MATCHED SAMPLES IN MEDICAL INVESTIGATIONS

BY

W. Z. BILLEWICZ

Research Group in Biometric Medicine and Obstetric Medicine Research Unit (M.R.C.), University of Aberdeen

In investigating the effects of a treatment or experience on a group of patients, it is usually necessary to observe a control group. The comparison between the two groups may be influenced by the characteristics of the subjects: for example, their sex, age, body weight, and original clinical state. In order to secure comparability for such relevant "extraneous" factors, the controls are frequently matched with the experimental subjects. An alternative is to select the controls at random from an appropriate population, and to rely on statistical analysis to make allowance for the effects of concomitant variables.

Medical research workers often prefer to select matched instead of randomly selected controls. The usual reasons seem to be that:

(a) Matching improves the clarity of the experimental situation and simplifies analysis and exposition. A state of "other things being equal" is built into the design, and no special statistical adjustments are required during the analysis of results.

(b) It protects the investigator against "freak" samples. With small numbers, random sampling may yield treatment and control groups which are highly dissimilar in composition.

(c) It improves the precision of the comparison by reducing variability of response due to the effect of extraneous factors.

Little has been written about the theory of matching procedures in medical research, and practice is often faulty. In twenty recently published medical investigations (not necessarily a representative selection), the number of matching variables ranged from one to six, two or three being the commonest number. In four of the examples, the original matching criteria had to be relaxed in order to complete the investigations. The method of statistical analysis appropriate to a study using matched subjects differs from that for an unmatched study; though this is made clear by most statistical texts, nine studies had incorrect analyses.

In the present investigation, two questions were examined. First, does matching possess the advantages claimed, and if so, under what conditions? Second, what are the administrative costs and practical aspects of matching procedures? These questions were studied empirically by simulating appropriate experimental situations on a computer. The procedure will be described briefly later on, but first some concepts and definitions will be discussed.

CONCEPTS AND DEFINITIONS

In the type of investigation under consideration, the treatment group has an experience to which the other, the control group, is not subjected. The differential experience may either (a) be imposed by the investigator (as in a drug trial), or (b) have taken place outside the investigator's control (for example, asphyxia in babies). Its effect is to be assessed in terms of the difference between the treatment and the control group with respect to some relevant response variable, e.g. blood pressure, degree of pain relief, or frequency of patients discharged as fit. The response variable should not be confused with the treatment effect, i.e. the difference of response between comparable treated and untreated subjects, which may be zero.

In medical journals two procedures are labelled as "matched samples". In the first, the experimental and control subjects are selected in pairs. In this, the matched pair method, each pair may be regarded as forming a separate small experiment in its own right. In the second procedure, frequency matching, the
identity of pairs is not preserved, but the composition of the experimental and control groups as a whole is so manipulated that the distribution of sampling units between the matching categories is the same in each group. The experiment is regarded as having been performed on two comparable groups rather than on several comparable pairs.

Randomly chosen experimental and control groups are likely to differ in many respects. The relevant respects, for matching purposes, are those which are correlated with the response variable. Such a correlation will be known or surmised on the basis of past experience; for example, if blood pressure is the response variable, it is known to be influenced by age and body weight. But such concomitant variables are usually correlated among themselves as well as with the response variable: thus, older subjects are likely to be heavier than younger subjects.

The degree of intercorrelation obviously affects the selection and number of matching criteria, since the law of diminishing returns may limit the number that can be usefully employed. If, for example, we have three intercorrelated variables and we classify the population in accordance with the first two, we automatically classify to some extent in accordance with the third, whether we introduce it into the matching scheme or not. The effect of introducing the third variable into the matching scheme will improve group homogeneity only in so far as it is not already accounted for through its correlation with the first two variables. Evidently, therefore, when a choice exists, matching variables should be selected from those which are least intercorrelated among themselves but are best correlated with the response variable.

