

This and the following paper also say "stop, look, and listen" to the researcher in any field of public health who may be less than adequately prepared in statistical method. All three papers are concerned, in general, with the deceptively simple matter of choosing controls, this one marking the pitfalls inherent in pairing or matching.

Matching in Analytical Studies*†

WILLIAM G. COCHRAN

*Professor of Biostatistics, School of Hygiene and Public Health,
Johns Hopkins University, Baltimore, Md.*

MOST of the following discussion will be confined to studies in which we compare two populations, which will be called the experimental population and the control population. The experimental population possesses some characteristic (called the experimental factor) the effects of which we wish to investigate: It may consist, for example, of premature infants, of physically handicapped men, of families living in public housing, or of inhabitants of an urban area subject to smoke pollution, the experimental factors being, respectively, prematurity, physical handicaps, public housing, and smoke pollution. I shall suppose that we cannot create the experimental population, but must take it as we find it, except that there may be a choice among several populations that are available for study.

The purpose of the control population is to serve as a standard of comparison by which the effects of the experimental factor are judged. The control population must lack this factor, and ideally it should be similar to the experimental population in anything else that might affect the criterion variables by which

the effects of the factor are measured. Occasionally, an ideal control population can be found, but, more usually, even the most suitable control population will still differ from the experimental population in certain properties which are known or suspected to have some correlation with the criterion variables.

When the control and experimental populations have been determined, the only further resource at our disposal is the selection of the control and experimental *samples* which are to form the basis of the investigation. Sometimes this choice is restricted, because the available experimental population is so small that it is necessary to include all its members, only the control population being sampled.

The problem is to conduct the sampling and the statistical analysis of the results so that any consistent differences which appear between the experimental and the control samples can be ascribed with reasonable confidence to the effects of the factor under investigation.

The first step in any matching process is to select those variables (called the covariables) on which the two samples are to be matched. I shall assume for the moment that this decision has been made; the principles in-

* Presented before the Statistics Section of the American Public Health Association at the Eightieth Annual Meeting in Cleveland, Ohio, October 23, 1952.

† Paper No. 288, Department of Biostatistics. Some of the theoretical results used in this paper were obtained under a research contract with the Office of Naval Research.

volved in the decision will be discussed briefly later.

PAIRING

Matching of the experimental and control samples with respect to the covariables can be accomplished in a number of ways. Conceptually, the simplest is the method of pairing. Each member of the experimental sample is taken in turn, and a partner is sought from the control population which has the same values as the experimental member (within defined limits) for each of the covariables. One way of doing this is to perform a multiple classifica-

advantages of pairing and of covariance analysis are usually demonstrated by means of a linear regression model. I shall present this analysis, but, as will be seen, there is reason to doubt whether the assumptions in the analysis are valid for many of the studies conducted in practice.

Let y denote the variable by which the effects of the experimental factor are measured, and x denote the covariable, assuming for simplicity that there is only one. The model assumes that y has a linear regression on x with the same slope β in each population. The equations are as follows:

$$\begin{array}{l} \text{Experimental population: } y = a + \beta x + d \\ \text{Control population : } y' = a' + \beta x' + d' \end{array} \quad \begin{array}{l} (1) \\ (2) \end{array}$$

tion of the control population by the variables. We then examine the first member of the experimental sample, pick the cell which contains all control members having the desired set of covariables, and choose as the partner one control member at random from this cell. This procedure is repeated for each member of the experimental sample.

If an occasional cell is found to be empty, it is usually preferable to choose the control partner from a neighboring cell, rather than to omit the experimental member. If numerous cells are found to be empty, this is a danger signal. Either the limits of variation allowed in the covariables are too narrow or the control population is not satisfactory.

The analysis of the results is very simple. The difference (experimental-control) is computed for each pair, and any t-tests are applied directly to this series of differences.

EFFECTIVENESS OF PAIRING WHEN WE HAVE AN IDEAL CONTROL

It is difficult to discuss the effectiveness of pairing in realistic terms. The

The variables x and d are independently distributed and the deviations d, d' have means zero in both populations. Further, it is assumed that the means \bar{X}, \bar{X}' of x in the two populations are equal, and that $(a - a')$ represents the true effect of the experimental factor, i.e., that no unsuspected biases are present.

