

STRATIFICATION BY A MULTIVARIATE CONFOUNDER SCORE

OLLI S. MIETTINEN¹

Miettinen, O. S. (Harvard School of Public Health, Boston, MA 02115). Stratification by a multivariate confounder score. *Am J Epidemiol* 104: 609-620, 1976.

The complexity and inefficiency of multiple cross-classification as a means of controlling confounding in etiologic research may be avoided upon summarizing the pattern of confounding factors for each subject in terms of a multivariate score. The control of confounding may be based on stratification by the score, with stratum-specific contingency tables obtained and analyzed in the usual manner.

epidemiologic methods; biometry

In epidemiologic analysis of nonexperimental data with a view to causation, the classical approach to the control of confounding is stratification (cross-classification) of the subjects by the confounding factors. Within each stratum, the comparison between exposed people and nonexposed reference subjects as to illness outcome or status (in follow-up and prevalence studies, respectively), or of cases of the illness and reference subjects as to exposure status (in case-referent studies), is regarded as free of confounding by the stratification factors. Thus, the analysis of the data is a matter of stratum-specific comparisons and of accumulation of comparative information over the strata to attain an overall comparison free of confounding by the stratification factors. In this classical approach, the primary statistical issues have been taken to be the assessment of a *p*-value for the residual

association between the exposure and the disease, and the estimation of the degree of residual association—most commonly in terms of the rate ratio (“relative risk”) parameter.

The classical analysis with stratification involves a characteristic problem: even with only a few confounding factors the cross-classification involves a large number of strata and, thus, great complexity and inefficiency of analysis. The inefficiency arises from the paucity of subjects within the strata, as this leads to wide variability in the propositus-referent ratio among the strata.

A solution to this problem has been sought through the use of multivariate models and analytic technics, but this solution involves problems of its own. For one, multivariate analysis depends on assumptions whose tenability is often a matter of considerable concern. Another shortcoming of multivariate analysis is that not only is it somewhat unintelligible to many epidemiologists but it generally leaves the investigator with little or no direct insight into the data behind the *p*-values and estimates provided by the analysis.

An approach designed to exploit the virtues and to minimize the drawbacks of both the classical stratification analysis

Received for publication April 20, 1976, and in final form July 30, 1976.

¹Departments of Epidemiology and Biostatistics, Harvard School of Public Health, and Department of Cardiology, Children's Hospital Medical Center, Boston, MA 02115.

The computations were carried out by R. K. Neff, using his superlative Console-oriented Programming System for multivariate analysis and other computing procedures that he has developed.

Supported by grants 5 P01 CA 06373 and HE 10436 from the National Institutes of Health.

and the modern multivariate technics is a combination of these two in the form of stratification by a multivariate confounder-summarizing score. In this approach, the confounding factors to be controlled are summarized in terms of a single (unidimensional) score, a few strata are formed on the basis of this confounder score, and the analysis then proceeds as in the classical analysis by cross-classification—as if, say, age alone were being controlled.

This mode of analyzing epidemiologic data has not yet been expressly delineated and discussed in the literature, even though its first published application (1) is already three years old. In the meantime, it has gained application in several other studies by us and others, but, at the same time, it has given rise to various questions about the principles involved and even about its very *raison d'être vis-a-vis* plain multivariate analysis. It is the purpose here to address these matters.

POOLABILITY

Stratification of the study subjects according to a confounder-summarizing score may be thought of as amounting to pooling the strata from complete cross-classification into a few, larger strata characterized by internal uniformity (or nearly so) of the score. Thus it is apparent that the principle of constructing the stratification score depends on the conditions on which two (or more) strata can be pooled without thereby introducing confounding by the stratification factor(s).

One sufficient condition for such poolability is that the strata in question have identical proportions of cases among the nonexposed. To see this, consider the propriety of pooling the stratum of “young” men with that of “old” women. As to follow-up and prevalence studies, sufficient component questions are whether for the exposed young men in the study, one could use the (nonexposed) old women as comparands (referents, “controls”), and

conversely, whether for the exposed old women in the study, the young men (without exposure) could serve as referents. Affirmative answers to these questions require interchangeability of referents from the two strata, i.e., identity of the proportions of affected subjects among the nonexposed between the two strata. This criterion must apply in case-referent studies as well, since these studies are quite analogous to follow-up and prevalence studies as to the estimability of parameters (2).