Having chosen the best matching criteria, it is necessary to divide them into classes: for example, age has to be divided into age classes. The product of the number of classes selected for each matching variable gives the number of matching categories, that is, of the sections of the population such that any two members belonging to the same category may be regarded as matched with respect to all criteria. For example, a scheme involving the two sexes and four age classes gives eight matching categories. If three social classes and three clinical state groups (e.g. "severe", "moderate", "mild") are then added, we have no less than 72 matching categories. It is easy, if the matching scheme is allowed to become too elaborate, for the number of matching categories to become impractically large.

**METHOD**

Appropriate experimental situations were simulated on an Elliott 803 computer. Briefly, this involved:

(a) Defining a "population" with response and matching variables having specified distributions and interrelationships;

(b) Choosing various numbers and boundaries of matching categories;

(c) Programming the computer to draw and analyse matched and simple random samples from this particular "population".

The smaller the variance of a difference in the values of the response variable between the experimental and control subjects, the more efficient is the method of sampling and analysis. Thus, the efficiency of matched sampling versus covariance analysis of simple random samples can be defined as the ratio of two variances. For example, if the variance of a difference in covariance analysis is 90 per cent. of that obtained with matching, the relative efficiency of matching is 90 per cent.

Suitable "treatment effects" were built into the simulation programme in order to compare the power of the two methods of sampling and analysis. Comparisons of power of two methods are based in the present context on the proportion of samples in which each method will detect the treatment effect as significant at the 5 per cent. level.

Over 14,000 matched and simple random samples, with ten to eighty "subjects" in each group, were drawn. The results will be described and discussed in a statistical journal (Biometrics). Here, only those comments and conclusions which are of importance to medical investigators will be given.

**STATISTICAL CONSIDERATIONS**

**METHODS OF ANALYSIS**

When an all-or-none response, such as "cure", is investigated a matched pair sample of N pairs can be represented as follows:

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Not Cured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated</td>
<td>Cured</td>
<td>Not Cured</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

The frequencies a, d, though relevant to the general rate of cure, give no information on the existence of a difference between groups. A test of significance of this difference should be based on the deviation of the frequencies b, c from equality. An exact test can...
be based upon a sum of binomial probabilities; if
b and c are not very small, a $\chi^2$ test with one degree
of freedom can be used, this being (Cochran, 1950):

$$\chi^2 = \frac{(bc - cl - 1)^2}{b + c}$$

In frequency matching, where the identity of pairs
is not preserved, each matching category provides a
simple $2 \times 2$ table, for which $\chi^2$ values can be
calculated in the usual manner and combined into
an overall test for a difference of proportions
(Cochran, 1954). This treatment obviously calls for
fairly large numbers in each matching category, and
therefore for large samples, even when only a few
matching categories are used.

In half of the examples found in medical journals,
mixed-pair and frequency-matched samples were
arranged as follows and tested by $\chi^2$ in the usual
manner:

<table>
<thead>
<tr>
<th>Result</th>
<th>Treated</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cured</td>
<td>A</td>
<td>C</td>
<td>A+B</td>
</tr>
<tr>
<td>Not Cured</td>
<td>N = A+C</td>
<td>B</td>
<td>C+D</td>
</tr>
<tr>
<td>Total</td>
<td>N</td>
<td>N</td>
<td>2N</td>
</tr>
</tbody>
</table>

For matched-pair samples $A, B, C,$ and $D$ are the
relevant marginal totals of the previous table. Such
amalgamation of all matching categories for the
purposes of testing for a difference of proportions of
"cure" between treatment and control groups is
legitimate only if the probability of "cure" (on the
null hypothesis) is assumed to be the same for all
matching categories; and if this is so there is no
point in matching. Fortunately, unless the differences
in the probability of "cure" between the matching
categories are large and highly correlated with
matching criteria the results of the simple treatment
versus control $\chi^2$ test are unlikely to be substantially
different from those obtained by the correct pro-
cedure.