In effect, this model postulates that we have been successful in finding an *ideal* control population, since the relation between y and x is the same in both populations and since x has the same average value in both populations.

With this model, the precision given by paired samples can be compared with that given by independent random samples drawn from the two populations. In either method, the effect of the experimental factor will be estimated by the difference $(\bar{y} - \bar{y}')$ between the means of the two samples. For the independent samples, each of size n , the variance V_1 of $(\bar{y} - \bar{y}')$ is

$$V_1 = \frac{2}{n} \sigma_y^2 \quad (3)$$

assuming for simplicity that σ_y is the same in both populations.

With samples paired for x , on the

TABLE 1
Values of $(1 - R^2)$

R	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$(1 - R^2)$	0.96	0.91	0.84	0.75	0.64	0.51	0.36	0.19

other hand, it follows from equations (1) and (2) that

$$\bar{y} - \bar{y}' = (a - a') + (\bar{d} - \bar{d}')$$

so that the variance V_p of this difference is

$$V_p = \frac{2}{n} \sigma_d^2, \quad (4)$$

This result may be expressed in a more useful form. From (1),

$$\sigma_y^2 = \beta^2 \sigma_x^2 + \sigma_d^2, \quad (5)$$

since x and d are assumed independent. If ρ is the correlation coefficient between y and x , then $\beta\sigma_x = \rho\sigma_y$. Thus (5) becomes

$$\sigma_y^2 = \rho^2 \sigma_y^2 + \sigma_d^2$$

giving a well-known result in theory,

$$\sigma_d^2 = \sigma_y^2 (1 - \rho^2)$$

Hence, finally, the variance of $(\bar{y} - \bar{y}')$ for the paired samples may be written, from (4)

$$V_p = \frac{2}{n} \sigma_y^2 (1 - \rho^2) \quad (6)$$

Comparison with (3) shows that the ratio of V_p to V_i is $(1 - \rho^2)$. A more concrete way of expressing this result is as follows. If n_p, n_i are the respective sample sizes which make the variances equal for paired and independent samples, then $n_p = n_i (1 - \rho^2)$. Thus, the ratio $(1 - \rho^2)$ shows how much we can afford to reduce the sample size, when pairing is employed, without any loss of precision.

If pairing is accomplished for several x -variables, all linearly related to y , the ratio becomes $(1 - R^2)$ where R is the

multiple correlation coefficient between y and the x 's. Values of this factor are shown in Table 1.

The reductions in variance are not large until R reaches 0.5. A reduction by one-half, which corresponds to an allowable reduction of the sample size by one-half, requires $R = 0.7$. Unfortunately, although many published papers have discussed the association between morbidity or mortality and such covariables as age, sex, economic status, family size, and degree of overcrowding, the associations tend to be described in general terms, and little information is available about the actual sizes of the correlations that are to be expected in field studies in public health. With some obvious exceptions (e.g., the association between chronic disease and age) my impression is that the multiple correlation coefficient is often below 0.5, so that the gain in precision from pairing is often modest.

SELECTION OF THE COVARIABLES

Table 1 is also relevant in the selection of the covariables on which the pairing is to be based. It is nearly always possible, with a little effort, to produce a substantial number of x -variables that *might* have some association with y , and this list must be reduced to a few x -variables which will actually be used in pairing. It follows from Table 1 that inclusion of a specific x -variable in the pairing is worth-while only if it decreases $(1 - R^2)$ by an appreciable amount (say 10 per cent), when this x -variable is added to those already selected. Although the practical use of this result requires an intimate knowledge of the relation between y and the x 's which is rarely possessed, the

result indicates that associations in which the correlation between a covariable and y is of the order of 0.1 — 0.3 are unlikely to produce useful gains in precision. Thus, in deciding whether to include a covariable, the important question is not “Is there an association?” but “How large is the correlation coefficient?”

