MATCHING IN EPIDEMIOLOGIC STUDIES: VALIDITY AND EFFICIENCY CONSIDERATIONS

by

Lawrence L. Kupper, John M. Karon,
David G. Kleinbaum, Donald K. Lewis
Department of Biostatistics
University of North Carolina, Chapel Hill, NC

and

Hal Morgenstern
Department of Epidemiology and Public Health
Yale University, New Haven, CT

Institute of Statistics Mimeo Series No. 1239

JULY 1979

MATCHING IN EPIDEMIOLOGIC STUDIES: VALIDITY AND EFFICIENCY CONSIDERATIONS

by

Lawrence L. Kupper[1], John M. Karon[1],
David G. Kleinbaum[1], Hal Morgenstern[2],
and
Donald K. Lewis[1]

[1]Department of Biostatistics
University of North Carolina, Chapel Hill, NC

[2]Department of Epidemiology and Public Health
School of Medicine, and Institution for Social and Policy Studies
Yale University, New Haven, CT

KEYWORDS: matching; follow-up and case-control studies; validity; efficiency.

ABSTRACT

This paper addresses both validity and efficiency issues with regard to the use of matching and random sampling as alternative methods of subject selection in follow-up and case-control studies. Only the simple situation involving dichotomous disease and exposure variables, and a single dichotomous matching factor, is considered; cost considerations (e.g., due to the loss of candidate subjects because of matching constraints) are ignored.

Given this framework, it is demonstrated that the decision to match or not should be motivated solely by efficiency considerations. An efficiency criterion based on a comparison of confidence intervals under matching and random sampling for the effect measure of interest (i.e., the risk ratio and risk difference in follow-up studies and the odds ratio in case-control studies) leads to the following conclusions. In follow-up studies, matching on a confounder is expected to lead to a sizeable gain in efficiency over random sampling, and matching on a non-confounder is not expected to result in a loss in efficiency. In case-control studies, matching on a confounder is expected to lead to a worthwhile gain in efficiency in most practical situations of interest, although not to the degree encountered in follow-up studies; and, matching on a non-confounder is expected to lead at worst to a loss in efficiency only in situations of little practical importance.

1. Introduction

In the field of epidemiology, there is probably no more misunderstood and hence misused technique than that of *matching*. This term itself is quite descriptive, since it refers to a method of subject selection which "matches" individuals in a *comparison* group (e.g., unexposed persons in a follow-up study or controls in a case-control study) with those in an *index* series (e.g., the exposed group in a follow-up study or the case group in a case-control study). The goal of such matching is to make the comparison group similar to the index group with respect to the distributions of one or more variables which, although not of primary concern, need to be controlled or adjusted for when describing an exposure-disease relationship of interest. Such extraneous factors are often referred to as potential confounding factors, since their presence, if ignored, may result in a distorted (or confounded) impression of the true exposure-disease relationship.

Many distinguished authors have discussed in one way or another the notions of confounding and matching (see, for example, Fisher and Patil [1974], McKinlay [1977], Miettinen [1970], and Seigel and Greenhouse [1973] to name just a few). In spite of these discussions, there still appears to be confusion concerning exactly what purpose matching serves with regard to the control of confounding and under what circumstances, if any, is matching likely to be a worthwhile enterprise.

It is the goal of this paper to address these two important issues in a quantitative way and to make some definitive recommen-

dations concerning the use of matching in follow-up and case-control studies as it pertains to issues of validity (lack of bias) and efficiency (i.e., precision and power). It will be seen that our discussions and subsequent conclusions and recommendations will very much depend on the type of epidemiologic study design being considered. In this regard, we will focus on the follow-up and case-control study designs, the two designs which most commonly involve the use of matched samples.

In the discussions to follow, we will utilize a probabilistic-based population model (to be described in the next section), and we will reach conclusions regarding the relative merits of random and matched samples by an examination of "expected" cell counts based on the given model. We will restrict our attention to the special case involving a dichotomous *disease* variable (with levels D for present and $\overline{D}$ for absent), a dichotomous *exposure* variable (with levels E and $\overline{E}$ for present or absent, respectively), and a single dichotomous *extraneous factor* F (with F_1 and F_0 denoting its two levels). This is the simplest situation that can be considered and is the one that has been almost exclusively examined by other investigators. Even so, its treatment has previously led to imprecise and sometimes even incorrect conclusions; we hope to remedy this situation here. Extensions of these concepts to the more general and realistic situation involving several mutually correlated extraneous factors is currently underway.

2. Probabilistic Model

In this section, the probabilistic framework is developed for the population model we will consider. The parameters to be utilized in subsequent discussions are best appreciated if introduced separately for the follow-up and case-control study situations.

2a. Follow-up study

In a follow-up study, subjects are selected from each of the two exposure groups. The following probabilities are of interest in this setting. For $i = 0$ and 1, define:

$$\alpha_i = Pr(D|EF_i) \ , \qquad \beta_i = Pr(D|\bar{E}F_i) \ ,$$

$$\theta_{1i} = Pr(F_i|E) \ , \qquad \theta_{0i} = Pr(F_i|\bar{E}) \ ,$$

and $$\psi = Pr(E) .$$

The association between any two variables, either conditional on or averaged over the levels of a third, can be expressed in terms of the parameters just defined.

For a follow-up study, the parameters used to measure the exposure-disease relationship are typically the risk ratio (RR) and the risk difference (RD). In particular, for the i-th stratum of the factor F,

$$RR_i = \frac{Pr(D|EF_i)}{Pr(D|\bar{E}F_i)} = \frac{\alpha_i}{\beta_i} \ ,$$

and

$$RD_i = Pr(D|EF_i) - Pr(D|\bar{E}F_i) = \alpha_i - \beta_i \ .$$

The corresponding *crude* effect measures are

$$cRR = \frac{Pr(D|E)}{Pr(D|\bar{E})} = \frac{\alpha_1\theta_{11} + \alpha_0\theta_{10}}{\beta_1\theta_{01} + \beta_0\theta_{00}} \ ,$$

and

$$cRD = Pr(D|E) - Pr(D|\overline{E}) = (\alpha_1\theta_{11} + \alpha_0\theta_{10}) - (\beta_1\theta_{01} + \beta_0\theta_{00}).$$

Relationships among the variables can also be expressed in terms of odds ratios (OR). In this regard, consider the following two tables based on joint probabilities:

F_1:

	E	$\overline{E}$
D	$\alpha_1\theta_{11}\psi$	$\beta_1\theta_{01}(1-\psi)$
$\overline{D}$	$(1-\alpha_1)\theta_{11}\psi$	$(1-\beta_1)\theta_{01}(1-\psi)$

F_0:

	E	$\overline{E}$
D	$\alpha_0\theta_{10}\psi$	$\beta_0\theta_{00}(1-\psi)$
$\overline{D}$	$(1-\alpha_0)\theta_{10}\psi$	$(1-\beta_0)\theta_{00}(1-\psi)$

By appropriate utilization of the cell and column marginal probabilities for these tables, it is possible to express all the odds ratios of particular interest to us in terms of the α's, β's, and θ's. For example, the exposure-disease odds ratios, conditional on the levels of factor F, are:

$$OR_1 \equiv (OR)_{de|f_1} = \frac{\alpha_1(1-\beta_1)}{\beta_1(1-\alpha_1)}$$

and

$$OR_0 \equiv (OR)_{de|f_0} = \frac{\alpha_0(1-\beta_0)}{\beta_0(1-\alpha_0)} \quad ;$$

note that $OR_i \doteq RR_i$ when α_i and β_i are small.

Without going into further detail, some other odds ratios of future interest in the follow-up study situation are:

$$(OR)_{ef} = \frac{\theta_{11}\theta_{00}}{\theta_{01}\theta_{10}} = \frac{\theta_{11}(1-\theta_{01})}{\theta_{01}(1-\theta_{11})} ;$$

$$(OR)_{df|e} = \frac{\alpha_1(1-\alpha_0)}{\alpha_0(1-\alpha_1)} ;$$

and,

$$(OR)_{df|\bar{e}} = \frac{\beta_1(1-\beta_0)}{\beta_0(1-\beta_1)} .$$

2b. Case-control study

In a case-control study, subjects are selected separately from the case and control groups. The following probabilities are relevant in this framework. For $i = 0$ and 1, define:

$$\varepsilon_i = Pr(E|DF_i) , \qquad \delta_i = Pr(E|\bar{D}F_i) ,$$

$$\gamma_{1i} = Pr(F_i|D) , \qquad \gamma_{0i} = Pr(F_i|\bar{D}) ,$$

and $$\phi = Pr(D).$$

For a case-control study, the parameter used to quantify the strength of the exposure-disease relationship is the odds ratio. Here, the stratum-specific odds ratios OR_1 and OR_0 defined earlier take the form

$$OR_1 = \frac{\varepsilon_1(1-\delta_1)}{\delta_1(1-\varepsilon_1)} \qquad \text{and} \qquad OR_0 = \frac{\varepsilon_0(1-\delta_0)}{\delta_0(1-\varepsilon_0)} ;$$

the corresponding *crude* odds ratio is

$$cOR = \frac{Pr(E|D)Pr(\bar{E}|\bar{D})}{Pr(E|\bar{D})Pr(\bar{E}|D)}$$

$$= \frac{(\varepsilon_1\gamma_{11} + \varepsilon_0\gamma_{10})[(1-\delta_1)\gamma_{01} + (1-\delta_0)\gamma_{00}]}{(\delta_1\gamma_{01} + \delta_0\gamma_{00})[(1-\varepsilon_1)\gamma_{11} + (1-\varepsilon_0)\gamma_{10}]} .$$