When the response variable is quantitative the
appropriate test for a difference between means of
matched-pair samples is the $t$-test of differences
between pairs. Frequency matching per se does very
little to improve the precision of the comparison;
all that is gained, in comparison to two randomly
selected groups, is that the means of the matching
variables in two frequency-matched groups should
be approximately equal. Examination of the differ-
ence between groups now requires an analysis of
variance between and within matching categories.
This is the method of analysis to adopt when
frequency matching has been used; at the stage of
planning an experiment, however, frequency match-
ing is seldom desirable. If there are many matching
criteria then frequency matching for qualitative
variables followed by covariance analysis with
respect to quantitative variables may be advan-
tageous.

In more than one-third of the papers scrutinized a
t-test was performed as if the samples were un-
matched. By doing this, the investigator forfeits
virtually all the improvement of precision which
matching could offer and the more effective the
matching the greater is the relative loss of precision
resulting from the use of an incorrect test procedure.

**EFFICIENCY OF MATCHING:**

**SAMPLES DRAWN FROM THE SAME POPULATION**

In situations where "structural" similarity of the
experimental and control groups cannot be assumed,
conceptual difficulties arise, and will be discussed
later. It will be assumed, for the present, that the
experimental and control subjects have been drawn
from the same population. That is, the relationships
between the matching variables themselves, and
between them and the response variable, are the
same for both groups. This is the situation in most
drug trials.

Suppose, for example, we are testing a drug which
may shorten labour. The response variable is
quantitative, duration of labour. To simplify things,
the trial is made on primigravidae only, but it is
thought that maternal age may influence the duration
of labour, and this is chosen as the matching variable.
If the effectiveness of the drug is independent of age,
we are simply testing for a distance between two
parallel lines or curves representing duration of
labour by age. But if the effect of the drug varies
with age, the lines are no longer parallel, and one of
the purposes of the trial should be to discern such
differential effects.

Experiments simulating these various possibilities
were made on the computer over a wide range of
correlations between the response and the matching
variable. For parallel linear regressions, matching
was always less efficient than covariance analysis of
random samples, relative efficiency decreasing from
98 to 74 per cent. for correlations from 0.2 to 0.8
respectively. With various types of parallel but non-
linear regression, the relative efficiency of matching
was about 90 per cent. In either situation, tests on
matched samples were less powerful than covariance
analysis in detecting the treatment effect. There is no
theoretical reason to suppose that the relative
efficiency of the two methods should be reversed for
matching schemes of greater complexity.

Matched samples and random samples were
able to show non-parallelism (in terms of the
above example, differential response to treatment
by age) with equal efficiency. But, in practice, covariance analysis has the great advantage that the significance tests for parallelism of regressions would be made routinely, whereas with matched samples this aspect would probably be overlooked and an important feature in treatment missed.

As already noted, the rather common practice of using matched-pair or frequency-matched samples, and then applying the ordinary $t$-test appropriate for unrelated groups, is no more efficient than such a test on completely unmatched samples.

It might be argued that with qualitative matching variables, such as social class or severity of condition before treatment, covariance analysis cannot be used unless these variables are quantified, which many research workers are reluctant to do. But most qualitative variables used in medicine are capable at least of being ordered, and experiment showed that the crude but simple device of allotting scores at equal intervals (e.g. labelling three social classes, 1, 2, 3) could be efficiently used for covariance analysis, even when conditions were so manipulated that unequal intervals would have been much more appropriate. The existence of qualitative variables is therefore not necessarily a good reason for using matched samples. The experiments in fact showed that matching was less efficient (87 per cent.) than covariance analysis based on simply scored qualitative variables.

It remains to consider the situation with an all-or-none response: for example, a new method of treatment is to be compared with a standard method, the criterion being proportion of cures obtained. Experience with the standard treatment has shown that the probability of cure is related to the patients’ sex and the severity of the disease, and we wish to ensure that both methods are applied to groups which are comparable in these respects. Under such conditions, matching should be at its best, since with random samples the available methods of analysis are poor. This type of situation was investigated using up to nine matching categories. The distributions were arranged to resemble real-life situations, and incorporated differences in the probability of “cure” in different categories which are probably more extreme than those usually encountered in practice. In the more complex situations matching appears to have a slight advantage over random sampling, in respect of both efficiency and power.