PAIRING VERSUS RANDOM SAMPLES WITH COVARIANCE

If, instead of pairing, we draw random samples of size n from each population and adjust the sample means by covariance, then, on the average,

$$V(\bar{y}_{adj} - \bar{y}'_{adj}) = \frac{2}{n} \frac{2}{y} \sigma^2 (1 - \rho^2) \left\{ 1 + \frac{1}{2(n-2)} \right\} \quad (7) *$$

The term in curly brackets represents an increase in variance due to errors in the covariance adjustment. If the covariance adjustment is made for k x -variates, by means of a multiple regression, this term becomes

$$1 + \frac{k}{(2n-k-3)}$$

Provided that the sample size n exceeds $10k$, this factor is close to unity, and covariance gives about the same precision as pairing.

In these circumstances there is not much to choose between pairing and covariance. Pairing has the advantage that the computations are simpler, particularly if the samples are paired on several x -variables. If the regression is nonlinear, pairing will give higher precision than covariance, unless the presence of nonlinearity is recognized in the covariance analysis and we go to

the trouble of fitting the appropriate type of regression curve. A difficulty which I have occasionally encountered with covariance is that some scientists have an inborn suspicion of adjustments to the data, and although the adjustments made in the covariance analysis are entirely objective, they may find a rather grudging acceptance.

Pairing has some limitations and disadvantages, the importance of which varies with the type of study. Pairing requires that data on the values of the covariables in the control population be readily accessible; this may not be the case. One disadvantage is the time

spent in constructing the pairs. If the experimental sample is small and the control population is large, or if the experimental sample becomes available one member at a time (as with newborn infants or admissions to a hospital) the pairing may be accomplished easily, but it can become tedious if the samples are large and it may impede the progress of the study. A small trial to estimate the time involved in pairing is sometimes advisable. If considerable attrition of the data is expected, as in a long-term follow-up study, the symmetry of pairing is lost. The simplicity of the analysis can be retained only by dropping all partners of “missing” sample members, which involves a loss of information. In order to avoid this loss, it is necessary to use a covariance analysis.

There is one further situation in which pairing may be highly effective. In some studies the experimental population has been drawn from some larger population by a mechanism which operates solely to select certain values of the covariables. For instance, suppose that the experimental population consists of families in public housing in a large city

* This result, which is a slightly different form of Dr. Greenberg's result, assumes that the x -variable is normally distributed in the population, but is approximately true even if x is not normal.

and that these families were selected from a larger population of approved applicants for public housing by some administrative rules. Suppose that the rules give preference to families of veterans and to families of certain sizes (since public housing is built with some preassigned and not necessarily average distribution of family sizes in mind), but are otherwise on a "first come, first served" basis. In this case the approved veteran applicants who are still waiting might constitute a good control population, except that the control and study samples need pairing or matching on family size. (It is not claimed that the selection of entrants to public housing actually operates in this way, the example being intended purely for illustration.)

In a situation of this kind, where x represents family size, the previous regression model might still apply, except that the population mean \bar{X} , \bar{X}' now differ. As a result, some difficulty may be experienced in finding control partners for the experimental sample, but if the pairs can be constructed, equation (6) continues to hold for the variance of $(\bar{y} - \bar{y}')$ in the paired samples.

With the covariance method, the corresponding variance may be shown to be approximately

$$V(\bar{y}_{adj} - \bar{y}'_{adj}) = \frac{2}{n} \frac{2}{y} \sigma^2 (1 - \rho^2) \left\{ 1 + \frac{1}{2(n-2)} + \frac{nD^2}{4(n-2)\sigma_x^2} \right\}$$

where $D = (\bar{X} - \bar{X}')$. The extra term involving D^2 appears because the covariance adjustment, $b(\bar{x} - \bar{x}')$, has become larger, since \bar{x} and \bar{x}' no longer have the same population means. The term in D^2 is almost independent of the sample size n . For $n > 20$, the variance for the covariance method is approximately

$$1 + \frac{D^2}{4\sigma_x^2} \tag{8}$$

times as large as that given by pairing, so that pairing is more efficient than covariance. The increase in precision from pairing relative to covariance is probably not great in practice. For instance, the variance ratio in (8) equals 1.25 when $D = \sigma_x$, that is when the population means \bar{X} , \bar{X}' are distant one standard deviation. This implies a fairly drastic selection operating on the x variable.

