

AMERICAN Journal of Epidemiology

Formerly AMERICAN JOURNAL OF HYGIENE

VOL. 91

FEBRUARY, 1970

NO. 2

COMMENTARY AND BRIEF REPORTS

MATCHING AND DESIGN EFFICIENCY IN RETROSPECTIVE STUDIES

OLLI S. MIETTINEN¹

A COMMENT AND A REACTION

In a recent article in *Biometrics* (1) I considered, among other matters, the efficiency of pairwise matching in experiments and in prospective nonexperimental studies. In a passing comment I cautioned against applying the statistical efficiency results to *retrospective* studies, pointing out that *with the retrospective approach matching generally tends to reduce efficiency and would therefore have to be motivated by the pursuit of validity alone*. That note has drawn a reaction in this journal from Bross (2), who attributes my view to the use of an "unrealistic" mathematical model, attempts to refute my position by presenting a "simple hypothetical counterexample", and maintains that even in retrospective studies it is reasonable to seek design efficiency by "matching out a strong factor" in exploring the effect of a weaker factor.

By reading my article (1), and possibly also the reference (3) given there and/or a subsequent discussion (4), the reader may verify that my reservations about the efficiency motive for matching in retrospective studies do not represent conclusions from any mathematical model but have a rather nonmathematical rationale.

As to the "refutation" of the assertion at

issue, in the "counterexample" Bross (2) does not actually compare the efficiency of the matched pairs *design* (involving constraints in the selection of control subjects together with a specific procedure of analysis) to that of using an independently selected control series. Instead, he merely compares two different ways of *analyzing* data from a study in which a matched control series was in fact used, an appropriate one (maintaining the original pairing) and a generally inappropriate one (with random re-pairing for analysis). Thus, Bross' "counterexample" illustrates the well-known analysis principle that efficiency is lost in hypothesis-testing if matching is ignored in the analysis, but it has no bearing on the question of whether matching should be applied in the *selection* of the control series.

THE "COUNTEREXAMPLE" RE-EXAMINED

The presentation of a proper numerical treatment of the "data" in Bross' "counterexample" (2, table 1) might help put his argument in perspective and clarify the assertion which inspired his commentary. In order to compare the efficiency of the matched pairs design to the use of an independent control series it is necessary to start out by considering the source populations of the *propositi* and controls. In reconstructing these populations for the present purpose we may assume, without loss of generality, that the sampling fraction for *propositi* was unity. Moreover, it is neces-

¹Departments of Epidemiology and Biostatistics, Harvard School of Public Health, and Cardiology Division, Department of Medicine, Children's Hospital Medical Center, Boston, Mass.

TABLE 1
Source populations for Bross' (8) "data"

		Category 1			Category 2		
		Exposure		Total	Exposure		Total
		+	-		+	-	
Disease	+	114	0	114	94	50	144
	-	$114(1 - P_1)/P_1$	0	$114(1 - P_1)/P_1$	$94(1 - P_1)/P_1$	$50(1 - P_2)/P_2$	$94(1 - P_1)/P_1 + 50(1 - P_2)/P_2$
Total		$114/P_1$	0		$94/P_1$	$50/P_2$	

sary to assume that the matching factor, a strong correlate of the exposure, bears no relation to the disease conditionally on the exposure, because otherwise it would be a confounding factor and would have to be controlled in the interest of validity while consideration of efficiency would become meaningless (see section below, *The case of confounding*). Finally, before setting up the source populations, Bross' tables 1-b and 1-c need to be adjusted the way he himself suggests (2, p. 363). The resulting source populations of *propositi* (disease +) and *controls* (disease -) are shown in table 1, where P_1 and P_2 denote the prevalences of the disease in the exposed and nonexposed subpopulations.

With these source populations, consider first the expected outcome of sampling when using pairwise matching. With this design we are merely concerned with the outcome with respect to discordant pairs. Let Z_{10} denote the number of pairs in which the case is exposed and the control is unexposed, and let Z_{01} denote the number of pairs with the opposite exposure pattern. Category 1 (see table 1) contributes nothing to either Z_{10} or Z_{01} . From category 2 the expected contribution to Z_{10} is $94[50(1 - P_2)/P_2]/[94(1 - P_1)/P_1 + 50(1 - P_2)/P_2] = 94[50P_1(1 - P_2)/(1 - P_1)P_2]/[94 + 50P_1(1 - P_2)/(1 - P_1)P_2]$. To evaluate this further, we observe that the ratio of exposed and nonexposed controls in category 2 (see table 1) may be estimated to be 78/66 on the basis of Bross' table 1-b after the adjustment already re-

