44

Odds ratio = 0.043

(b) Association conditional on non-exposure

	$D\|\bar{E}$	$\bar{D}\|\bar{E}$
z_1	8	111
z_2	20	12

Odds ratio = 0.043

It is not required that these odds ratios be the same; only that they differ from 1.

The corresponding illustration that the covariable and the exposure are associated, conditional on the disease status, is obtained from the following 2 × 2 tables.

https://doi.org/10.1017/S0001566000006644 Published online by Cambridge University Press

(a) Association conditional on disease status

	$E \mid D$	$\bar{E} \mid D$
z_1	20	8
z_2	185	20

Odds ratio = 0.270

(b) Association conditional on non-disease status

	$E \mid \bar{D}$	$\bar{E} \mid \bar{D}$
z_1	110	111
z_2	44	12

Odds ratio = 0.270

It is commonly stated that the association of the covariable with exposure should be based on the marginal distribution of exposure and not on the distribution conditional on disease status. The following counter-example shows a case in which the covariable is independent of the marginal distribution of exposure, but is still involved in confounding.

	z_1			z_2	
	D	$\bar{D}$		D	$\bar{D}$
E	100	9,900	E	925	9,075
$\bar{E}$	10	9,990	$\bar{E}$	100	9,900

Confounding is present in this case in that the odds ratio in each stratum is 10.09, but the odds ratio in the collapsed table is 9.77. The covariable, however, is not dependent on the marginal distribution of exposure. It is, however, dependent on the distribution of exposure, conditional on disease status.

PREVENTION OR CONTROL OF CONFOUNDING

What has been written so far deals mainly with identification of confounding when the appropriate multivariate data are on hand. Sometimes confounding can be anticipated at the stage of planning. This must be done to the extent, at least, of providing guidance as to the variables that need to be included in the data. Occasionally, background knowledge gives a reasonably clear indication of what is needed. Suppose a study is to be made of the use of oral contraceptives as a risk factor for venous thrombo-embolism, and we wish to consider the variable parity as a potential confounding factor. It is related to venous thrombo-embolism, since repeated pregnancies predispose to the condition. Parity would be expected to have this relation whether the subject uses oral contraceptives or not, that is, parity is presumably related to the disease conditional on exposure. It is also related to the use of oral contraceptives, since regular users are likely to have fewer children and this would presumably be independent of the eventual disease status. Parity could therefore induce an association between the exposure and the disease even if there were no association at each separate parity level. If parity were ignored, there would be a risk of negative confounding, that is under-estimating any positive association between use of oral contraceptives and thrombo-embolism, or even, perhaps, suggesting that oral contraceptives protected against the disease.

It is true in general, however, that confounding cannot usually be removed by measures taken in the design of a study; much must be left to the stage of analysis. The danger in relation to confounding at the planning stage is that a relevant variable might simply be ignored. If the data are obtained, methods of analysis are available to control the effects of confounding.

A false step that might be taken at the stage of analysis is to treat as a confounding factor something that is an intermediary between the exposure and the disease. To control this intermediary is to remove the means by which the exposure leads to the disease. If, for example, a certain diet leads to hypercholesterolemia, which in turn leads to disease, then it would be misleading to match on blood cholesterol level or to control it in the analysis. It is sometimes claimed that biological insight is needed to detect this problem. To the extent that this is true, the insight comes from previous inspection of data. Unfortunately, in many epidemiological studies emphasis has to be, or at any rate is, placed on an end-point. In cohort studies, it is possible to observe intermediate points as well; and more opportunity should be taken to do this.

MATCHING BY MEANS OF TWINS

Matching in epidemiological studies is no longer considered the key to the control of confounding. At the stage where methods of data analysis had reached a point of providing an alternative to matching, a literature arose in which the latter was criticized as curbing the possibility of a full analysis of all the data relating to the matching factors themselves. The interactions of these factors with unmatched factors could be studied from paired data, but the main effect of the factors themselves could not be investigated. A certain amount of group matching went on inevitably, even when it was not specifically acknowledged, since it would be pointless to attempt to compare two groups with widely different distributions of factors such as age and sex.

The climate of criticism in relation to matching encouraged the discussion of issues such as over-matching [9]. Questions were also raised about the expense and trouble of matching. There were analytical difficulties too, in that an unmatched factor might come to be considered important during the course of a long study, and the investigator was then faced with an analysis that involved a mixture of matched and unmatched factors.

These analytical difficulties have been overcome [11], and the positive aspects of matching are once more relevant. The inefficiency that results from data that are imbalanced with respect to confounding variables can be greatly reduced. The issue of over-matching must, however, be considered.

Over-matching arises in a case-control study when the matching factor is strongly associated with the exposure. The cases who have been exposed will mostly have the matching factor because of the strong association between the two. When the cases are positive for the matching factor, the matched controls must also be, since this is the meaning of matching. This leads to an aggregation of matched pairs of the type in which case and control are concordant for exposure; but the McNemar test for association in matched studies is based only on the discordant pairs; and since these are scarce, the test of association, though valid, will have low efficiency.

These considerations do not apply in a cohort-type study where the matching is between exposed and non-exposed. All studies in which the matching is based on twins [2] are cohort in design, even when they are partly retrospective in timing. There may certainly be difficulty in obtaining a large enough group of twins to provide pairs with an exposed and non-exposed member, but when that can be achieved, the ensuing analysis does not involve the problem of over-matching or, indeed, any special difficulty resulting from the nature of the matching.

REFERENCES

1. Chapin FS (1947): "Experimental Designs in Sociological Research." New York: Harper and Brothers.
2. Cederlof R, Friberg L, Lundman T (1977): The interactions of smoking, environment and heredity and their implications for disease etiology. A report of epidemiological studies on the Swedish Twin Registries. Acta Med Scand Suppl 612:1–128.
3. Freedman D, Pisani R, Purves R (1978): "Statistics." W. W. Norton and Co., New York.
4. Greenwood E (1945): "Experimental Sociology: A Study in Method." New York: King's Crown Press.
5. Lane-Claypon JE (1926): A further report on cancer of the breast. Reports on Public Health and medical subjects. No. 32, Ministry of Health. London: His Majesty's Stationery Office.
6. Locket S, Swann PG, Grieve WSM (1951): Methonium compounds in the treatment of hypertension. Br Med J 1:778–784.
7. Lombard H, Doering C (1928): Cancer studies in Massachusetts. 2. Habits, characteristics and environment of individuals with and without cancer. N Engl J Med 198:481–487.
8. Pearl R (1923): The age at death of the parents of the tuberculous and the cancerous. Am J Hyg 3: 71:89.
9. Samuels ML (1979): How case-matching can reduce design efficiency in retrospective studies. Technical Report No. 50, Divisoin of Biostatistics, Stanford University.
10. U.S. Public Health Service (1979): Smoking and health: A report of the Surgeon General U.S. Department of Health, Education and Welfare. Public Health Service, DHEW Publication No. 79-50066.
11. Yanagawa T (1979): Designing case-control studies. Environ Health Perspect 32:143–156.

Correspondence: Professor Colin White, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06510, USA.