Odds ratios which will be useful to us in a case-control study setting are

$$(OR)_{ef|d} = \frac{\varepsilon_1(1-\varepsilon_0)}{\varepsilon_0(1-\varepsilon_1)}$$

and

$$(OR)_{ef|\bar{d}} = \frac{\delta_1(1-\delta_0)}{\delta_0(1-\delta_1)} \quad .$$

Also, with some manipulation of conditional probabilities, it can be shown that the odds ratios $(OR)_{df|e}$ and $(OR)_{df|\bar{e}}$ defined earlier can be equivalently expressed as

$$(OR)_{df|e} = \frac{\varepsilon_1\delta_0}{\varepsilon_0\delta_1}\left(\frac{\gamma_{11}\gamma_{00}}{\gamma_{10}\gamma_{01}}\right)$$

and

$$(OR)_{df|\bar{e}} = \frac{(1-\varepsilon_1)(1-\delta_0)}{(1-\varepsilon_0)(1-\delta_1)}\left(\frac{\gamma_{11}\gamma_{00}}{\gamma_{10}\gamma_{01}}\right) ,$$

where $\gamma_{11}\gamma_{00}/\gamma_{10}\gamma_{01} = (OR)_{df}$.

We have now described the probabilistic structure of our population. In our subsequent discussions on matching, we will be looking at "expected" cell frequencies based on selecting random and matched samples from this population, and such expected frequencies will clearly depend on the study design used.

However, before we can examine matching in any detail, we need to discuss the phenomenon known as *confounding*. Such discussion is necessary in order for us to see what connection, if any, matching has with regard to issues of validity. We will consider the follow-up and case-control study design situations separately.

3. Confounding

3a. Follow-up study

One recommended method for establishing the presence or absence of confounding in a set of data (e.g., see Miettinen [1974] and Rothman [1975]) is to compare the crude effect measure with a "standardized" effect measure; there is said to be confounding or no confounding in the data depending on whether or not the crude and standardized measures differ in value. For follow-up studies, the standardized measure is a weighted sum of the stratum-specific risk ratios or risk differences, with the weights typically being chosen to reflect the distribution of the extraneous factor over the strata among either the unexposed or the exposed subjects. In terms of the risk ratio, the former choice of weights leads to Miettinen's "externally" standardized risk ratio [1972, 1979]

$$s'RR = \frac{\beta_1\theta_{01}\left(\frac{\alpha_1}{\beta_1}\right) + \beta_0\theta_{00}\left(\frac{\alpha_0}{\beta_0}\right)}{\beta_1\theta_{01} + \beta_0\theta_{00}} ;$$

the latter choice of weights leads to his "internally" standardized risk ratio (the well-known standardized mortality or morbidity ratio)

$$sRR \equiv SMR = \frac{\beta_1\theta_{11}\left(\frac{\alpha_1}{\beta_1}\right) + \beta_0\theta_{10}\left(\frac{\alpha_0}{\beta_0}\right)}{\beta_1\theta_{11} + \beta_0\theta_{10}} .$$

(For notational simplicity, we will avoid putting "hats" on parameters to denote sample estimates, although we wish to emphasize that confounding is generally considered to be a property of the sample.)

It is now easy to specify the necessary and sufficient conditions for which $cRR = s'RR$ and for which $cRR = sRR$. In particular, the

former equality holds if and only if either $\alpha_1 = \alpha_0$ or $\theta_{11} = \theta_{01}$; and, the latter equality holds if and only if either $\beta_1 = \beta_0$ or $\theta_{11} = \theta_{01}$. From Section 2, it is clear that the condition $\alpha_1 = \alpha_0$ is equivalent to the condition $(OR)_{df|e} = 1$, that the condition $\beta_1 = \beta_0$ is equivalent to the condition $(OR)_{df|\bar{e}} = 1$, and that the condition $\theta_{11} = \theta_{01}$ is equivalent to the condition $(OR)_{ef} = 1$. The fact that $(OR)_{df|e}$ and $(OR)_{df|\bar{e}}$ represent *conditional* measures of association, as opposed to an *unconditional* measure like $(OR)_{ef}$, is a crucial distinction that has often been overlooked in previous investigations.

It is clear that the conditions for no confounding, as stated above, depend on the choice of weights used to form the standardized measure, which is not an entirely desirable feature. Also, the use of any sort of standardized measure can be seriously questioned when the stratum-specific values vary across the strata. Indeed, the use of a summary index is to be recommended only when there is *approximate* uniformity of the effect measure over the strata. The analogy to the situation when the presence of interaction precludes any worthwhile interpretation of main effects in regression analysis should be apparent. It is our opinion that severe lack of uniformity makes any assessment regarding the presence or absence of confounding somewhat superfluous. In such circumstances, it is best simply to list the stratum-specific values of the effect measure, along with an assessment, if possible and relevant, of a trend in the measure over "levels" of the strata.

McKinlay [1977] illustrated via numerical examples the problems

attendant with using summary indices in the presence of interaction in pair-matched data, but she provides no theoretical treatment of confounding. Seigel and Greenhouse [1975], on the other hand, attempt to develop some purely theoretical results regarding confounding and pair-matching, but they make some misleading statements (e.g., see the Appendix).

If we assume that there is uniformity with respect to risk ratio, then it is easy to see that

$$s'RR = sRR = \frac{\alpha_1}{\beta_1} = \frac{\alpha_0}{\beta_0}$$

and that

$$cRR = \frac{\alpha_1}{\beta_1}\left[\frac{\beta_0 + (\beta_1-\beta_0)\theta_{11}}{\beta_0 + (\beta_1-\beta_0)\theta_{01}}\right] .$$

Thus, it is clear that the necessary and sufficient conditions for no confounding are either $\theta_{11} = \theta_{01}$ or $\beta_1 = \beta_0$ (which implies $\alpha_1 = \alpha_0$, and vice versa). In other words, if either $(OR)_{ef} = 1$ or if $(OR)_{df|\bar{e}} = 1$ (or, equivalently, if $(OR)_{df|e} = 1$), then there is no confounding; otherwise (i.e., these odds ratios all differ from 1 in value), then, strictly speaking, there is confounding.

With regard to risk difference, it is straightforward to show that the assumption $RD_1 = RD_0$ leads to exactly the same necessary and sufficient conditions for no confounding that we just obtained for the risk ratio. In passing, we note that the assumption $(\alpha_1 - \beta_1) = (\alpha_0 - \beta_0)$ implies that $\frac{\alpha_1}{\beta_1} \neq \frac{\alpha_0}{\beta_0}$ unless $\beta_1 = \beta_0$; and, in general, an assumption of uniformity of the risk difference implies non-uniformity with respect to risk ratio, and vice versa. This point has been stressed by Miettinen [1974], Mantel, Brown and Byar [1977],

and Kupper and Hogan [1978], and bears repeating.

3b. Case-control study

If we assume uniformity with respect to odds ratio, then the necessary and sufficient conditions for which

$$OR_1 = OR_0 = cOR$$

are either that

$$\frac{\delta_1(1-\delta_0)}{\delta_0(1-\delta_1)} = 1 \qquad \text{or} \qquad \frac{\varepsilon_1\delta_0}{\varepsilon_0\delta_1}\left(\frac{\gamma_{11}\gamma_{00}}{\gamma_{10}\gamma_{01}}\right) = 1 \ .$$

Since $\frac{\delta_1(1-\delta_0)}{\delta_0(1-\delta_1)} = \frac{\varepsilon_1(1-\varepsilon_0)}{\varepsilon_0(1-\varepsilon_1)}$ and $\frac{\varepsilon_1\delta_0}{\varepsilon_0\delta_1} = \frac{(1-\varepsilon_1)(1-\delta_0)}{(1-\varepsilon_0)(1-\delta_1)}$ under the uniformity assumption, it is clear that the above necessary and sufficient conditions can be expressed in a number of equivalent ways.

Now, from Section 2, we know that the condition $\frac{\delta_1(1-\delta_0)}{\delta_0(1-\delta_1)} = 1$ equivalently means that $(OR)_{ef|\bar{d}} = 1$ (i.e., that there is no exposure-factor F association in the non-diseased group); and, similarly, the condition $\frac{\varepsilon_1(1-\varepsilon_0)}{\varepsilon_0(1-\varepsilon_1)} = 1$ is equivalent to $(OR)_{ef|d} = 1$.