An alternative sufficient condition for poolability of strata is that the proportions of exposed subjects are the same among them in the domain of noncases. The propriety of this criterion is most readily seen by considering a case-referent study: for young male cases the referents can include old female noncases as long as the latter reveal the exposure rate of young male noncases; similarly, for old female cases the referents may include young male noncases as long as the two categories have the same exposure rates among the reference subjects. And again, the applicability of the criterion cannot be confined to one type of study only, since comparable estimates are yielded by follow-up and case-referent studies (2).

Further insight into these and other criteria may be gained by considering in detail the nature of confounding (3) and the quantitation of its effect (4). In particular, the following may be helpful: if in the j^{th} one of the strata at issue there are a_j and b_j exposed and nonexposed cases, respectively, and c_j and d_j exposed and nonexposed noncases, respectively, then the confounder-attributable rate ratio from a case-referent study—ordinarily incidence density ratio (2)—is, as has been explained elsewhere (4), $[(\sum b_j c_j / d_j) / \sum b_j] / (\sum c_j / \sum d_j)$, and this is seen to be unity (implying no confounding) if either b_j / d_j or c_j / d_j is constant over j . In the context of a cohort follow-up or prevalence study the corresponding quantity—cumulative incidence

or prevalence ratio (2)—is $[(\sum b_j c_j / d_j) / (\sum b_j c_j / d_j + \sum c_j)] / [\sum b_j / (\sum b_j + \sum d_j)]$ and this too, is unity when either b_j / d_j or c_j / d_j is constant over j . Another point worthy of special note is that the estimate from the pooled data can be thought of as a weighted average of the estimates obtainable from the elementary strata, and that the weights inherent in these two criteria are determined by the distributions of the exposed and the nonexposed, respectively.

The first one of these criteria suggests the principle of pooling the study subjects across the elementary strata on the basis of a function of the confounders which permits ranking of the study subjects according to how “case-like” they are. The function has no external meaning here; it only has to do with the data at hand (3). More specifically it describes “smoothened” data; it is the “fitted” function which has to do with the proportion of cases among the study subjects, conditional on the realizations for the component confounding factors and the assumption of nonexposure. The consideration of smoothened data is not invoked as a condition for validity but in the interest of extending poolability.

The second criterion for the poolability of subjects across the elementary strata implies that, alternatively, the individual subjects may be pooled according to their values for a scoring function that has to do with the proportion of exposed subjects, conditional on being a noncase, again regardless of whether the condition actually obtains or not.

SCORING

The pooling of the study subjects into larger strata according to the above principles presupposes the development of a scoring function which will then be evaluated individually for each subject in the light of her/his particular profile in terms of the original confounding factors.

According to the above criteria for pool-

ing, the scoring function is either an *outcome function*, indicating how “case-like” a subject is (conditionally on nonexposure), or an *exposure function*, summarizing correlates of exposure (conditionally on not being a case). An outcome function is ordinarily preferable in the sense that it elucidates, as an aside, the determinants of outcome, which tend to be more interesting than those of exposure. However, in case-referent studies with the reference series matched, the control of unmatched confounders in the analysis theoretically requires, for reasons of validity (5), the employment of an exposure function (involving as “independent” variates the matching factors, the added confounders and an indicator of outcome).

The type of (either outcome or exposure) function that suggests itself on the basis of both familiarity and simplicity of fitting (noniterative least squares procedure) is the linear discriminant function for the separation between either the cases and the noncases or the exposed and the nonexposed. The potential confounding factors may be entered as additive and/or interactive (product) terms. Quantitative variates may be entered as such, with appropriate transformations (usually power or logarithmic), or in terms of indicators of various categories. Qualitative variates are treated in terms of indicator variates (whose number is one less than the number of categories). Experience has shown the discriminant function analysis to apply quite adequately even in the case of dichotomous variates (6). Formal evaluations have led to similar conclusions (7).

On the other hand, one might prefer to employ an alternate function, perhaps the quadratic discriminant function (8), the log-linear function (9, 10) or, equivalently, the logistic regression function (11-13). The latter approaches afford a somewhat more thorough control of confounding, but they have the drawbacks of relative unfamiliarity, some conceptual subtlety,

and the need for an iterative procedure of fitting.