ferred to. This yields $P_1(1 - P_2)/(1 - P_1)P_2 = 94(66)/50(78) = 1.59$, and the corresponding estimate of the expected value of Z_{10} becomes 43. By analogous computation, the estimate of the expected value of Z_{01} becomes 27. These are, as they should be, the values in Bross' table 1-a. The McNemar test (5), without Yates' correction, yields

$$\chi^2(1) = (43 - 27)^2/(43 + 27) = 3.66 \quad (i)$$

The maximum likelihood estimate of the relative risk ρ is (6)

$$\hat{\rho} = 43/27 = 1.59, \quad (ii)$$

and the exact 90 per cent confidence interval for ρ becomes (7)

$$1.03 < \rho < 2.47. \quad (iii)$$

Next, suppose the expected data with pairwise matching, derived as in the above or taken from Bross' table 1-a, are analyzed as if the two series were independent. The ordinary chi square test, without Yates' correction, would yield $\chi^2(1) = [208(66) - 50(192)]^2/516/(258)^2/400(116) = 2.85$. Alternatively, if the subjects were randomly repaired, the expected values of Z_{10} and Z_{01} would be $208(66)/258 = 53.2$ and $50(192)/258 = 37.2$, respectively, and the corresponding value of the McNemar statistic would be $(53.2 - 37.2)^2/(53.2 + 37.2) = 2.83$. This value is, as expected, very close to the above 2.85. Bross considered the ratio of the correct chi square value in (i) and this incorrect result, $3.66/2.83 = 1.29$, and regarded this as a measure of a presumed effi-

ciency gain by using the matched pairs design (2, p. 363). In point of fact, however, this ratio—to the extent that it has anything to do with efficiency—is an indicator of the efficiency gain by *proper analysis* (maintaining the original pairing) relative to an inappropriate analysis (ignoring the original pairing in the selection) in a case where a matched control series was used. The inappropriateness of the alternative analysis that Bross considered manifests itself also in the fact that the estimation of relative risk from the randomly re-paired data is biased: $53.2/37.2 = 1.43$ as opposed to the correct value $\hat{\rho} = 1.59$ in (ii). Similarly, the 2×2 table which ignores the original pairing gives $208(66)/50(192) = 1.43$ instead of the correct 1.59.

Let us now turn from this alternative and inappropriate *analysis* of data from a study with a matched control series to the actual alternative *design*—the use of an independent control series. The series of *propositi* would remain unaltered, i.e., there would again be 258 cases, of whom 208 would be exposed and 50 would be unexposed (cf. table 1 or Bross' table 1-a). As to the controls, the proportion in the source population (see table 1) falling into category 1 is $[114(1 - P_1)/P_1]/[(114 + 94)(1 - P_1)/P_1 + 50(1 - P_2)/P_2] = 114/[114 + 94 + 50P_1(1 - P_2)/(1 - P_1)P_2]$, and using the earlier estimate 1.59 for $P_1(1 - P_2)/(1 - P_1)P_2$ the estimated proportion of control population falling in category 1 becomes 0.397. Thus, the expected proportion of exposed individuals in the independently selected control series has the estimate $0.397(1.000) + (1 - 0.397)[94(1 - P_1)/P_1]/[94(1 - P_1)/P_1 + 50(1 - P_2)/P_2] = 0.397 + 0.603(94)/[94 + 50(1.59)] = 0.724$. This corresponds to 187 exposed and 71 unexposed individuals among a series of 258 controls. With these data the ordinary chi square statistic without Yates' correction takes on the value

$$\chi^2(1) = [208(71) - 50(187)]^2 / 516 / (258)^2 395(121) = 4.75 \quad (\text{iv})$$

The estimate of the relative risk is computed (8) as

$$\hat{\rho} = 208(71)/50(187) = 1.58, \quad (\text{v})$$

and an approximate 90 per cent confidence interval for ρ is (9, eqns. 10 and 11)

$$1.12 < \rho < 2.23. \quad (\text{vi})$$

The relative merits of the two designs may now be evaluated by comparing the results in (i), (ii) and (iii) to (iv), (v) and (vi), respectively. It may be noted first that, with the constraint that the matching factor, a strong correlate of the exposure, was not a risk factor of the disease conditionally on exposure, both designs are valid: the estimates of relative risk in (ii) and (v) differ only to such an extent as the use of integer frequencies necessitates. On the other hand, a difference in efficiency is apparent both in the value of the chi square statistic and the width of the confidence interval. The differences illustrate the loss of efficiency which tends to arise in retrospective studies from matching which is unnecessary for validity.