Furthermore, the condition $\frac{\varepsilon_1\delta_0}{\varepsilon_0\delta_1}\left(\frac{\gamma_{11}\gamma_{00}}{\gamma_{10}\gamma_{01}}\right) = 1$ is equivalent to the condition $(OR)_{df|e} = 1$; and, similarly, $\frac{(1-\varepsilon_1)(1-\delta_0)}{(1-\varepsilon_0)(1-\delta_1)}\left(\frac{\gamma_{11}\gamma_{00}}{\gamma_{10}\gamma_{01}}\right) = 1$ is equivalent to $(OR)_{df|\bar{e}} = 1$.

If we do *not* make the assumption $OR_1 = OR_0$, then, as was the case for risk ratio, the specification of necessary and sufficient conditions for no confounding requires consideration of some sort of standardized odds ratio. Without going into any details, if s'OR and sOR denote Miettinen's externally and internally

standardized odds ratios, respectively, then it can be shown that s'OR = cOR if and only if either $(OR)_{ef|\bar{d}} = 1$ or $(OR)_{df|e} = 1$ and that sOR = cOR if and only if either $(OR)_{ef|\bar{d}} = 1$ or $(OR)_{df|\bar{e}} = 1$. Thus, as with risk ratio, the no confounding conditions vary with the choice of standardized measure used.

This now completes our discussion on confounding. A summary of the conditions for no confounding by study design type and by choice of standardized effect measure is presented in Table 1.

TABLE 1

Conditions for No Confounding in Follow-Up and Case-Control Studies by Choice of Standardized Effect Measure Used.

Study Design Type	Follow-up		Case-control	
Standardized Effect Measure	s'RR	sRR	s'OR	sOR
Conditions For No Confounding	$(OR)_{df\|e} = 1$ or $(OR)_{ef} = 1$	$(OR)_{df\|\bar{e}} = 1$ or $(OR)_{ef} = 1$	$(OR)_{df\|e} = 1$ or $(OR)_{ef\|\bar{d}} = 1$	$(OR)_{df\|\bar{e}} = 1$ or $(OR)_{ef\|\bar{d}} = 1$

NOTE: In the absence of interaction for case-control data, $(OR)_{df|e} = (OR)_{df|\bar{e}}$ and $(OR)_{ef|d} = (OR)_{ef|\bar{d}}$.

In conclusion, we reiterate that, practically speaking, confounding is an issue only when it manifests as an attribute of the sample, and that a sample of individuals may exhibit confounding with respect to one or more factors even when there is no "confounding" in the population. Such an undesirable occurrence is simply a manifes-

tation of the method by which the individuals to be studies are selected from that population; e.g., matching and even the vagaries of random sampling can introduce confounding into a sample.

One final comment is in order regarding the question of how to decide on what set of extraneous factors to consider as potential confounders in a study. It is our belief that this set should be restricted to include only those extraneous factors considered by the investigator to be risk factors (i.e., disease determinants). This list should be decided on at the design stage of the study, and the decision should be based on previous empirical evidence and on theoretical knowledge concerning the disease process under investigation. Allowing only risk factors to be potential confounders follows logically from the desired study objective, namely, to assess the effect of the exposure variable on the disease process after controlling for the effects of *established* disease determinants. Note that a risk factor F would necessarily be associated with a non-null value for $(OR)_{df|\bar{e}}$.

In our forthcoming discussions on matching, we will utilize the quantification of confounding given in this section. This is because we will be comparing matching to random sampling as a method for selecting subjects from our population, and such a comparison will be based on examining sample properties determined using expected cell frequencies. We will again treat follow-up and case-control studies separately. Section 4 will discuss matching with regard to issues of validity, while Section 5 will deal with questions of efficiency.

4. Matching: Validity Considerations

4a. Follow-up study

Suppose we select random samples of N_1 exposed and N_0 unexposed individuals from our population. Then, we would obtain, on the average, the following "expected" cell frequencies:

F_1:

	E	$\bar{E}$
D	$N_1\alpha_1\theta_{11}$	$N_0\beta_1\theta_{01}$
$\bar{D}$	$N_1(1-\alpha_1)\theta_{11}$	$N_0(1-\beta_1)\theta_{01}$
	$N_1\theta_{11}$	$N_0\theta_{01}$

F_0:

	E	$\bar{E}$
D	$N_1\alpha_0\theta_{10}$	$N_0\beta_0\theta_{00}$
$\bar{D}$	$N_1(1-\alpha_0)\theta_{10}$	$N_0(1-\beta_0)\theta_{00}$
	$N_1\theta_{10}$	$N_0\theta_{00}$

Because these frequencies are those expected from random sampling, it should be clear that the properties of these "typical" data would *exactly* duplicate those of our hypothetical population. Our reasons for including these rather obvious random sampling results are, firstly, to highlight the contrast with the structure of the corresponding tables based on selecting matched samples, and, secondly, to help in our discussions on efficiency to be presented in Section 5.

To examine the consequences of choosing a matched sample of unexposed subjects, suppose we select, as before, a random sample of N_1 exposed subjects, but now choose the group of N_0 unexposed individuals in such a way that the distribution of the factor F is the same in the sample of N_0 unexposed persons as it is in the sample of N_1 exposed people. Under this sampling scheme, we would obtain the following tables of expected cell counts:

F_1:

	E	$\bar{E}$
D	$N_1\alpha_1\theta_{11}$	$N_0\beta_1\theta_{11}$
$\bar{D}$	$N_1(1-\alpha_1)\theta_{11}$	$N_0(1-\beta_1)\theta_{11}$
	$N_1\theta_{11}$	$N_0\theta_{11}$

F_0:

	E	$\bar{E}$
D	$N_1\alpha_0\theta_{10}$	$N_0\beta_0\theta_{10}$
$\bar{D}$	$N_1(1-\alpha_0)\theta_{10}$	$N_0(1-\beta_0)\theta_{10}$
	$N_1\theta_{10}$	$N_0\theta_{10}$

	E	$\bar{E}$
D	$N_1(\alpha_1\theta_{11}+\alpha_0\theta_{10})$	$N_0(\beta_1\theta_{11}+\beta_0\theta_{10})$
$\bar{D}$	$N_1[(1-\alpha_1)\theta_{11}$ $+(1-\alpha_0)\theta_{10}]$	$N_0[(1-\beta_1)\theta_{11}$ $+(1-\beta_0)\theta_{10}]$
	N_1	N_0

Strictly speaking, the matching scheme described here is that of category or frequency matching. However, the same expected cell counts would have been obtained by first matching on an individual basis (i.e., pair matching) and then ignoring the individual matches within each of the two strata of the factor F (see the Appendix). Most other authors (e.g., Miettinen [1970], Seigel and Greenhouse [1973], and McKinlay [1977]) have chosen to retain the matched pairs in exactly the setting we have described. Although we could have easily done the same, we feel that a matched pairs analysis is inappropriate here because of the inherent nonuniqueness of the pairs formed within each level of the extraneous factor; e.g., within each of the two strata of F, any member of the unexposed group can be paired up with any member of the exposed group without altering the basic within-stratum structure. Such "random" pairing is clearly artificial, and leads (within the framework we are considering) to a smaller χ^2 statistic value (based on McNemar's test) than the appropriate one involving just two strata.

An inspection of the previous tables of expected cell frequencies based on the category matching of unexposed to exposed subjects reveals the following. First of all, the "expected" value of the risk ratio based on the "expected" cell counts for the overall (pooled) table is equal to $(\alpha_1\theta_{11} + \alpha_0\theta_{10})/(\beta_1\theta_{11} + \beta_0\theta_{10})$, which is exactly Miettinen's internally standardized risk ratio sRR. The conclusion to be made here is that the act of category matching itself leads to a valid estimate of the population SMR, without the necessity of performing a stratified analysis. So, the fact that a matched sample has been selected can be ignored at the analysis stage without introducing a bias in the estimation of the SMR. Again, we hasten to reiterate our concern as to the relevance of such a summary index when the stratum-specific risk ratios vary considerably in value. In this regard, note that the two sub-tables do provide the correct stratum-specific risk ratios, namely α_1/β_1 and α_0/β_0; and, if there is uniformity (as was assumed by Seigel and Greenhouse), then, as we noted earlier, $SMR = \alpha_1/\beta_1 = \alpha_0/\beta_0$.