Even though the scoring has to do with proportions conditional on either nonexposure (outcome scores) or not being a case (exposure scores), one can nevertheless fit the function to the entire set of subjects as long as the conditioning variate is included in the model (possibly with terms of interaction with the confounding factors). The actual scoring function—an outcome function conditional on (hypothetical or actual) nonexposure or the exposure function reflecting lack of the illness—is then obtained by appropriately fixing the conditioning variate in the fitted function.

(The need to include the conditioning variate in the model in the fitting stage has been puzzling to many epidemiologists, even in the face of the poolability principles delineated above. The consideration of an example has been helpful in such instances. If the effect of smoking on the risk of lung cancer is evaluated, the set of potential confounders to be included in the model might consist of age, sex, and “yellow finger.” Were the exposure, smoking, not included in a function separating cases from noncases, “yellow finger” would tend to have a substantial contribution to the function—even though it is not a confounder (3). On the other hand, if smoking is incorporated in the model, then “yellow finger” tends not to contribute to it, since conditionally on the smoking habit, considered in sufficient detail, “yellow finger” has little or nothing to do with the risk of lung cancer. Thus, the fitted function which incorporates the conditioning variate corresponds to the a priori desideratum of “yellow finger” not being included in the confounder-summarizing function.)

With the initial, full model fitted to the data, the statistical significance of the coefficients for many of the (potential) confounding factors is often found to be quite low. In these situations there may be a temptation to reduce the model in a step-

wise fashion until all the remaining terms have (nominally) “significant” coefficients. Such reduction of the model would tend to defeat the purpose of multivariate control of confounding, since “nonsignificance” does not mean lack of confounding, and the deletion of many “nonsignificant” terms from the model may lead to substantial confounding by the aggregate of the deleted factors. Moreover, the deletion, even if not detrimental, would not serve a purpose of parsimony analogous to that in other contexts. For, the (potential) confounding factors are extraneous to the real issue—conditional association between the exposure and the disease—and no inferences need to be made about these controlled variates.

On the other hand, model reduction does tend to serve efficiency of the analysis (maximization of information about the effect parameter) in situations where the reference series is similar in size to the index series. In these instances it may be desirable to reduce, say, an outcome function so as to make it minimally discriminating between cases and noncases—consistent with no material reintroduction of confounding. The latter may be monitored, in simple terms, by having only the “main effect” of exposure in the model and observing the coefficient (b_e) of this variate during the reduction. The antilog of that coefficient ($\exp(b_e)$) is an estimate (involving rather demanding assumptions) of the exposure-odds ratio between cases and noncases and, therefore, depending on the design, of either the incidence-density ratio, the cumulative-incidence-odds ratio or prevalence-odds ratio between the exposed and the nonexposed (14, 2). Thus, $\exp(b_e) - 1$ is a measure of the effect under study, and in the model reduction one might allow, say, up to a 10 per cent change in it.

Having derived the fitted, appropriately reduced scoring function, its value is computed for each subject.

STRATIFICATION

The score distributions of *propositi* and *referents* are examined, and, based on these distributions, stratum boundaries are specified.

The comparison of *propositi* and *referents* should be confined to that range of the score which is common to both series. The common range of the score to be considered in the stratification analysis might be defined literally as the score interval from the greater of the score minima for *propositi* and *referents* to the lesser of score maxima in these groups. Alternatively, some "outliers" in the score distribution might be excluded before ascertaining the common range of the score (among the retained subjects).

Within the range of comparison, some five strata should generally be enough for adequate control of confounding, given that they are properly defined (15, 16). A simple and commonly satisfactory procedure of specifying the actual strata is to consider first deciles of the distribution of the *propositi* within the common range as the boundaries, to examine the contingency tables relating the illness to the exposure within each of these 10 strata, and then to combine adjacent strata so as to have finally some five strata only. The combinations are made according to the criteria in the section "Poolability." As a check on the admissibility of this final reduction of the number of strata it is helpful to compute the measure of association before and after the reduction to make sure that only a tolerable degree of relative change in the effect measure resulted. By the same token, the measure from the reduced stratification should be compared to the crude one (corresponding to the use of a single stratum only) to make sure that the reduced stratification is relevant. As an alternative approach, one may apply the checking separately to each aggregation of the initial strata, and as the monitoring

criterion one might use a measure of confounding itself (4) instead of a measure of residual association.