The ordinary $\chi^2$ test on frequency-matched samples gave virtually the same results as those obtained with matched pairs. But, as has been noted above (“Methods of Analysis”) the use of this test may, theoretically, result in loss of power and is best avoided.

Efficiency of Matching:

**Samples drawn from different populations**

When subjects drawn from a population are allocated at random to treatment or control groups, we can reasonably expect that the two groups will be roughly balanced not only in terms of matching variables but also in terms of factors not included in the matching scheme. But this is not so when “experimental” subjects are self-selected and the controls have to be chosen from a different population. For example, we may wish to study the effect of premature birth on intelligence.

The experimental group is then premature births, and the controls must be selected from non-premature babies. Social class, which is known to be correlated with intelligence test scores, may be chosen as a matching variable. The incidence of prematurity is also correlated with social class, and groups of premature and non-premature babies are likely to differ in their social class composition, which gives an additional reason for matching. But this does not overcome the fundamental difficulty that groups of premature and non-premature babies are likely to differ in many respects other than social class, which may also be correlated with intelligence (e.g. obstetric history, family size). Since we cannot choose our premature and non-premature babies in advance from the same population, the allocation to these two groups is not random. Consequently, we cannot expect that the groups will be comparable in terms of concomitant variables, other than those included in the matching scheme; and even with an elaborate matching scheme we can never afford to assume that “other things are equal”. There is no way round this difficulty and it is essential to be extremely cautious when interpreting observed differences.

The fact that matching variables are differently distributed in the experimental and control populations may:

(a) Produce an observed difference even when the experience (prematurity, in the example) had no effect;

(b) Attenuate the effect of the experience, if the distributions of the matching variables and the effects of the experience “oppose” one another;

(c) Show an exaggerated effect of the experience if the different distributions act in the same direction.

Situations were simulated to give these three types of effect. The “population” was constructed to correspond roughly with the distribution of premature and non-premature births between three maternal social classes and three maternal height
groups in Aberdeen. The analysis of random and matched samples of eighty subjects in the experimental and in the control groups showed that for quantitative response variables matching was nearly as efficient and just as powerful as covariance analysis.

It should be remembered, however, that when the two populations are different, the relationships between the matching variables and the response need not be similar in each. This possibility would be examined as a matter of course in covariance analysis but would probably not be noticed with matched samples. And yet, the existence of substantial differences in the relationships between relevant variables in the two populations would cast doubt on the logic of matching itself.

With an all-or-none response, and with randomly-selected controls, no method of analysis which allows for the effect of associated variables is satisfactory unless both the experimental and control groups are unusually large. Experiments showed that with samples of 80 "subjects" matching was clearly superior. With randomly-selected controls, the usual \( \chi^2 \) test gave an excessive number of significant results in cases (a) and (c) above, and a considerable deficiency in case (b). Tests based on matched samples gave the expected proportions of significant results in all three situations.

As noted in connexion with samples drawn from the same population, it is theoretically better to use matched-pair than frequency-matched samples, although, experimentally, the two methods gave closely similar results.

**The Mechanics of Matching**

**Determination of Matching Categories**

The smaller the variance of a difference in response between matched pairs, the more efficient the procedure in demonstrating an effect of treatment.

For qualitative matching variables, the possible class boundaries are determined by the original classification: for example, by a social classification in which there are five classes. If differences of response between adjacent classes are small (e.g. if Social Classes I and II or IV and V are similar in their behaviour) it is advantageous to amalgamate such adjacent classes. This gives a more compact classification without appreciable loss of efficiency.

For quantitative variables (e.g. body weight) the class boundaries must be specified by the investigator. If more than three or four classes per matching variable are used, matching may become difficult. Empirically, it was found that for three classes the choice of boundaries was not at all critical. An interval of width about one standard deviation in the centre of the distribution appeared to be nearly optimal for the middle class whatever the value of the correlation between the response and the matching variables. An increase of the number of classes from three to four did not give a worthwhile increase of efficiency.