Note that the matching process itself has insured that there will be no confounding in these data due to the factor F, thus obviating the need for a stratified analysis on validity grounds (but not necessarily with regard to efficiency). This is because the unexposed subjects have been selected specifically so that there will be no exposure-factor F association in the data; i.e., the odds ratio is 1 for the table

	E	$\bar{E}$
F_1	$N_1\theta_{11}$	$N_0\theta_{11}$
F_0	$N_1\theta_{10}$	$N_0\theta_{10}$
	N_1	N_0

And, as we know from Section 3a, this is a sufficient condition for no confounding in a follow-up study. Clearly, this matching process eliminates any chance to examine the true exposure-factor F association in the population. In fact, in the population $(OR)_{ef}$ may be much different from 1 in value and sRR may not equal cRR, even though there is no evidence of confounding in the sample. For random sampling, on the other hand, "confounding" in the population would generally be reflected as confounding in the observed data, thus necessitating a stratified analysis.

To summarize, regardless of whether or not we category match in a follow-up study, we can come up with a valid estimate of the population standardized risk ratio. If we match, we are *not* required to perform a stratified analysis on validity grounds; if we do not match, we may have to conduct a stratified analysis to control for confounding. Thus, since a valid estimate of effect can be obtained in either situation, the decision to match or not must necessarily be based solely on efficiency considerations (e.g., issues regarding precision in estimation of effect and/or power in hypothesis testing).

We will now move on to a discussion of the validity issue with regard to category matching in case-control studies. In this setting as well, although we can no longer ignore at the analysis stage the

fact that a matched sample has been selected, we will find that the decision to match or not must be based solely on efficiency considerations.

4b. Case-control study

Suppose we select random samples of N_1 cases and N_0 controls from our population. Then, we would obtain the following tables of "expected" cell frequencies:

F_1:

	E	$\bar{E}$	
D	$N_1\varepsilon_1\gamma_{11}$	$N_1(1-\varepsilon_1)\gamma_{11}$	$N_1\gamma_{11}$
$\bar{D}$	$N_0\delta_1\gamma_{01}$	$N_0(1-\delta_1)\gamma_{01}$	$N_0\gamma_{01}$

F_0:

	E	$\bar{E}$	
D	$N_1\varepsilon_0\gamma_{10}$	$N_1(1-\varepsilon_0)\gamma_{10}$	$N_1\gamma_{10}$
$\bar{D}$	$N_0\delta_0\gamma_{00}$	$N_0(1-\delta_0)\gamma_{00}$	$N_0\gamma_{00}$

As we know, these expected frequencies based on the random sampling of cases and controls provide an exact description of relevant population associations. We will refer again to these tables in Section 5.

To examine the effects of category matching, suppose we select a random sample of N_1 cases, and then choose the N_0 controls in such a way that the distribution of the factor F is the same in the sample of controls as it is in the sample of cases. Under this sampling scheme, we would obtain the following tables of expected cell counts:

F_1:

	E	$\bar{E}$	
D	$N_1\varepsilon_1\gamma_{11}$	$N_1(1-\varepsilon_1)\gamma_{11}$	$N_1\gamma_{11}$
$\bar{D}$	$N_0\delta_1\gamma_{11}$	$N_0(1-\delta_1)\gamma_{11}$	$N_0\gamma_{11}$

F_0:

	E	$\bar{E}$	
D	$N_1\varepsilon_0\gamma_{10}$	$N_1(1-\varepsilon_0)\gamma_{10}$	$N_1\gamma_{10}$
$\bar{D}$	$N_0\delta_0\gamma_{10}$	$N_0(1-\delta_0)\gamma_{10}$	$N_0\gamma_{10}$

	E	$\bar{E}$	
D	$N_1(\varepsilon_1\gamma_{11}+\varepsilon_0\gamma_{10})$	$N_1[(1-\varepsilon_1)\gamma_{11}+(1-\varepsilon_0)\gamma_{10}]$	N_1
$\bar{D}$	$N_0(\delta_1\gamma_{11}+\delta_0\gamma_{10})$	$N_0[(1-\delta_1)\gamma_{11}+(1-\delta_0)\gamma_{10}]$	N_0

An examination of these tables indicates that the effect of category matching in a case-control study is decidedly different from that in a follow-up study. In particular, the "crude" odds ratio obtained from the combined table (i.e., by ignoring the matching) is

$$cOR_m = \frac{(\varepsilon_1\gamma_{11}+\varepsilon_0\gamma_{10})[(1-\delta_1)\gamma_{11}+(1-\delta_0)\gamma_{10}]}{(\delta_1\gamma_{11}+\delta_0\gamma_{10})[(1-\varepsilon_1)\gamma_{11}+(1-\varepsilon_0)\gamma_{10}]} ;$$

this quantity, like cOR, is not a valid estimator of the common stratum-specific odds ratio when there is confounding. Thus, it is apparent that one *cannot* ignore at the analysis stage the fact that a matched sample has been chosen in a case-control study, and still be assured (as in follow-up study) that a valid estimate of the effect measure of interest will be obtained due solely to the matching process itself.

In fact, it can be shown that category matching in a case-control study accomplishes nothing more than random sampling with regard to the control of confounding, and, in certain circumstances, could even introduce confounding into the observed data when random sampling would not. To see all this, one need only compare the odds

ratios $(OR)_{df|\bar{e}}$ and $(OR)_{ef|\bar{d}}$ for the tables of expected cell counts based on matching to those based on random sampling. Thus, category matching in a case-control study has nothing to recommend it over random sampling on validity grounds, since, under either sampling scheme, a stratified analysis would be required to control for any confounding present in the data. Clearly, then, any decision to match in a case-control study must be based solely on efficiency considerations; such considerations will be pursued in Section 5.

5. Matching: Efficiency Considerations

Based on our earlier findings, it is clear that any recommendations regarding whether or not to match either in a follow-up study or in a case-control study will necessarily hinge on considerations of efficiency. Since a properly analyzed random sample preserves validity as well as a correctly handled category matched sample, while also providing a better representation of the population under study, the extra effort required in choosing a matched referent group is worthwhile only if there is expected to be a reasonable gain in statistical efficiency over random sampling. Such a gain in efficiency will be reflected, for example, in terms of increased power of a statistical test procedure (like a Mantel-Haenszel χ^2), or equivalently, in terms of a tighter confidence interval for the effect measure of interest. Our discussions on efficiency will focus on the latter criterion. *Also, although we are aware that issues of cost (e.g., due to labor, time, and to the loss of candidate subjects because of matching constraints) often need to be considered when deciding whether to match or not, we have chosen on grounds of*

simplicity not to deal with such issues in this paper. However, our efficiency studies are currently being extended to deal with such cost considerations.

Before proceeding, however, we wish to point out that previous work concerning the efficiency of matched samples for the case of dichotomous factors (e.g., Worcester [1964], Billewicz [1965], Miettinen [1968, 1969, 1970] and McKinlay [1975]) has not considered as we will here for both follow-up and case-control studies the comparison of confidence intervals for important effect measures (e.g., risk ratios, risk differences, or odds ratios) for matching versus random sampling (with stratification when needed). Worcester [1964], for example, focused on power considerations only and gave conditions for which the McNemar χ^2 statistic for pair-matching would exceed the ordinary crude χ^2 for unmatched data. Billewicz [1965] used simulation techniques exclusively to compare variances of difference effect measures for pair-matched versus unmatched stratified analyses. Using Monte Carlo techniques to compare χ^2 values, McKinlay [1975] contrasted pair-matching to stratification of independent samples, and found that pair-matching did not always lead to a more powerful test.

Miettinen [1968, 1969] considered for follow-up studies only the power efficiency (for difference effect measures) of pair-matched analyses versus unmatched and unstratified analyses in the situation where "matching is not required for validity"; based on our previous conclusions, the phrase in quotes is not meaningful. Moreover, Miettinen's power calculations in the unmatched case were based on

the assumption that "subjects in the two independent comparison series (of equal size) are randomly paired and that the data are then analyzed as in the matched pairs design". Miettinen [1970] has offered intuitive arguments leading to conditions for which random sampling without stratification is preferable to pair-matching in case-control studies; but he provides no quantitative justification and, moreover, does not consider the comparison of category matching to stratification without matching when the latter is required for the control of confounding.

Our work on efficiency involves comparing category matching to random sampling coupled, when required, with stratification, and considers an appropriate efficiency criterion (one based on the variance of the estimator of the effect measure of interest).

Speaking generally for the moment, suppose we let μ denote the unknown parameter representing either the common risk ratio or the common odds ratio in the population (assuming uniformity), as the case may be. If $\hat{\mu}_1$ and $\hat{\mu}_0$ denote the corresponding stratum-specific estimates of μ, then the weighted linear combination

$$\hat{\ell} = w_1(\ell n\hat{\mu}_1) + w_0(\ell n\hat{\mu}_0) \ ,$$

with $(w_1 + w_0) = 1$, would be used to estimate $\ell n\mu$, leading to a $100(1 - \alpha)\%$ large sample confidence interval for μ of the form

$$\exp\left\{\hat{\ell} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{\text{var}}(\hat{\ell})}\right\} .$$

The reason for working with logs is because, for moderate to large samples, $\ell n\hat{\mu}_1$ and $\ell n\hat{\mu}_0$ will tend to be more closely normally

distributed than $\hat{\mu}_1$ and $\hat{\mu}_0$. Further justification (based on extensive simulation work) for the use of this type of transformation has been provided by Katz, Baptista, Azen, and Pike [1978].