EXAMINATION OF THE STRATA

With the subjects grouped into a few strata of the confounder score, characteristics of these groups can, and should, be examined.

One purpose of reviewing the profiles is to evaluate the extent to which stratification by the confounder score indeed secured control of the individual confounders. To this end, the cases and noncases within each stratum of the score should be compared as to their distributions by the various confounding factors incorporated in the score, given that an outcome function was used; when using an exposure function, the exposed would be compared to the nonexposed. If the scoring function is very good and the stratification tight enough, the distributions within any given stratum have, on the null hypothesis, only random differences with regard to each confounding factor incorporated in the score, with the differences averaging out to zero over the strata. In the non-null situation the similarity is expected to hold, strictly speaking, only conditionally—in the context of an outcome function among the nonexposed, and, when dealing with an exposure function, among the noncases.

Another purpose for the examination is to evaluate the generalizability of stratum-specific results. This purpose is served by an overall characterization of each stratum as to the confounding factors and possibly with regard to other factors as well. A refined evaluation of the subjects' profiles for this purpose would be focused on the exposed only, because the effect apparent from any study is, inherently, an effect among exposed subjects in it.

EVALUATION OF THE RESIDUAL ASSOCIATION

The main focus of the analysis is, of course, the evaluation of the residual asso-

ciation between the disease and the exposure, i.e., of the association which remains upon the control of the confounding factors incorporated into the summary score.

The reduced data generally take the form of a set of stratum-specific contingency tables relating the disease and the exposure. In currently established terms, the p -value for the residual association would be based on the (asymptotic) Mantel-Haenszel test (17)—a refinement of a test proposed by Cochran (18)—or Mantel's extension of it (19). The "exact" counterpart of the Mantel-Haenszel test (18) is also feasible by the use of a computer (20). The uniformity of the association parameter might be assessed in terms of either the ordinary test for the slope of a weighted regression line relating the association parameter (or a suitable metameter of it) to scores characterizing the strata; alternatively, an asymptotic likelihood ratio test for the regression coefficient might be employed (1). For the usual instance where the assumption of uniformity of the exposure-odds ratio is adopted in a case-referent study, the technics of estimation have been developed and described in some detail (17, 21, 2).

EXAMPLE OF APPLICATION

The problem

As an example of the application of this approach, consider the case-referent study by Jick et al. (1) concerned with the possible role of coffee-drinking in the etiology of myocardial infarction (MI). On the basis of preliminary analysis it was deemed desirable to consider as potential joint confounding factors the following: age (three separate decades), sex, smoking (current cigarette-smoker, past cigarette-smoker, never smoked, other), history of MI (positive, negative), history of the use of antianginal drugs (positive, negative), history of the use of digitalis (positive, negative), diabetes (present, absent), religion (Jewish, other), season (January

–May, June–September) and hospital (each of 24 hospitals of ascertainment). Despite the large number of subjects—440 cases and 12,319 referents—analysis with stratification in the classical manner would have been overwhelmingly complex and very inefficient, as there were a total of 36,864 distinct elementary strata (with one case for every 84 of them).

Analysis with stratification by multivariate score

Analysis with stratification by a multivariate confounder score was employed in the study. A linear discriminant function for the separation of cases and referents was constructed. It involved each of the above confounding factors as additive terms, all but age treated in terms of indicator variates (one for sex, three for smoking, etc.), together with two additive terms indicating level of coffee-consumption (zero, one-to-five or at least six cups per day); also included were all first-order interaction (product) terms among histories of MI, antianginal drugs and digitalis. The model was fitted to the entire series, i.e., including the exposed as well as the nonexposed. No reduction of the model was performed, as the number of referents was ample in all strata (see below). The fitted outcome function was transformed to the scoring function by fixing the exposure variate (coffee-consumption) at zero cups per day, i.e., by deleting the exposure terms. The score value was computed for each subject, and these values were used as a basis for constructing five strata—in the score range shared by cases and referents—with about one-fifth of the cases in this range in each stratum. For the illustration here, the original data were reanalyzed to include in the analysis all subjects irrespective of coffee-drinking habits, whereas the original multivariate analysis was confined to those drinking either zero or at least six cups of coffee per day. No other changes were made.