**Selection of Subjects**

Consider first the situation in which the experimental group is identified in advance. For example, in a study of mental development of premature and non-premature children, or of social adjustment of delinquent and non-delinquent children, the experimental group (premature or delinquent) is identified from hospital or court records. In the prematurity investigation, the controls could be selected from the remainder of the same large pool of hospital records from which the treatment group was extracted. In this situation matching presents little technical difficulty, the distribution of the treatment group and of the remainder of the cases with respect to a set of prospective matching variables can be ascertained, and this will provide information as to how closely and for how many variables it is feasible to match. However, in the delinquency investigation, there is no convenient pool of records from which to draw control cases; although we know the distribution of the treatment group, we have only a vague idea about the distribution of the remainder of the population with respect to the matching variables. To provide controls, we must obtain a large random sample (not containing individuals who should be in the treatment group), and determine the distribution of this sample with respect to matching characteristics; we can then decide the matching scheme and sub-sample for controls who will be assessed in terms of the response variable.

Next consider the situation in which the differential experience to which the treatment and the control groups are subjected can be imposed by the investigator, as in a trial of two competitive drugs. If the patients are admitted to trial as they present themselves at a clinic, we must consider not only whether controls can be found if a given matching scheme is adopted but also how long it will be until a suitable control subject presents himself at the clinic. A matching scheme with many rare categories will be difficult to implement, and an analysis of past records will serve as a guide to simplification. Whatever the final matching scheme is, it is advisable to minimize the "waiting time" between comparable cases admitted to the treatment and control groups. The shorter the waiting time, the greater the chance that the general management of patients and the
level of subjective assessments will be similar for comparable treatment and control subjects. For some conditions (e.g. hyperthyroidism) spontaneous improvement or deterioration of the condition of the subjects follows a seasonal pattern—a good reason to have treatment and control cases closely spaced in time. Finally, an allocation system which minimizes the waiting time between comparable treatment and control subjects ensures that the two groups are similar in composition at any point of time during the progress of the investigation. Consequently, should the investigation be terminated earlier than originally planned, the available cases will provide two groups reasonably well balanced with respect to all matching criteria.

A simple “minimum waiting time” method of allocating subjects to treatment and control groups is as follows. For any matching category, the first subject is allocated at random to treatment or control. If the first subject was allocated to the treatment group the next subject belonging to this matching category is automatically allocated to the control group and vice versa. The third subject in this matching category is again allocated at random and his allocation determines the lot of subject number four, and so on. With this system of allocation the number of treatment and control subjects will differ at most by one in any matching category and at any point of time during the progress of the experiment.

While, for any given matching scheme, such a procedure minimizes waiting time, it does not overcome the fundamental difficulty that complex schemes with rare matching categories are likely to involve long delays in the completion of pairs. Thus, if a matching category covers only 5 per cent. of the population, and the weekly intake is limited to four subjects, it may easily be 10 weeks before a pair can be completed. The only solution to this difficulty is to use matching categories which are not unduly rare.

**Duration of Experiments**

Just before the completion of the sample there will usually be a number of subjects who, although suitable for study, cannot be admitted since they do not correspond to the outstanding unpaired subjects. It may be thought that this will not bother the experimenter to any great extent. Assume, however, that the number of suitable subjects in any week is a Poisson variable with mean 3 and that the resources are such that there is no upper limit to the number of subjects admitted each week. These assumptions were included in the simulation programme to investigate the effect of matching on the time required to complete a sample. In about 40 per cent. of occasions a longer period than 27 weeks (eighty subjects at an average rate of three per week) was necessary to complete a random sample. For a simple matching scheme of three or four categories, the completion of the series required only about a week longer. For six and nine matching categories, the completion required on average three and six weeks more respectively than without matching; thus on average nine and eighteen suitable subjects respectively will pass through the clinic without being included in the investigation. If special medical or laboratory facilities have to be provided for the investigation, it is necessary to estimate how long they will be needed. In the examples, to be 90 per cent. certain that the sample will be completed within a stated period the investigator has to allow 31 weeks when simple random sampling is used and 43 weeks when he decides to match for nine categories. If limited resources set a maximum to the number of subjects that can be accepted in one week, matching appears still more disadvantageous since early filling of the quota for a week may cause rejection of a subject who would have completed a pair.