To compare category matching with random sampling in regard to efficiency, it is meaningful to consider $\text{Var}(\hat{\ell})$ under the two different sampling schemes. Now, if $\sigma_1^2 = \text{Var}(\ell n\hat{\mu}_1)$ and $\sigma_0^2 = \text{Var}(\ell n\hat{\mu}_0)$, then it can be shown that the choices for w_1 and w_0 which minimize $\text{Var}(\hat{\ell})$ are $\sigma_0^2/(\sigma_1^2 + \sigma_0^2)$ and $\sigma_1^2/(\sigma_1^2 + \sigma_0^2)$, respectively. For this particular weighting scheme, it follows that

$$\text{Var}(\hat{\ell}) = \sigma_0^2\sigma_1^2/(\sigma_0^2 + \sigma_1^2) \ .$$

Similarly, if μ denotes the risk difference, then $\hat{\ell} = w_1\hat{\mu}_1 + w_0\hat{\mu}_0$, the expressions for the weights which minimize $\text{Var}(\hat{\ell})$ are as given except that $\sigma_i^2 = \text{Var}(\hat{\mu}_i)$, and the $100(1-\alpha)\%$ confidence interval is $\hat{\ell} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\hat{\text{Var}}(\hat{\ell})}$.

Now, let σ_M^2 and σ_R^2 denote the values of $\text{Var}(\hat{\ell})$ under category matching and random sampling, respectively; our efficiency results will be based on a theoretical examination of the difference $(\sigma_M^2 - \sigma_R^2)$ and on a comparison of "expected" confidence intervals involving σ_M^2 and σ_R^2 .

However, before proceeding with our discussions of the follow-up and case-control situations, we hasten to point out here that the use of such stratified analysis - based intervals can be both unnecessary and inefficient when there is no confounding; in that instance, the appropriate analysis involves the unstratified (or crude) data layout.

5a. Follow-up study

We will first dispense with the "no confounding" situation. In particular, if either $(OR)_{ef} = 1$ (i.e., $\theta_{11} = \theta_{01}$) or if $(OR)_{df|\bar{e}} = 1$ (i.e., $\beta_1 = \beta_0$), then matching is a futile exercise because the appropriate unstratified analysis results will be the same for matching and random sampling; this follows directly from an inspection of the tables at the beginning of Section 4a.

To make efficiency comparisons when confounding is present, we need to return to the stratified analysis framework introduced earlier. In particular, the large-sample Taylor series approximation to σ_i^2, the variance of $\ell n\widehat{RR}_i$, can be written as

$$\text{Var}(\ell n\widehat{RR}_i) \doteq \frac{[1 - \Pr(D|EF_i)]}{N_{1i}^* \Pr(D|EF_i)} + \frac{[1 - \Pr(D|\bar{E}F_i)]}{N_{0i}^* \Pr(D|\bar{E}F_i)}, \qquad (1)$$

where N_{1i}^* and N_{0i}^* are the "expected" numbers of exposed and unexposed subjects appearing in the i-th stratum, $i = 0, 1$. Similarly, the variance of $\widehat{RD}_i$ can be expressed as

$$\text{Var}(\widehat{RD}_i) = \frac{\Pr(D|EF_i)[1 - \Pr(D|EF_i)]}{N_{1i}^*} + \frac{\Pr(D|\bar{E}F_i)[1 - \Pr(D|\bar{E}F_i)]}{N_{0i}^*}. \qquad (2)$$

The "expected" sample sizes involved in expressions (1) and (2) will vary depending on whether a random or a matched sample is chosen. In fact, N_{1i}^* and N_{0i}^* are the only quantities in expressions (1) and (2) which depend on the sampling scheme; so, any gain in efficiency due to matching in a follow-up study can be attributed directly to the fact that the category matching process itself has introduced some stratum-specific symmetry with regard to subject allocation.

If we let σ^2_{iR} and σ^2_{iM} denote expression (1) with N^*_{1i} and N^*_{0i} replaced by their values under random sampling and category matching, respectively, then an inspection of the tables at the start of Section 4a leads to the following expressions for σ^2_{1R}, σ^2_{0R}, σ^2_{1M}, and σ^2_{0M}:

$$\sigma^2_{1R} = \frac{(1-\alpha_1)}{(N_1\theta_{11})\alpha_1} + \frac{(1-\beta_1)}{(N_0\theta_{01})\beta_1} \quad ;$$

$$\sigma^2_{0R} = \frac{(1-\alpha_0)}{(N_1\theta_{10})\alpha_0} + \frac{(1-\beta_0)}{(N_0\theta_{00})\beta_0} \quad ;$$

$$\sigma^2_{1M} = \frac{(1-\alpha_1)}{(N_1\theta_{11})\alpha_1} + \frac{(1-\beta_1)}{(N_0\theta_{11})\beta_1} \quad ;$$

$$\sigma^2_{0M} = \frac{(1-\alpha_0)}{(N_1\theta_{10})\alpha_0} + \frac{(1-\beta_0)}{(N_0\theta_{10})\beta_0} \quad .$$

Now, for notational simplicity, define

$$a_i = \frac{(1-\alpha_i)}{\alpha_i} \quad , \qquad b_i = \frac{N_1}{N_0}\frac{(1-\beta_i)}{\beta_i} \quad ,$$

$$\rho = \frac{\theta_{11}}{\theta_{01}} \quad , \quad \text{and} \quad \rho' = \frac{(1-\theta_{11})}{(1-\theta_{01})} \quad .$$

Then, $\frac{a_0}{a_1} = (OR)_{df|e}$, $\frac{b_0}{b_1} = (OR)_{df|\bar{e}}$, and $\frac{\rho}{\rho'} = (OR)_{ef}$.

With $\sigma^2_M = \sigma^2_{0M}\sigma^2_{1M}/(\sigma^2_{0M} + \sigma^2_{1M})$ and with σ^2_R defined analogously, some straightforward algebraic manipulation leads to the expression

$$(\sigma^2_M - \sigma^2_R) = K(1-\rho)\left\{\frac{(a_0+b_0)b_1}{(a_1+b_1)b_0}\rho(a_0+b_0\rho') - \rho'(a_1+b_1\rho)\right\} , \qquad (3)$$

where K is a positive constant depending on the a's, b's, θ's and N_1.

Expression (3) pertains to the risk ratio. It should be clear from (1) and (2) that the corresponding expression for the risk difference has exactly the form (3) with a_i replaced by $a_i' = \alpha_i(1 - \alpha_i)$ and b_i replaced by $b_i' = \frac{N_1}{N_0}\beta_i(1 - \beta_i)$.

To proceed further with expression (3), suppose (without loss of generality) we specify F_1 and F_0 so that $\beta_1 > \beta_0$. If $(\beta_0 + \beta_1) < 1$, which is always the situation in practice, then it can be shown (we omit the proof) that the behavior of the sign of expression (3) as a function of $\rho = \theta_{11}/\theta_{01}$ is as depicted in Figure 1.

FIGURE 1

Under these conditions, it can be shown that the region for which $(\sigma_M^2 - \sigma_R^2) > 0$ shrinks as the risk ratio decreases toward 1, and expands as the risk ratio increases away from 1. Also, some numerical results indicate that the magnitudes of the positive values of $(\sigma_M^2 - \sigma_R^2)$ tend to be substantially smaller than the magnitudes of the negative values.

Thus, based just on the above findings, we would be inclined to recommend matching over random sampling as a method of subject selection in follow-up studies when confounding is present and when RR is the effect measure because:

i) the region for which matching is better is larger than the region for which it is worse;

ii) the closer the risk ratio is to 1, the more likely it is that matching will be advantageous; and, this is the very situation in which the variance of the effect estimator needs to be as small as possible to detect only a moderately large risk ratio (say, between

1.5 and 2.5 in value); and,

iii) the negative values of $(\sigma_M^2 - \sigma_R^2)$ are large in magnitude compared to the positive values, indicating that the possible gain in precision is large compared to the possible loss.

As far as the risk difference is concerned, the region for which $\sigma_M^2 > \sigma_R^2$ has a shape similar to that for the RR (see Figure 1), although it can appear below the diagonal. Further theoretical results, however, were much less useful than those for the risk ratio.