The distributions of cases and referents according to the confounder score, and the construction of the strata, are illustrated in figure 1. The total range common to both series, from -0.37 to 6.19 , was considered in the analysis. This meant the exclusion of five cases (with highest scores) and 481 potential referents (with lowest scores). The inter-stratum boundaries—approximate quintiles of the cases in the common range—were 1.04 , 1.56 , 2.47 and 3.47 .

Intrastratum profiles of the subjects in terms of a selected subset of the potential confounding factors are shown in table 1, separately for cases and referents. First, an indication of the discriminating "power" of the summary confounder is obtained by comparing the case-referent ratios among the strata. This is seen to range from $88/6811 = 0.013$ to $82/552 = 0.15$. Even though the trend is substantial, it did not call for attempts at model reduction, since even in the most critical stratum the reference group is still $1/0.15 = 6.7$ -fold relative to the index group. (Case-referent ratios in the vicinity of unity, say in the 0.25 to 4 range, would have prompted efforts to reduce the model according to the princi-

ples discussed in the section "Scoring.") The table also shows the degree of intrastratum comparability between cases and referents with regard to selected individual factors controlled through the composite score. It is seen, for example, that whereas the crude frequencies of positive history for MI for cases and referents were 34 per cent and 11 per cent, respectively, and whereas the general frequency ranged from less than 1 per cent in the first stratum to about 96 per cent in the last stratum, yet within each stratum the frequencies are very similar between cases and referents. In fact, as shown in the table, the standardized overall frequencies based on the five strata (with a uniform standard) were 34 per cent and 37 per cent, respectively. There was, however, some residual incomparability in the first stratum in terms of various factors supposedly controlled by the stratification procedure. This resulted from the fact that the stratum was too wide: the case-referent ratio within that stratum is likely to have been non-uniform, since this ratio had a major trend from the first stratum ($88/6811 = 0.013$) to the next one ($86/2429 = 0.035$). (Instead of using quintile boundaries for the strata, it is better to start from deciles and then to combine adjacent strata according to the principles outlined in the section "Poolability"; see section "Stratification.")

The evaluation of the residual association between coffee consumption and MI is summarized in table 2. Within the strata, the rate (incidence density) ratio estimates are computed in the ordinary manner (2): for those drinking one to five cups per day in the first stratum, it is $62(1596)/16(4424) = 1.4$; etc. The estimates do not show any statistically significant linear trend with the stratum index (1 through 5) when evaluated in terms of an asymptotic likelihood ratio test. It is therefore reasonable to entertain the assumption that the parameter is uniform over the strata and to estimate it as outlined by Gart (21). For signifi-

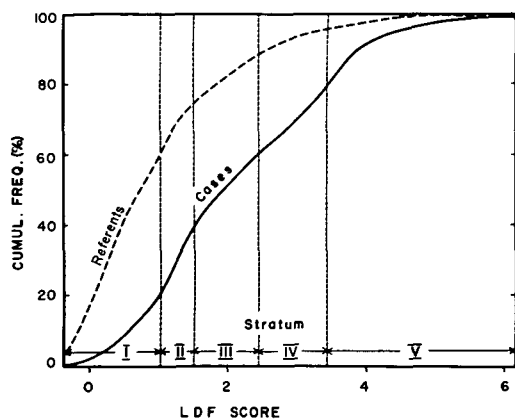


FIGURE 1. Distributions of cases of myocardial infarction and reference patients according to the score for a linear discriminant function (LDF) separating the two series, and the specifications for stratum boundaries (quintiles of the score distribution of cases in the common range).