The above review should dispose of Bross' (2) argument and give insight to the assertion which he found "confusing", "misleading" and "wrong". In my view, his commentary illustrates a rather general confusion that still prevails when it comes to the fundamentals of matching in retrospective epidemiologic studies. There thus seems to be a need for an attempt to clarify the issues over and beyond Bross' "simple hypothetical counterexample".

THE ENDS AND THE CONFUSION ABOUT THE MEANS

The objectives of matching in nonexperimental research of disease etiology are to increase the validity and/or efficiency of a study (10). The *validity* objective is generally considered to be the predominating one. It concerns the removal of bias from the estimation of the effect under study (or the control of the level of significance-testing), i.e., it has to do with the "comparability"

of *propositi* and controls. In the present context we are not specifically concerned with the principles of increasing validity by matching. Suffice it to point out that these principles do not depend much on whether the prospective or the retrospective approach is used, and that if matching for a given factor is to increase validity, this factor has to be predictive, even under the null hypothesis, of both the exposure (or trait) and the disease under investigation (3, 4).

Our concern here is with the secondary objective of matching, *efficiency*. This is usually discussed and defined without regard to the "cost" of applying the selection constraints. Bross (2) also did this, and the same viewpoint shall be maintained here for the sake of comparability and simplicity. Design efficiency thus refers to the precision of the estimation of the effect (or the power of testing the hypothesis of no effect) with a given total number of study subjects.

As to the principles of seeking efficiency by matching, it seems to be commonplace to assume that one may apply to retrospective studies the ideas which relate to the experimental procedure of grouping ("blocking"). Thus, it is felt that by matching for risk factors of the disease it is possible even with the retrospective approach to remove a presumed blurring effect of variability between the matching categories. Bross (2) even claimed empirical justification for this when he, in the context of discussing retrospective studies, made the assertion that: "From extensive past experience, it has been found that strong etiologic factors often mask the effects of weaker factors. It has also been found that by matching out the effects of a strong factor, such as age, it is possible to improve the chances of detecting a real—but relatively weak—relationship in a secondary factor. In other words, matching out a strong factor can improve the design efficiency of a study". No substantiating evidence was given over and beyond the "simple hypothetical counterexample", but, as shown above, that example, as he treated it, does not relate to the point at issue.

In fact, however, the efficiency principles of experimentation and of prospective non-experimental studies do not apply to retrospective studies. To the extent that matching in retrospective studies influences efficiency, the effect tends to be a *loss* of efficiency rather than a gain. This loss arises only if the matching factors are correlates of the *exposure*. Otherwise efficiency is not influenced by matching. A justification of this position is given below—not in terms of any mathematical model nor by reference to "extensive experience" but by presenting its rationale.

TO MATCH OR NOT TO MATCH

Overview of conditions, complications and implications

In a typical retrospective study the *propositi* are cases of the disease whose etiology is being investigated (11). The use of *matched* controls has nothing to do with the characteristics of the cases; it can only influence the characteristics of the control series. It assures that the control series will have the same structure as the series of *propositi* with respect to the matching factors. The problem is to understand the secondary effects of this similarity on the expected pattern of exposure in the control series and to correctly interpret the implications of these effects.

The general implications of matching in a retrospective study depend on the matching factor's relationship to both the exposure and the disease, as follows:

When the matching factor is

- 1) *unrelated to exposure*, no gain or loss of either validity or efficiency is possible through matching; matching is futile.
- 2) *related to exposure*
 - a) *and also related to disease* (conditionally on exposure), matching serves validity; efficiency is not, therefore, natural to consider here, but matching does tend to reduce the likelihood of a "significant" result; matching is

motivated by the need to avoid confounding.

- b) *but unrelated to disease* (conditionally on exposure), matching is irrelevant for validity but reduces efficiency; matching is to be avoided in the interest of efficiency (avoidance of overmatching).

The case of futility

Let us first consider case 1 in the above scheme. It is obvious that if the matching factor is unrelated to the exposure, then the approach of matching in the selection of controls (and maintaining the original matched groups in the analysis) is interchangeable with the procedure of drawing the control subjects independently of the cases and then randomly grouping the cases and controls for analysis. The latter approach, in turn, is equivalent to the use of independent controls without pairing in the analysis, apart from the case of very small samples (12, 13). Matching, no matter how strongly the matching factor may be related to the disease, would therefore have no implication for either validity or efficiency. An example of this situation is provided by the recent retrospective study of Jick et al. (14), concerning the etiologic role of blood group O in venous thromboembolic disease in women. And indeed, there was no matching (or other control) for even age or sex. Both of these characteristics may be considered to be "strong etiologic factors", but the point is that they are practically unrelated to the exposure, blood group O.