If the experimental population involves selection on several x variables, pairing removes the disturbing effects of this selection, provided that *all* the x variables are included in the pairing.

OTHER METHODS OF MATCHING

As we have seen, pairing is most easily done when one of the populations (usually the control population) is large and the samples are small. If the samples are large, pairing may be time consuming, and if the available populations are not much larger than the desired size of samples, pairing may be impossible. In these circumstances, methods of matching which are less thorough deserve consideration.

In the technic known as *balancing*, we do not pair individually, but select the control sample so that its means agree with those of the experimental sample for each of the covariables. Balancing can usually be carried out more quickly than pairing; it gives the same precision as pairing if the regression of y on the covariables is linear, although balancing is less precise if the regression

is nonlinear. As Dr. Greenberg states in his article, balancing requires the use of a covariance analysis in order to perform tests of significance—a point which has often been overlooked.

Another method of obtaining a less rigorous matching is to divide the range of any covariable into three or four classes. Thus, if matching is being done for three covariables, the number of cells produced will lie somewhere between 27 and 64 (some of which may empty). The sample from any cell in the experimental population is drawn at random, and its size is proportional to the number of entries in the cell. The sample from any cell in the control sample is also drawn at random and is made to be the same size as that for the corresponding cell in the experimental sample. If International Business Machines' equipment is used and the samples are large, this method, sometimes called *stratified matching*, is more expeditious than pairing, because it requires a much less detailed breakdown of the population into cells. It is less precise than pairing, since the covariables do not have exactly the same set of values in the control and experimental samples. However, with at least three classes for any covariable, it may be shown by theory that the loss precision is small unless the multiple correlation coefficient exceeds about 0.7. The analysis of the results is slightly more complicated, because we can compare the experimental and control results separately for each cell. In return, this analysis focuses attention on any variation that occurs in the effects of the experimental factor as the levels of the covariables change—in other words, on the interactions of the experimental factor with the covariables. Such information may broaden the results of the study.

There are several variants of this method. For instance, if there is particular interest in interactions, the sam-

ples from each cell may be made equal so far as is feasible.

The experimental and control samples need not be the same size. With pairing, we could select r control partners for each member of the experimental sample. The factor $2/n$ in the variance of the mean difference is then reduced to $(r + 1)/rn$. The cost of the study is, of course, increased, but the device may be profitable when the experimental sample is small and the cost of obtaining and processing the control data is not prohibitive.

SITUATION WHEN THE CONTROL POPULATION IS NOT IDEAL

It is worth reiterating that the previous discussion of the effectiveness of matching or covariance assumes that the control population is ideal, in the sense that the control and experimental populations differ at most through selection on certain covariables which are included in the matching or covariance.

When the two populations differ in other ways, we do not know how effective matching is. It is almost certain to be less effective than when the control population is ideal, and it may be practically ineffective. What is likely to happen is that the regressions of y on the covariables will differ in the two populations. In this event, matching no longer removes all the disturbing effects of the covariables on which we match. Further, there are likely to be other variables with respect to which the populations differ. In so far as these variables are uncorrelated with the matching covariables, their disturbing effects are unchanged by matching. The net result is that the difference between the means of the matched samples is a biased estimate of the effects of the experimental factor. The bias can be expected to be smaller with matched than with unmatched samples, but it may be only slightly smaller.

DISCUSSION AND SUMMARY

The principal conclusion from the preceding discussion is that the selection of the control population is a more crucial step than the selection of a method of matching or of the covariables on which to match. Matching removes the deficiencies of a poor control to only a limited extent.

In turn, this suggests that, whenever feasible, any study should start with a comparison of the experimental and control populations. This is by no means the rule in practice. Frequently, the experimental sample is chosen first and the control population is then searched in order to find the partners. If partners seem hard to find, some misgivings about the control population may arise, but if the pairs can be found, sometimes by selecting an extreme sample from the control population, the study proceeds. This kind of matching may leave us with a control population that does not resemble the study population and a control sample that is a very extreme sample from the control population.