TABLE 1
Profiles of cases of myocardial infarction and hospital referents in terms of some factors controlled through linear discriminant function score, by strata of the score. Shown are percentages of subjects in selected categories of the factors and total numbers of subjects

Characteristic	Category	Stratum										Total													
		1		2		3		4		5		Crude		Standardized*											
		Cases	Ref- erents	Cases	Ref- erents	Cases	Ref- erents	Cases	Ref- erents	Cases	Ref- erents	Cases	Ref- erents	Cases	Ref- erents										
Percentages of subjects																									
Age	60-69	46	27	32	36	47	48	33	48	43	53	40	34	40	42										
History of myocardial infarction	Positive	1.1	.4	1.1	1.8	13	18	63	67	94	96	34	11	34	37										
History of angina†	Positive	0	.3	0	.9	3	10	24	26	32	27	11	4	12	13										
History of congestive heart failure‡	Present	2	3	2	3	10	11	15	18	11	9	8	5	8	9										
Diabetes	Present	5.7	1.9	5.8	6.6	27	19	13	19	23	27	15	7	15	15										
Religion	Jewish	3	2	8	8	24	22	20	13	18	21	15	7	15	13										
		88		6,811		86		2,429		92		1,375		87		671		82		552		435		11,838	

* The standard was taken to be uniform distribution over the strata.
 † Use of antianginal drugs.
 ‡ Use of digitalis.

TABLE 2
Distributions of cases of myocardial infarction and referent patients according to level of coffee consumption, by level of linear discriminant function score in the range common to cases and referents. Estimates of rate (incidence density) ratio (RR) are also given

Discriminant function category (stratum)	Series	Level of coffee use (cups per day)		
		0	1-5	6+
1	Cases	16	62	10
	Referents	1596	4424	791
	\hat{RR}	(1.0)	1.4	1.3
2	Cases	8	65	13
	Referents	492	1602	335
	\hat{RR}	(1.0)	2.5	2.4
3	Cases	16	63	13
	Referents	323	940	112
	\hat{RR}	(1.0)	1.4	2.3
4	Cases	19	52	16
	Referents	182	423	66
	\hat{RR}	(1.0)	1.2	2.3
5	Cases	11	56	15
	Referents	153	349	50
	\hat{RR}	(1.0)	2.2	4.2
Overall evaluation of RR				
	Gaussian deviate*, †		3.34	4.60
	ML estimate‡	(1.0)	1.6	2.3
	Standardized RR estimate§	(1.0)	1.6	2.4
	90% confidence limits‡		1.3, 2.0	1.7, 3.1

* Mantel-Haenszel test (17).
 † For the dose-response trend—with exposure scored 0, 1 and 2—the corresponding Mantel statistic (19) is 4.54.
 ‡ Following Gart (21). (ML, maximum likelihood.)
 § The values were obtained with each of two standards, all referents and non-coffee-using referents, as described by Miettinen (22).

cance testing, the Gaussian deviates given in the table were based on the Mantel-Haenszel statistic (17). (The square of such a deviate is the Mantel-Haenszel chi square.)

Analysis under a multivariate model

The summary statistics presented on the bottom of table 2 are reproduced in table 3,

augmented by the Mantel test statistic for a linear dose-response trend (19).

Table 3 also shows the corresponding results from an analysis under a multivariate model. The Gaussian deviates are the (linear) discriminant function coefficients for the indicators of exposure categories divided by their standard errors. The dose-specific deviates were derived with a model involving two indicator variates for the three levels of coffee consumption, whereas the test statistic value (4.22) for an overall dose-response trend was obtained by the use of a quantitative exposure term with scoring 0, 1 and 2 (as in the Mantel test). The point estimates for rate (incidence density) ratio under the assumption of uniformity were obtained as the antilogs of the coefficients for the indicator of the level of exposure (14). The corresponding confidence limits were derived as antilogs of the limits for the respective coefficients, with the limits of the coefficients based on the

TABLE 3
Case-referent study of myocardial infarction in relation to coffee use; comparison of results from an analysis with stratification by multivariate score and analysis conducted completely under a multivariate model

Contrast of coffee consumption (cups/day)	Statistic*	Analysis	
		Stratified by multivariate score	Completely multivariate
1-5 vs. 0	Gaussian deviate	3.34	3.24
	RR estimates		
	Uniform RR (assumed)		
	point	1.59	1.48
	90% interval	1.26, 2.01	1.21, 1.81
6+ vs. 0	Nonuniform RR		
	SMR†	1.60	1.22
	Gaussian deviate	4.60	3.97
	RR estimates		
	Uniform RR (assumed)		
6+ vs. 1-5 vs. 0	point	2.29	2.07
	90% interval	1.68, 3.13	1.53, 2.80
	Nonuniform RR		
	SMR	2.35	1.67
	Gaussian deviate	4.54	4.22

* For specifications, see text.
 † Standardized morbidity ratio.

point estimates and their standard errors on the assumption of Gaussian sampling distributions. The "standardized morbidity ratio" (SMR) estimate was computed as the ratio of the observed number of exposed cases to the "expected" number of cases among the exposed, with the latter estimated by the use of the logistic function (6) with sample "priors" and the argument (discriminant function) evaluated at nonexposure.