The case of confounding

If a matching factor is related to the exposure (case 2), there is a need to consider the question of whether it is also associated with the disease. Consider first the situation where the factor is associated both with the exposure and the disease (case 2a). A typical example of this is provided by the problem which Bross (2) dealt with, namely age in

the exploration of the possible etiologic role of lactation in breast cancer. A factor which is related to both the exposure and the disease is a classical confounding factor, so that matching in this situation serves the purpose of increasing the validity of the etiologic inference.

Whether matching in this case also improves efficiency may not be equally clear, however. In my view it is meaningless to be concerned with the relative precisions of estimations with and without matching when the estimate without matching is known to involve a bias of *unknown* magnitude. In other words, the validity objective so dominates the consideration of efficiency that relative efficiency can be defined meaningfully only when both designs are valid, i.e., when the potential matching factor does not have the confounding property. On the other hand, if one chooses to think in terms of the power of the significance test without regard to validity (as Bross (2) did), it is apparent that the removal of bias (the "spurious" component of the total association between the exposure and the disease) by matching would usually tend to *reduce* the association and therefore the likelihood of obtaining a "significant" result, particularly if a "strong" confounding factor is matched for in exploring a "weak" etiologic factor. Indeed, as to the very example at hand, there were various early observations of association between breast cancer and lack of lactation experience, suggesting an etiologic connection, but later studies with careful matching (e.g., MacMahon and Feinleib (15)) have failed to reproduce it. As explained in the above, I do not consider this a loss of efficiency in view of the validity implication of the matching, even though the matching constraints apparently resulted in *reduced* "significance" of the statistical test results. With his "simple hypothetical counterexample" Bross (2) attempted to show an increase in "significance" but, as already noted, he compared different types of analysis, not a matching design to the use of an independent control series.

The case of overmatching

It remains to examine the case where the matching factor is related to the exposure but would not be related to the disease under the null hypothesis of no etiologic connection between the exposure and the disease (case 2b). This is a common situation in practice. The factor at issue may be time of birth, place of residence, occupation, social class, family (in the use of sibling controls), or the like.

As an example, let us consider the matching problems in a retrospective study concerned with the hypothesis that induced abortion is conducive to ectopic implantation in a subsequent pregnancy (16). Factors such as number of pregnancies would seem to be related to both the exposure (induced abortion) and the "disease" (ectopic implantation). They would, therefore, be considered confounding factors and would be matched on in the interest of validity.

But let us focus on the problem of whether it is desirable, in this particular study, to match on year of birth, knowing that this factor would serve as a predictor of the exposure, due to an increase in the frequency of induced abortions during the time period covered in the sampling frame (a large series of hospital patients enrolled in another study). There is no reason to think that there was a time trend in the risk of ectopic implantation over and beyond what may have been secondary to the trend in the frequency of induced abortions. Suppose that there indeed was none. In this situation, matching for year of birth would be irrelevant for validity (this is not case 2a). Under the null hypothesis matching would have no efficiency implication as far as hypothesis-testing is concerned, because the probability of rejecting the null hypothesis would equal the significance level regardless of whether matching is used or not. However, the precision of the estimation of relative risk (which here equals unity) would be reduced by matching, because it is estimated only from the discordant pairs (6, 7), and

these tend to decrease in number with increasing correlation between the matching factor and the exposure (1). Under the non-null hypothesis in a situation of this type, the disease tends to occur where the cause (exposure) is. Thus, due to the increasing trend in induced abortions, cases of ectopic pregnancy would tend to have later birth dates than noncases. This difference in birth date distributions is here a manifestation of nothing but the causal connection being investigated. Matching in the selection of controls would completely mask this manifestation, and the result would be a loss of information. This is most obvious when considering the extreme case where the matching factor is a perfect predictor of the exposure: with, say, the matched pairs design no discordant pairs could occur, and thus there would be no basis for the estimation of relative risk or for hypothesis-testing. In less extreme situations the loss of efficiency as well as the irrelevance for validity is easy to verify numerically; an example of this was provided in an earlier section of this discussion when dealing with Bross' "counterexample".