The moral is that if an investigation has to be completed within a given time, it may be possible to use random samples which are much larger than matched samples. Under such circumstances trials showed that the increase of power gained from larger sample size may more than outweigh the advantages of matching.

Similar conclusions obtain for the use of matching in sequential trials since, as pointed out by Billewicz (1956) and Armitage (1960), the reduction of average sample size due to matching is likely to be small unless the differences between strata are exceptionally large.

**Discussion**

It may be helpful at this point to re-examine the arguments in favour of matching. The argument that matching improves the clarity of the experimental design is weak. To construct a good matching scheme requires as much information about relevant variables and their interaction, as is necessary to plan the collection of data with a view to covariance analysis. With a matching scheme, the importance of the variables is pre-judged, and since our knowledge is always incomplete some of the variables may in fact be unnecessary. In covariance analysis only those variables will be used which improve the precision of the experiment.

That matching simplifies statistical analysis cannot be questioned. Yet covariance analysis is by no
MATCHED SAMPLES IN MEDICAL INVESTIGATIONS

means difficult, and is much more informative. It may be overlooked that the collection of matched samples is sometimes difficult and time-consuming, and it may be more economic to undertake a more complex analysis at the end of the investigation than to invest in a complex matching plan. Even when we have to deal with a mixture of qualitative and quantitative variables, it seems that simple scoring of the former is at least as effective as matching. When the number of really useful matching variables is inconveniently large, we may frequency match the groups with respect to the more “awkward” and perform covariance analysis with respect to the remainder.

The danger that random selection may result in samples so “freakish” as to affect the efficiency of covariance analysis is by no means as great as is often supposed. In the simulated experiments, using 5,600 random and matched samples of ten to eighty “subjects” each, the “treatment effect” was recognized as significant at the 5 per cent. level in 4,491 random and 4,283 matched samples. In these experiments, the correlation between the response and matching variable ranged from 0.2 to 0.6; with higher correlations, matching is very ineffective. It appears, then, that the chance of missing a real treatment effect is about 20 per cent. greater with matched samples than with covariance analysis.

There is no doubt that matching improves the precision of comparisons*, but with quantitative response variables an even greater improvement can be achieved with covariance analysis. With all-or-none response, matching is advantageous, particularly when the samples are small and especially when the treatment and control groups have to be drawn from different populations. Under these circumstances, matching may assist precision and guards against the possibility of bias in the estimation of the treatment effect. When the experimental and control groups are chosen from the same population (as in a drug trial), matching is preferable if the sample size is fixed. But larger random than matched samples can usually be collected in a given time and the increase in sample size may outweigh the advantages of matching.

The choice whether to match or to use randomly selected subjects depends upon practical as well as theoretical considerations. Matching can be costly in time and effort; before a final decision is made the probable duration of a trial based on matched samples, and the gap in time between the admission of comparable subjects should be carefully considered. Random sampling is always easier to administer and may also be more efficient from a statistical point of view.

The danger of matching is that it may appear to excuse the investigator from looking at more than just one aspect of the data. After matching, it is generally assumed that the groups are comparable in all relevant respects. At least one investigator (Douglas, 1960) found that the omission of the less obvious matching criteria might have led to completely erroneous conclusions. In relation to matching subjects at the beginning of a longitudinal study, the same author observed that matching criteria which seem satisfactory at the beginning of the inquiry are not necessarily those relevant for a study of the temporal progress of the subjects.

This paper is not meant to be a wholesale condemnation of the use of matching procedures in medical studies; it is rather a plea for preserving matching for situations in which it is advantageous.

I am greatly indebted to my colleagues in the Research Group in Biometric Medicine and in the Obstetric Medicine Research Unit, and to Dr J. Crooks for comments and discussions. I am particularly grateful to Dr A. M. Thomson for assistance in drafting this paper.

REFERENCES


* Mathen (1963) examined the usefulness of matching in comparative public health studies. On topics on which the two papers overlap, his theoretical approach leads to similar conclusions to those obtained empirically in the present study.