

AMERICAN Journal of Epidemiology

Formerly AMERICAN JOURNAL OF HYGIENE

VOL. 93

FEBRUARY, 1971

NO. 2

COMMENTARY

MATCHING IN RETROSPECTIVE STUDIES

ROBERT J. HARDY AND COLIN WHITE¹

In the recent correspondence between Bross and Miettinen (1, 2), reference has been made by the latter to the method of selection of variables on which to match in order to design a valid retrospective study. This topic has not been exhausted.

It is often taken for granted that in retrospective studies one should match on factors that affect the incidence of the disease. Worcester considers this to be the prevailing view among those who design retrospective studies: "When a disease group is being compared with another group, matching is usually done on variables known to be related to the disease rather than on variables related to the outcome" (3). There are many books and articles in which such a procedure is explicitly laid down or tacitly assumed. It is, however, incorrect, as Miettinen and others have stated. In a retrospective study, the presence or absence of disease merely defines the two groups that are to be compared. The random variable that serves to measure the outcome is the factor that is being studied as a putative etiologic agent of the disease—for example, smoking in studies of cancer of the lung, or radiation of pregnant

women in studies of childhood leukemia. The factors on which matching should be considered must be related to this outcome variable; otherwise they do not affect the measure that is the basis for a decision about the association between the putative etiologic agent and the disease. For instance; there is good evidence that the ABO blood group is unrelated to sex or, indeed, to any factor other than race, and certain diseases; and in a retrospective study of the association between blood group and carcinoma of the cervix, male controls would be quite acceptable if they were healthy and were matched to the patients on race. The mistaken view that one should match on factors that affect the incidence of the disease probably arose from confusion with follow-up studies. In these it is proper to match on factors that are correlated with the disease, since the occurrence of disease is the outcome variable in follow-up studies.

Miettinen makes the further stipulation that one should match only on those factors that are correlated with both the outcome variable and the incidence of the disease. Since there are practical difficulties in matching on several variables, this recommendation, which has the effect of reducing the number of factors on which matching is attempted, is commended by its convenience. We do have a reservation about it, however.

¹Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, Connecticut 06510.

This work was supported by Biometry Training Grant USPHS 5-T01-GM-0047.

Miettinen is correct in principle: a factor that is related to the outcome variable will not, in general, affect the difference between the outcome for the cases and that for the controls, if it involves the same proportion of cases as controls. However, one should not rest assured that when a given variable fails to influence the development of a disease, matching on that variable is automatically unnecessary. For example, a disease may attack males and females with equal frequency in the long run but a particular sample of cases may, by chance, include a higher proportion of males than the control group. If the outcome happens to be correlated with sex, a spurious finding may result from the study. Our position is that whenever a factor is strongly correlated with the outcome, one should match on this factor, or take it into account in the analysis, even if, a priori, it is thought not to affect the incidence of the disease.

An important practical detail still remains to modify decisions about the variables on which to match. Often it is uncertain whether a variable is, or is not, correlated with the outcome variable. In this case a decision to match or not might properly be influenced by whether the variable is known to affect the incidence of the disease. If it clearly has such an effect, the investigator may decide to match on it rather than to run a risk of bias. The alternative is to control for this factor in the analysis.

REFERENCES

1. Bross IDJ: How case-for-case matching can improve design efficiency. *Amer J Epidem* 89: 359-363, 1969
2. Miettinen OS: Matching and design efficiency in retrospective studies. *Amer J Epidem* 91:111-118, 1970
3. Worcester J: Matched samples in epidemiologic studies. *Biometrics* 20:840-848, 1964