THE PROSPECTIVE-RETROSPECTIVE CONTRAST

According to Bross (2), "The key to a fair evaluation of matching procedures in retrospective epidemiologic studies is to adopt a 'descriptive' rather than a 'prescriptive' approach to the problem." In the above I have made an attempt to analyze and describe, for retrospective studies, the implications of matching under various relevant conditions. From that description it should be apparent that, as far as design *efficiency* is concerned, the principles and potentials of matching in retrospective epidemiologic studies are quite different from those of grouping in experiments or matching in prospective studies. In the latter situation efficiency may be gained by the control of predictors (risk factors) of the *disease*; apart from the case where the exposure influences the matching factor, nothing needs to be said about the relationship of the grouping factor to the ex-

posure, the presence or absence of which is determined by either interventive allocation or selective inclusion by the investigator. In contrast to this, it was shown in the above analysis that the key to the efficiency implications of matching in retrospective studies is the relationship of the matching factor to the *exposure*. Efficiency is lost by matching for nonconfounding factors related to the exposure; little or nothing (depending on validity) needs to be said about its relation to the disease, whose presence or absence is based on the investigator's selection.

When moving from the experiment to its rather distant relative, the retrospective epidemiologic study, it is natural and a matter of definition that the outcome (criterion, response) variate shifts from the disease to the exposure. The corresponding shift in the relevant grouping factors from correlates of the disease to correlates of the exposure should also be easy to accept. There may be, however, some difficulty in internalizing the resultant shift from a gain in efficiency to a loss of it. I find it helpful to consider grouping by a perfect determinant of the outcome. In this situation there are basically two kinds of groups. In one, all members of all groups are destined, under the null hypothesis, to show the criterion outcome (the disease in an experiment and in a prospective study, or the exposure in a retrospective study); in the other type, none would show it. The crucial difference now is this: In experiments and in prospective studies, one possible component in sufficient causal constellations, the exposure under study, varies by design within the otherwise homogeneous groups, and this exposure has a chance, a perfect one indeed in this extreme case, to show its effect on the disease. In the retrospective study, on the other hand, there will be no such remaining element to cause variability in the homogeneous exposure patterns within the matched groups. In particular, the disease generally is not part of any causal complex of the exposure and it can not, therefore, induce variability in the

exposures within groups that are homogeneous with respect to the true determinants of exposure.

To put it differently, it is essential in etiologic research that there be variation of exposure within the elementary study groups. In experiments with grouping and in prospective nonexperimental studies with matching, the elementary study groups are the blocks and the matched groups, respectively, and within these elementary groups the exposure varies by design. In retrospective studies only the disease varies by design, and one should permit maximum natural variation of exposure within the matched groups rather than limit it by matching when this is not required for validity.

REFERENCES

1. Miettinen, O. S. The matched pairs design in the case of all-or-none responses. *Biometrics*, 1968, **24**: 339-352.
2. Bross, J. D. J. How case-for-case matching can improve design efficiency. *Amer. J. Epid.*, 1969, **89**: 359-363.
3. Miettinen, O. S. Some basic theory for matching designs in nonexperimental research on causation. Ph.D. Thesis, Univ. of Minn. University Microfilms, Inc. (publisher), Ann Arbor, Mich., 1968.
4. Miettinen, O. S. Under- and overmatching in epidemiologic studies. *Atti del 5° Congresso Internazionale di Igiene e Medicina Preventiva*, 1968, **1**: 49-61.
5. McNemar, Q. Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika*, 1947, **12**: 153-157.
6. Cornfield, J. and Haenszel, W. Some aspects of retrospective studies. *J. Chronic Dis.*, 1960, **11**: 523-534.
7. Miettinen, O. S. Estimation of relative risk from individually matched series. *Biometrics*, 1970, **26**: in press.
8. Cornfield, J. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *J. Nat. Cancer Inst.*, 1951, **11**: 1269-1275.
9. Gart, J. J. On the combination of relative risks. *Biometrics*, 1962, **18**: 601-610.
10. Cochran, W. G. The planning of observational studies of human populations. *J. Roy. Stat. Soc. A*, 1965, **128**: 234-265.
11. White, C. and Bailar, J. C. Retrospective and

- prospective methods of studying associations in medicine. *Amer. J. Public Health*, 1956, *46*: 35-44.
12. Wald, A. *Sequential Analysis*. John Wiley and Sons, Inc., New York and London, 1947, p. 108.
 13. Youkeles, L. H. Loss of power through ineffective pairing of observations in small two-treatment all-or-none experiments. *Biometrics*, 1963, *19*: 175-180.
 14. Jick, H., Slone, D., Westerholm, B., Inman, W. H. W., Vessey, M. P., Shapiro, S., Lewis, G. P., and Worcester, J. Venous thromboembolic disease and ABO type: a cooperative study. *Lancet*, 1969, *1*: 539-542.
 15. MacMahon, B. and Feinleib, M. Breast cancer in relation to nursing and menopausal history. *J. Nat. Cancer Inst.*, 1960, *22*: 733-753.
 16. Miettinen, O. S. Individual matching with multiple controls in the case of all-or-none responses. *Biometrics*, 1969, *25*: 339-354.