A comparison of the two sets of statistics in table 3 indicates relative conservatism in the results from the multivariate analysis. The differences are minor, except for the SMR estimates involving the use of the logistic function. This function, when applied to the example at hand, gives gross exaggeration of the proportion of cases in the highest-scores stratum, owing to the positive skewness of the distributions of the discriminant score (figure 1 and table 2).

RELATIVE MERITS

Analysis with stratification by a multivariate confounder-summarizing score represents a hybrid between the classical analysis employing detailed stratification (cross-classification) by the confounders and analysis performed completely under a multivariate model.

When the number of detailed, elementary strata remains small, say a dozen or less, the classical cross-classification is the approach of choice—for reasons of simplicity and wide intelligibility, coupled with still reasonable retention of efficiency.

When the number of elementary strata is large, the classical analysis is generally cumbersome; and in particular, unless one of the compared series (usually the reference series) is several-fold relative to the other series within each of the elementary strata, that type of analysis is also wasteful of information (inefficient). These problems are solved not only by the use of a confounder-summarizing score as the basis for stratification but by plain multivariate analysis as well. This raises the question of

the relative merits of these latter two approaches.

Validity

Control of confounding. The scoring function for stratification is, procedurally, interchangeable with the provision for confounder control in plain multivariate analysis. For example, both may be based on (iterative) maximum likelihood fitting of a multiple logistic function (11-13). Yet adequacy of the control in plain multivariate analysis depends on the model providing a faithful description of the *actual proportion*, smoothened and conditional, of cases or exposed subjects, as the case may be (see section "Scoring"). By contrast, for the stratification approach it suffices to have proper *ranking* of the subjects according to that proportion, i.e., this approach is more robust with regard to any assumptions underlying the multivariate model. On the other hand, if overly heterogeneous strata are employed, the control of confounding can be poor relative to plain multivariate analysis.

Significance testing. In plain multivariate analysis the validity of significance testing, i.e., proper performance of the test statistic in the null case, requires that the sampling distribution of the coefficient of the exposure variate in an outcome function, or of the outcome variate in an exposure function, is Gaussian. (The respective tests are interchangeable given that the same set of confounders are involved.) In particular, exact significance-testing is not feasible. By contrast, stratification by a confounder score permits the use of the exact counterpart (18, 20) of the Mantel-Haenszel test (17) in the evaluation of the conditional (residual) association.

Estimation. A fitted multivariate function can yield, theoretically, a proper point estimate of a particular epidemiologic parameter under certain assumptions. Thus, in a follow-up or prevalence study the coefficient of the exposure variate of a fitted "binary regression" model (23) for

outcome is an estimate of the rate difference over a unit change in the exposure variate—given that the model itself is appropriate. Similarly, if a (multiple) logistic function is fitted to any type of data, the antilog (base e) of the coefficient of the exposure variate in an outcome function, or of the outcome variate in an exposure function, is an estimate of the rate-odds ratio corresponding to a unit change in the exposure variate (14)—given that the model is tenable. The same interpretation can be given to the corresponding discriminant function coefficient insofar as it is feasible to assume (with reliance on robustness) that the variates involved have a Gaussian joint distribution with identical dispersion matrices in the discriminated groups (14, 24). The adequacy of point estimation in the stratification approach is less dependent on the model: stratification always provides for point estimation of the residual association, in terms, e.g., of the maximum likelihood principle. Valid interval estimation is generally inherent in the validity of significance testing.