Although the above findings concerning the magnitude and sign of $(\sigma_M^2 - \sigma_R^2)$ suggest that matching will often lead to some gain in efficiency in follow-up studies when there is confounding, they do not address the practical issue of whether or not the expected gain is large enough to make any difference in real-life situations. In order to find out, we compared numerically the "expected" 95% confidence intervals based on matching and random sampling for various sets of values of the parameters. For example, we first specified a set of values for β_0, β_1/β_0, N_1, N_0/N_1, and RR (or RD). Then, for the 420 points $(\theta_{11}, \theta_{01})$ over the grid generated by θ_{11} and θ_{01} each ranging from 0.10 to 0.90 in increments of 0.04 with $\theta_{11} \neq \theta_{01}$, we recorded the following quantities:

n_{00} = number of times out of 420 when *both* confidence intervals included the null value (1 for RR and 0 for RD);

n_{01} = number of times out of 420 when only the interval based on random sampling covered the null value;

n_{10} = number of times out of 420 when only the interval based on matching covered the null value; and,

n_{11} = number of times out of 420 when *neither* interval covered the null value.

Thus, for a given set of parameter values for which $RR > 1$ (or $RD > 0$), a large value for n_{00} suggests that neither interval is very sensitive at detecting a true non-null effect, while a large value for n_{11} suggests that both intervals are sensitive in this regard. Furthermore, if $n_{01} > n_{10}$, this is evidence that matching is preferable to random sampling, whereas random sampling gets the nod if $n_{10} > n_{01}$.

We considered the following parameter values:

$$\beta_0 = 0.005,\ 0.05; \quad \beta_1/\beta_0 = 2,\ 3; \quad N_1 = 100,\ 250,\ 500;$$

$$N_0/N_1 = 1,\ 2,\ 3; \quad RR = 1.5,\ 2.5,\ 4.0.$$

The choices for RD were determined as the values of $(\alpha_0 - \beta_0)$, where $\alpha_0 = \beta_0(RR)$; thus, the α_1 values must necessarily be different for the RR and RD cases in order for the uniformity assumption to hold.

Since there are a large number of parameter value combinations under consideration, we have only presented in Tables 2 and 3 illustrative subsets of the outcomes for RR and RD, respectively. However, based on all our numerical findings, the following general statements can be made.

For RR, we can distinguish the following two extreme cases:

i) N_1 and N_0 are small, the α's and β's are small (so that the ratios $(1-\alpha)/\alpha$ and $(1-\beta)/\beta$, which appear in the expression for the variance of the estimator of $\ell n RR$, are large), and the true RR value is not much greater than 1. In this situation, n_{00} is very large, n_{11} is zero, and n_{01} and n_{10} are either zero or very small in value. This suggests that neither matching nor random sampling provides a "sensitive enough" interval.

ii) N_1 and N_0 are fairly large, the α's and β's are moderately large (say, $\beta_0 \geq 0.05$), and RR is fairly large (say, RR ≥ 2.5). Then, n_{11} is very large, n_{00} is zero, and n_{01} and n_{10} are either zero or very small in value. This suggests that either method of subject selection will provide a "sensitive enough" interval.

For situations between these two extremes, the n_{00} and n_{11} values are typically smaller in value (so that n_{01} and/or n_{10} are necessarily larger), and n_{01} *is always greater than* n_{10} (see Table 2 for a typical set of results). **These findings suggest that matching is worthwhile in such fairly common circumstances.**

For RD, a completely analogous pattern emerges, except that the variance of the estimator of RD increases or decreases with the sizes of the products $\alpha(1-\alpha)$ and $\beta(1-\beta)$; e.g., see Table 3.

Also, in each case considered for RR and RD, the confidence interval width based on matching is shorter, on the average, than the corresponding width for random sampling.

Given the limitations to generality imposed by the special framework in which we are working (and ignoring cost considerations), then, based on all our findings, *we recommend matching as a method of subject selection in follow-up studies.* **One can expect a meaningful gain in efficiency when matching on a confounder, and can anticipate no loss in efficiency when matching on a nonconfounder.**

5b. Case-control study

We will address the "no confounding" situation first. If $\delta_1 = \delta_0$, or equivalently $(OR)_{ef|\bar{d}} = 1$, then, from an inspection of the tables at the beginning of Section 4b, it follows that the "expected" unstratified data layouts for matching and for random

sampling will be identical. Hence, matching is unnecessary, since random sampling provides comparable efficiency. The futility of matching in this situation is clear: the equality $(OR)_{ef|\bar{d}} = (OR)_{ef|d} = 1$ means that matching on F will have absolutely no effect on the distribution of the exposure variable in the D and $\bar{D}$ groups.

In contrast, if $(OR)_{ef|\bar{d}} \neq 1$ but $(OR)_{df|\bar{e}} = 1$ (which also implies "no confounding"), then a stratified analysis is required for matched data but not for randomly selected data (again, see the tables at the beginning of Section 4b). This is because the unstratified data layout based on matching provides the crude effect measure cOR_m (defined in Section 4b), which will not equal the uniform odds ratio value (OR, say) unless $\delta_1 = \delta_0$. The need to consider these two case-control study "no confounding" conditions *separately* with regard to efficiency is an added complexity which does not arise in the follow-up study situation.

To compare *equitably* the efficiency of matching to random sampling when $(OR)_{df|\bar{e}} = 1$ but $\delta_1 \neq \delta_0$, we will contrast (using the computer-based approach employed in the follow-up study situation) the "expected" 95% confidence interval based on stratification after matching to the "expected" 95% confidence interval based on the unstratified data layout for random sampling. This latter interval has the form

$$[(OR)e^{-1.96\sigma}, \quad (OR)e^{+1.96\sigma}], \tag{4}$$

where

$$\sigma^2 = \{N_1 Pr(E|D)[1 - Pr(E|D)]\}^{-1} + \{N_0 Pr(E|\bar{D})[1 - Pr(E|\bar{D})]\}^{-1},$$

$\Pr(E|D) = \varepsilon_1\gamma_{11} + \varepsilon_0\gamma_{10}$, and $\Pr(E|\bar{D}) = \delta_1\gamma_{01} + \delta_0\gamma_{00}$. The former interval has a more complex structure for σ than the latter, and so we will defer the discussion of our numerical comparison until we have described the forms of confidence intervals for OR based on stratification after matching and after random sampling.

In particular, the large-sample Taylor-series approximation to σ_i^2, the variance of $\ell n \hat{OR}_i$, can be written as

$$\begin{aligned} \mathrm{Var}(\ell n\, \hat{OR}_i) \doteq & \left\{N_{1i}^* \Pr(E|DF_i)[1 - \Pr(E|DF_i)]\right\}^{-1} \\ & + \left\{N_{0i}^* \Pr(E|\bar{D}F_i)[1 - \Pr(E|\bar{D}F_i)]\right\}^{-1}, \end{aligned} \tag{5}$$

where N_{1i}^* and N_{0i}^* are the "expected" numbers of cases and controls appearing in the i-th stratum, $i = 0, 1$. Then, based on (5) and on the tables at the beginning of Section 4b, the following expressions can be obtained for the four stratum-specific variances:

$$\sigma_{1R}^2 = \frac{1}{(N_1\gamma_{11})\varepsilon_1(1-\varepsilon_1)} + \frac{1}{(N_0\gamma_{01})\delta_1(1-\delta_1)};$$

$$\sigma_{0R}^2 = \frac{1}{(N_1\gamma_{10})\varepsilon_0(1-\varepsilon_0)} + \frac{1}{(N_0\gamma_{00})\delta_0(1-\delta_0)};$$

$$\sigma_{1M}^2 = \frac{1}{(N_1\gamma_{11})\varepsilon_1(1-\varepsilon_1)} + \frac{1}{(N_0\gamma_{11})\delta_1(1-\delta_1)};$$

$$\sigma_{0M}^2 = \frac{1}{(N_1\gamma_{10})\varepsilon_0(1-\varepsilon_0)} + \frac{1}{(N_0\gamma_{10})\delta_0(1-\delta_0)}.$$

Thus, the "expected" 95% confidence intervals for OR based on stratification after matching and after random sampling are of the form (4) with σ^2 replaced by

$$\sigma_M^2 = \sigma_{0M}^2\sigma_{1M}^2/(\sigma_{0M}^2 + \sigma_{1M}^2) \quad \text{and} \quad \sigma_R^2 = \sigma_{0R}^2\sigma_{1R}^2/(\sigma_{0R}^2 + \sigma_{1R}^2),$$

respectively.

The confidence interval evaluation in the "no confounding" situation when $(OR)_{df|\bar{e}} = 1$ but $\delta_1 \neq \delta_0$ compares interval (4) to that same form of interval with σ_M replacing σ; the confidence interval study in the "confounding" situation when both $(OR)_{df|\bar{e}} \neq 1$ and $\delta_1 \neq \delta_0$ compares (4) using σ_M for σ to (4) using σ_R for σ.

For each comparison, the following combinations of parameter values were utilized: OR = 1.5, 2.5, 4.0; N_1 = 25, 50, 100; N_0/N_1 = 1, 2, 3; δ_0 = 0.10, 0.30; $\delta_1 = 1.5\delta_0$, $2\delta_0$, $2.5\delta_0$, $3\delta_0$; $(OR)_{df|\bar{e}} = 1$ for the "no confounding" comparison; $(OR)_{df|\bar{e}}$ = 2, 3, 4, 5 for the "confounding" comparison. For each combination of parameter values, 81 pairs of $(\gamma_{11}, \gamma_{01})$ values were utilized with γ_{11} varying between 0.10 and 0.90 in steps of 0.01 and with γ_{01} then varying subject to the specified constraint on the value of $(OR)_{df|\bar{e}}$. Tables 4 and 5 illustrate typical results of these two comparisons for particular combinations of parameter values.