The kind of comparison which I am recommending was conducted by Densen, *et al.*,¹ as a preliminary to a proposed study on the penicillin treatment of cardiovascular syphilis. From hospital records, it was planned to select a sample of patients who had not received penicillin as a control for an experimental sample of patients who had been treated with penicillin. In a comparison of the available patients from two hospitals, Densen, *et al.*, found many differences between the experimental and control cases and concluded that any matching process would be hard to carry out and that its results would be suspect.

A comparison of this kind does not provide proof that the control population is ideal, since only the covariables can be studied, but not their relations with y . If, however, a number of covariables are included, it is reassur-

ing to find agreement, or at most minor disagreement, between the two populations on all these covariables. Where more substantial discrepancies appear, their implications as to the suitability of the control can be considered.

If the control population appears satisfactory, the next step is to select the covariables to be used in matching. At this point any available information about the sizes of the correlations between y and the x 's is relevant. Since my impression is that these correlations are frequently low in public health field studies, with some obvious exceptions, my recommendation, in cases of doubt, is to omit an x -variable from the matching rather than to include it. Covariance adjustments can be made should this variable later prove important.

If the samples are small and at least one of the populations is large, so that pairing is not too laborious, pairing has much to recommend it. If the samples are large, stratified matching may be preferable. If the samples are not much smaller than the available populations, random samples should be drawn and covariance adjustments applied.

In conclusion, in observational studies, where it is not feasible to assign a subject at random to the control or experimental sample, we can never be sure that some unsuspected disturbance does not account, in large part, for the observed difference between the two samples. Consequently, the results of tests of significance must be interpreted with more caution in observational studies than in experiments where randomization can be employed. One good practice in observational studies is to check any theory at as many points and in as many ways as ingenuity can devise. An illustration occurs in a study of the relation between inoculation with pertussis vaccine and poliomyelitis, reported by Hill and Knowelden.² The experimental sample consisted of children with poliomyelitis and the control sample of

children without poliomyelitis, paired for age and sex and living in the same area. The experimental sample showed a marked excess of inoculations during the month preceding onset of poliomyelitis, but no excess of inoculations at intervals larger than a month. Second, during the month preceding onset, paralysis occurred at the site of inoculation in the great majority of cases, but with the interval larger than a month, the site of inoculation was involved in paralysis in only a small minority of cases. The point is that these two independent results greatly strengthen the evidence for a causal relationship. If

the same kind of result appears repeatedly when the data are analyzed from widely different points of view, it becomes successively more difficult to imagine any "disturbance" that will explain away *all* the results. Where it can be employed, this technic does much to overcome the handicap under which we all labor in observational studies.

REFERENCES

1. Densen, P. M., et al. Studies in Cardiovascular Syphilis II. Methodologic Problems in the Evaluation of Therapy. *Am. J. Syph., Gonorr. & Ven. Dis.* 36, 1:64-76 (Jan.), 1952.
2. Hill, A. Bradford, and Knowelden, J. Inoculation and Poliomyelitis. A Statistical Investigation in England and Wales in 1949. *Brit. M. J.* ii:1 (July), 1950.

The Journal 25 Years Ago

TERMINAL DISINFECTION

Terminal disinfection which was discontinued by Charles V. Chapin, M.D., for certain communicable diseases in the City of Providence as early as 1905, was nevertheless still inveighed against in the *Journal* nearly a quarter of a century ago. Inspired by a survey and report on disinfection as practiced in various countries by Dr. Carlos Chagas of the Office International d'Hygiene Publique, an editorial takes pride in the fact that "as far as our knowledge goes, Dr. Chapin preceded all others in these ideas and in the practical demonstration of their soundness; but we welcome the clear presentation given by Dr. Chagas in corroboration of what we have believed and practiced for a number of years."

Dr. Chagas's report had indicated that terminal disinfection was almost always useless, based on incorrect ideas, and the results not commensurate with the expense. Conclusions reached, the editor says, by Dr. Chapin and others many years ago. (*A.J.P.H.* 18, 9:1132 (Sept.), 1928.)