Efficiency

The sensitivity of significance testing and the precision of estimation are, theoretically and empirically, similar between the two approaches as long as the scoring function is not overly discriminating. If it is, the stratification procedure is less efficient, as it does not exploit, in terms of the model, the information outside the common range of the scores between *propositi* and *referents*. It is to be noted, though, that this efficiency advantage of plain multivariate analysis derives from faith in the model, from *a priori* premises generally based on few or no data and untestable even in the framework of the data at hand.

Verifiability and intelligibility

The adequacy of the control of confounding and the validity of the assessment of the residual association are largely matters of faith, with little opportunity for direct

verification, when the analysis is conducted completely under a multivariate model. By contrast, with the stratification approach the residual confounding may be evaluated directly (see sections “Examination of the Strata” and “Example of Application,” subsection “Analysis with stratification by multivariate score”); also, valid evaluation of the residual association is always feasible (see section “Relative Merits,” subsection “Validity”), with no assumptions of the multivariate model involved. Thus, with the stratification approach an understanding of multivariate analysis is not a prerequisite for evaluation; one has a direct “feel” for both the control of confounding and the residual association.

Overall preference

The above considerations imply that, when dealing with the effect of a particular exposure, plain multivariate analysis might be confined to the preliminary stage of determining what factors are to be jointly controlled. With this decision taken, the ultimate analysis would usually be preferable if based on the proposed principles of stratification by a confounder-summarizing score.

REFERENCES

1. Jick H, Miettinen OS, Neff RK, et al.: Coffee and myocardial infarction. *N Engl J Med* 289:63-67, 1973
2. Miettinen OS: Estimability and estimation in case-referent studies. *Am J Epidemiol* 3:226-235, 1976
3. Miettinen OS: Confounding and effect-modification. *Am J Epidemiol* 100:350-353, 1974
4. Miettinen OS: Components of the crude risk ratio. *Am J Epidemiol* 96:168-172, 1972
5. Miettinen OS: Matching and design efficiency in retrospective studies. *Am J Epidemiol* 91:111-118, 1970
6. Truett J, Cornfield J, Kannel W: A multivariate analysis of the risk of coronary heart disease in Framingham. *J Chronic Dis* 20:511-524, 1967
7. Gilbert ES: On discrimination using qualitative variables. *J Am Stat Assoc* 63:1399-1412, 1968
8. Gilbert ES: The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. *Biometrics* 25:505-515, 1969
9. Birch MW: Maximum likelihood in three-way

- contingency tables. *J Roy Stat Soc B* 26:220-233, 1963
10. Worcester J: The relative odds in the 2³ contingency table. *Am J Epidemiol* 93:145-149, 1971
 11. Cox DR: The regression analysis of binary sequences. *J Roy Stat Soc B* 20:215-232, 1958
 12. Cox DR: Some procedures connected with the logistic qualitative response curve. *Research papers in statistics: Essays in honour of J. Neyman's 70th birthday*. Edited by FN David. London, Wiley, 1966
 13. Walker SH, Duncan DB: Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54:167-179, 1967
 14. Seigel DG, Greenhouse SW: Multiple relative risk functions in case-control studies. *Am J Epidemiol* 97:324-331, 1973
 15. Cox DR: Note on grouping. *J Am Stat Assoc* 52:542-547, 1957
 16. Cochran WG: The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24:295-313, 1968
 17. Mantel N, Haenszel W: Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 22:719-748, 1959
 18. Cochran WG: Some methods of strengthening the common χ^2 tests. *Biometrics* 10:417-451, 1954
 19. Mantel N: Chi square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *J Am Stat Assoc* 58:690-700, 1963
 20. Thomas DT: Exact and asymptotic methods for the combination of 2×2 tables. *Comput Biomed Res* 8: 423-446, 1975
 21. Gart JJ: Point and interval estimation of the common odds ratio in the combination of 2×2 tables with fixed marginals. *Biometrika* 57:471-475, 1970
 22. Miettinen OS: Standardization of risk ratios. *Am J Epidemiol* 96:383-388, 1972
 23. Elwood JH, MacKenzie G: The measurement and comparison of infant mortality risks by binary multiple regression analysis. *J Chronic Dis* 24:93-106, 1971
 24. Halperin M, Blackwelder WC, Verter JI: Estimation of the multivariate logistic risk function: A comparison of the discriminant function and maximum likelihood approaches. *J Chronic Dis* 24:125-158, 1971