In the "no-confounding" situation, random sampling always gives a shorter confidence interval than matching for all

combinations of parameter values considered. However, the difference in interval length is of no practical importance except in fairly uncommon situations when δ_0 and OR are both large in value; more specifically, when OR ≥ 2.5 and the exposure is quite common (e.g., 30% and more of the individuals in each one of the four stratified D and $\overline{D}$ groups possess the attribute), then matching can sometimes lead to a meaningful loss in efficiency.

Table 4 provides one example of the following general pattern observed for sets of parameter values of practical importance:

i) if the sample sizes are small, n_{00} is large and n_{11} is zero;

ii) if the sample sizes are large, n_{00} is zero and n_{11} is large;

iii) n_{01} is always zero;

iv) for intermediate sample sizes, n_{10} may be non-zero, but is most likely to be small.

In summary, matching on a particular type of non-confounder can sometimes lead to a loss in efficiency, although such a loss will only be of practical importance when considering an unusually common exposure. Furthermore, since $(OR)_{df|\bar{e}} = 1$ is the no confounding condition under consideration, and this is characteristic of a *non*-risk factor, the policy of considering only disease determinants in a study would help to avoid such over-matching.

In the confounding situation, interpretations of the results are also complicated by the fact that they depend on the sizes of δ_0 and OR.

First, suppose that $\delta_0 = 0.10$. Then, we can identify the following two extreme situations:

i) If the sample sizes and OR are small, then n_{00} is quite large and n_{11} is zero, which means that neither sampling method can be expected to detect an OR just slightly greater than 1.

ii) If the sample sizes and OR are large, then n_{00} is zero and n_{11} is quite large, which means that both methods for choosing controls are sensitive at detecting an OR appreciably greater than 1.

In situations intermediate between these two extremes, n_{01} may be fairly large but n_{10} is always zero (e.g., see Table 5). And, in every case examined for $\delta_0 = 0.10$, the confidence interval based on matching was shorter than that based on random sampling. These findings tend to favor matching over random sampling.

Let us now consider the case when $\delta_0 = 0.30$. If OR is small (about 1.5 or so), then neither interval detects such a non-null value even with fairly large samples. If OR is moderate in size (say, about 2.5), then neither interval is good for small samples, but both are good for intermediate to large samples. However, in the situation when both δ_0 and OR are large, random sampling often provides shorter confidence intervals for OR, the advantage becoming more pronounced with very high exposure rates (e.g., on the order of 60 to 70% or more).

In summary, we would say that matching in case-control studies can provide an important gain in efficiency in the presence of confounding when the exposure is not overly common and when OR is not extremely large; however, the gain will not be to the degree expected in follow-up studies. When the exposure is quite common and OR is large, then random sampling can yield a shorter confidence interval.

Thus, given the limitations to generality imposed by the special framework in which we are working (and ignoring cost considerations), *we recommend that matching be employed in case-control studies as a method of subject selection for small to moderate samples, intermediate values for OR, and for small to moderate exposure rates; this is the set of circumstances most often considered in a case-control study situation.*

5c. Summary of efficiency studies

Table 6 provides a summary of the various analysis comparisons made between matching and random sampling as a function of the type of study design and the nature of the confounding. Table 7 summarizes the various conclusions we have drawn about efficiency based on these comparisons.

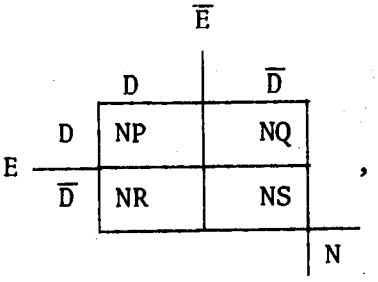
Acknowledgments

Work on this project was partially supported by NIEHS Research Career Development Award #5-K04-ES00003 and by NIEHS Training Grant #1-T32-ES07018.

APPENDIX

Follow-up study

The 2×2 table of "expected" cell frequencies based on the (artificial) *pair matching* of N unexposed subjects to a random sample of N exposed subjects can be shown to have the following structure:



where $P = \alpha_1\beta_1\theta_{11} + \alpha_0\beta_0\theta_{10}$, $Q = \alpha_1(1 - \beta_1)\theta_{11} + \alpha_0(1 - \beta_0)\theta_{10}$,

$R = (1 - \alpha_1)\beta_1\theta_{11} + (1 - \alpha_0)\beta_0\theta_{10}$, and $S = 1 - (P + Q + R)$.

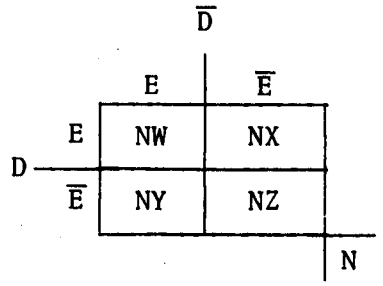
The maximum likelihood estimator of the population *risk ratio* (assuming uniformity of effect) is calculated as the ratio of the observed marginal frequency for row 1 to that of column 1, and has an "expected" value of

$$\frac{(P + Q)}{(P + R)} = \frac{\alpha_1\theta_{11} + \alpha_0\theta_{10}}{\beta_1\theta_{11} + \beta_0\theta_{10}} = \text{SMR} \quad \text{(see Section 3a)}.$$

In the face of non-uniformity, this estimator can be quite misleading (see our earlier remarks on the use of such standardized measures and also the comments and numerical examples of McKinlay [1977]). In such a situation, appropriate analysis of the paired data within each of the two levels of F leads to the correct stratum specific values α_1/β_1 and α_0/β_0.

Case-control study

The 2×2 table of "expected" cell frequencies based on the (artificial) *pair matching* of N controls to a random sample of N cases has the following structure:



,

where $W = \varepsilon_1\delta_1\gamma_{11} + \varepsilon_0\delta_0\gamma_{10}$, $X = \varepsilon_1(1 - \delta_1)\gamma_{11} + \varepsilon_0(1 - \delta_0)\gamma_{10}$,

$Y = (1 - \varepsilon_1)\delta_1\gamma_{11} + (1 - \varepsilon_0)\delta_0\gamma_{10}$, and $Z = 1 - (W + X + Y)$.

The maximum likelihood estimator of the population odds ratio (assuming uniformity of effect) is calculated as the ratio of observed off-diagonal cell counts, and has an "expected" value of

$$\frac{X}{Y} = \frac{\varepsilon_1(1 - \delta_1)\gamma_{11} + \varepsilon_0(1 - \delta_0)\gamma_{10}}{(1 - \varepsilon_1)\delta_1\gamma_{11} + (1 - \varepsilon_0)\delta_0\gamma_{10}}$$

$$= w_1 OR_1 + w_0 OR_0 ,$$

where $w_i = \delta_i(1 - \varepsilon_i)\gamma_{1i} \Big/ \sum_{i=0}^{1} \delta_i(1 - \varepsilon_i)\gamma_{1i}$ for $i = 0, 1$. Thus, as in the follow-up study situation, this estimator loses meaning when the stratum-specific odds ratios differ considerably in value. In any case, stratification with respect to the factor F will provide the correct stratum-specific values OR_1 and OR_0.

REFERENCES

BILLEWICZ, W.Z. (1965). "The Efficiency of Matched Samples: An Empirical Investigation, *Biometrics*, Vol. 21, 623-644.

FISHER, L. and PATIL, K. (1974). "Matching and Unrelatedness", *American Journal of Epidemiology*, Vol. 100, 347-349.

KATZ, D., BAPTISTA, J., AZEN, S.P., and PIKE, M.C. (1978). "Obtaining Confidence Intervals for the Risk Ratio in Cohort Studies", *Biometrics*, Vol. 34, 469-474.

KUPPER, L.L. and HOGAN, M.D. (1978). "Interaction in Epidemiologic Studies", *American Journal of Epidemiology*, Vol. 108, 477-453.

MANTEL, N., BROWN, C., and BYAR, D.P. (1977). "Tests for Homogeneity of Effect in an Epidemiologic Investigation", *American Journal of Epidemiology*, Vol. 106, 125-129.

McKINLAY, S.M. (1975). "A Note on the Chi-square Test for Pair-Matched Samples", *Biometrics*, Vol. 31, 731-735.

McKINLAY, S.M. (1977). "Pair-Matching: A Reappraisal of a Popular Technique", *Biometrics*, Vol. 33, 725-735.

MIETTINEN, O.S. (1968). "The Matched Pairs Design in the Case of All-or-None Responses", *Biometrics*, Vol. 24, 339-352.

MIETTINEN, O.S. (1969). "Individual Matching with Multiple Controls in the Case of All-or-None Responses", *Biometrics*, Vol. 25, 339-355.

MIETTINEN, O.S. (1970). "Matching and Design Efficiency in Retrospective Studies", *American Journal of Epidemiology*, Vol. 91, 111-118.

MIETTINEN, O.S. (1972). "Standardization of Risk Ratios", *American Journal of Epidemiology*, Vol. 96, 383-388.

MIETTINEN, O.S. (1974). "Confounding and Effect-Modification", *American Journal of Epidemiology*, Vol. 100, 350-353.

MIETTINEN, O.S. (1979). *Principles of Epidemiologic Research*. Unpublished manuscript.

ROTHMAN, K.J. (1975). "A Pictorial Representation of Confounding in Epidemiologic Studies", *Journal of Chronic Diseases*, Vol. 28, 101-108.

SEIGEL, D.G. and GREENHOUSE, S.W. (1973). "Validity in Estimating Relative Risk in Case-Control Studies", *Journal of Chronic Diseases*, Vol. 26, 219-225.

WORCESTER, J. (1964). "Matched Samples in Epidemiologic Studies", *Biometrics*, Vol. 20, 840-848.

TABLE 2

Comparisons of Confidence Intervals for RR Based on Matching and Random Sampling: Confounding Present, RR = 1.5, $\beta_0 = 0.05$

N_1	N_0	β_1	$(OR)_{df\mid \bar{e}}$	n_{00}	n_{01}	n_{10}	n_{11}
100	100	.10	2.11	420	0	0	0
250	250	↓	↓	420	0	0	0
500	500	↓	↓	151	110	9	150
100	200	↓	↓	420	0	0	0
250	500	↓	↓	420	0	0	0
500	1,000	↓	↓	0	51	0	369
100	300	↓	↓	420	0	0	0
250	750	↓	↓	280	87	0	53
500	1,500	.10	2.11	0	19	0	401
100	100	.15	3.35	420	0	0	0
250	250	↓	↓	354	51	6	9
500	500	↓	↓	39	51	21	309
100	200	↓	↓	420	0	0	0
250	500	↓	↓	159	84	21	156
500	1,000	↓	↓	0	10	0	410
100	300	↓	↓	420	0	0	0
250	750	↓	↓	100	59	0	261
500	1,500	.15	3.35	0	2	0	418

TABLE 3

Comparisons of Confidence Intervals for RD Based on Matching and Random Sampling: Confounding Present, RD = 0.075, $\beta_0 = 0.05$

N_1	N_0	β_0	$(OR)_{df\|\bar{e}}$	n_{00}	n_{01}	n_{10}	n_{11}
100	100	.10	2.11	420	0	0	0
250	250	↓	↓	0	11	0	409
500	500	↓	↓	0	0	0	420
100	200	↓	↓	300	82	0	38
250	500	↓	↓	0	1	0	419
500	1,000	↓	↓	0	0	0	420
100	300	↓	↓	200	95	0	125
250	750	↓	↓	0	0	0	420
500	1,500	.10	2.11	0	0	0	420
100	100	.15	3.35	420	0	0	0
250	250	↓	↓	0	22	0	398
500	500	↓	↓	0	0	0	420
100	200	↓	↓	360	45	0	15
250	500	↓	↓	0	4	0	416
500	1,000	↓	↓	0	0	0	420
100	300	↓	↓	280	76	0	64
250	750	↓	↓	0	1	0	419
500	1,500	.15	3.35	0	0	0	420

TABLE 4

Comparisons of Confidence Intervals for OR Based on Matching and Random Sampling: No Confounding, OR = 2.5, $\delta_0 = 0.10$, $(OR)_{df|\bar{e}} = 1$

N_1	N_0	δ_1	$(OR)_{ef|\bar{d}}$	n_{00}	n_{01}	n_{10}	n_{11}
25	25	.15	1.59	81	0	0	0
50	50	↓	↓	81	0	0	0
100	100	↓	↓	0	0	0	81
25	50	↓	↓	81	0	0	0
50	100	↓	↓	4	0	1	76
100	200	↓	↓	0	0	0	81
25	75	↓	↓	81	0	0	0
50	150	↓	↓	0	0	0	81
100	300	.15	1.59	0	0	0	81
25	25	.20	2.25	81	0	0	0
50	50	↓	↓	78	0	2	1
100	100	↓	↓	0	0	0	81
25	50	↓	↓	81	0	0	0
50	100	↓	↓	0	0	0	81
100	200	↓	↓	0	0	0	81
25	75	↓	↓	81	0	0	0
50	150	↓	↓	0	0	0	81
100	300	.20	2.25	0	0	0	81

TABLE 5

Comparisons of Confidence Intervals for OR Based on Matching and Random Sampling: Confounding Present, $OR = 2.5$, $\delta_0 = 0.10$, $(OR)_{df|\bar{e}} = 3$

N_1	N_0	δ_1	$(OR)_{ef|\bar{d}}$	n_{00}	n_{01}	n_{10}	n_{11}
25	25	.15	1.59	81	0	0	0
50	50	↓	↓	81	0	0	0
100	100	↓	↓	0	0	0	81
25	50	↓	↓	81	0	0	0
50	100	↓	↓	5	37	0	39
100	200	↓	↓	0	0	0	81
25	75	↓	↓	81	0	0	0
50	150	↓	↓	0	0	0	81
100	300	.15	1.59	0	0	0	81
25	25	.20	2.25	81	0	0	0
50	50	↓	↓	80	1	0	0
100	100	↓	↓	0	0	0	81
25	50	↓	↓	81	0	0	0
50	100	↓	↓	0	18	0	63
100	200	↓	↓	0	0	0	81
25	75	↓	↓	81	0	0	0
50	150	↓	↓	0	0	0	81
100	300	.20	2.25	0	0	0	81

TABLE 6

Summary of Analysis Comparisons as a Function of the Type of Study Design and the Nature of Confounding

		MATCHING	RANDOM SAMPLING
FOLLOW-UP STUDY	NO CONFOUNDING	UNSTRATIFIED	UNSTRATIFIED
	CONFOUNDING	STRATIFIED*	STRATIFIED
CASE CONTROL STUDY	$(OR)_{ef\|\bar{d}} = 1$	UNSTRATIFIED	UNSTRATIFIED
	$(OR)_{ef\|\bar{d}} \neq 1$, $(OR)_{df\|\bar{e}} = 1$	STRATIFIED	UNSTRATIFIED
	CONFOUNDING	STRATIFIED	STRATIFIED

*Although a stratified analysis is not required on validity grounds, it is generally more efficient than an unstratified analysis in this situation.

TABLE 7

Summary of Conclusions about Efficiency Based on Comparisons in Table 6

FOLLOW-UP STUDY	NO CONFOUNDING: No Expected Loss From Matching		CONFOUNDING: Expected Gain From Matching	
CASE-CONTROL STUDY	$(OR)_{ef\|\bar{d}} = 1$	$(OR)_{ef\|\bar{d}} \neq 1$, $(OR)_{df\|\bar{e}} = 1$	OR and exposure rates are small to moderate in value	OR and exposure rates are large in value
	No expected loss from matching	No expected loss from matching except when OR and exposure rates are large in value	Expected gain from matching	Expected gain from random sampling

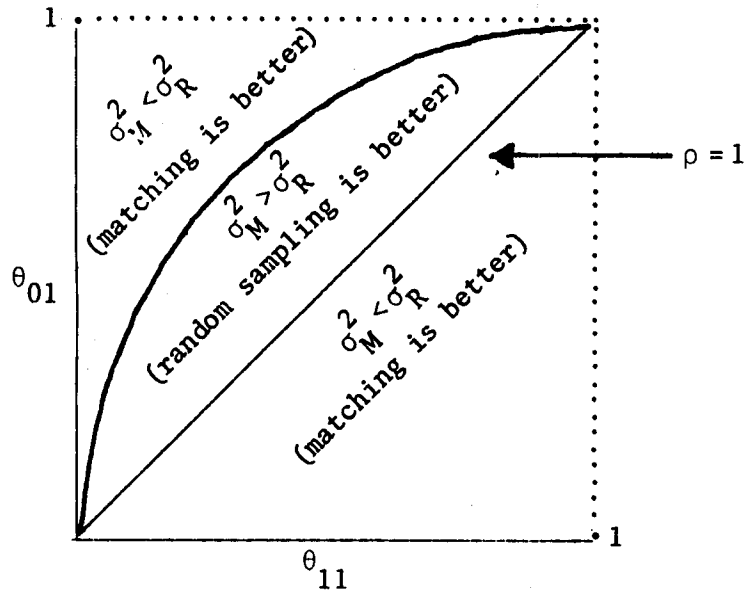


Figure 1

Behavior of the sign of $(\sigma_M^2 - \sigma_R^2)$ as a function of $\rho = \theta_{11}/\theta_{01}$.