Table 7.9 (contd)

C. Dose, duration and other variables

				EST	DOS2	DOS3	DUR1	DUR2	DUR3	DUR4	DUR× DOS	GALL	GALL× EST	NON	OB
1	7	127.63	6.25[d]	2.020	−0.024	1.175	−0.961	−0.131	0.251	0.404					
			5.35[e]	(2.90)	(−0.05)	(2.07)	(−1.18)	(−0.19)	(0.33)	(0.57)					
2	8	127.04	0.59	2.024	−0.835	−0.416	−0.395	−0.555	−0.464	0.254	0.179				
				(2.91)	(−1.03)	(−0.53)	(−0.35)	(−0.39)	(−0.61)	(0.19)	(0.77)				
3	8	118.91	9.54	2.083	−0.725	−0.019	0.470	0.283	−0.049	1.136		1.498			
				(2.81)	(−0.86)	(−0.03)	(0.59)	(0.38)	(−0.09)	(1.95)		(2.93)			
4	9	117.06	1.87	2.433	−0.708	−0.034	0.357	0.276	0.000	1.111		2.531	−1.519		
				(2.99)	(−0.85)	(−0.05)	(0.45)	(0.37)	(0.00)	(1.92)		(2.72)	(−1.35)		
5	9	116.42	2.33	1.951	−0.694	−0.008	0.525	0.220	−0.076	1.114		1.521		0.936	
				(2.59)	(−0.82)	(−0.01)	(0.65)	(0.29)	(−0.14)	(1.90)		(2.96)		(1.50)	
6	9	113.58	5.22	2.195	−0.908	−0.140	0.356	0.231	−0.228	1.242		1.423			1.059
				(2.86)	(−1.04)	(−0.19)	(0.43)	(0.29)	(−0.41)	(2.06)		(2.73)			(2.24)

[a] Based on 54 matched sets, 263 observations having known values for both dose and duration of conjugated oestrogen use
[b] Score test relative to preceding model in each Part, unless otherwise indicated
[c] Relative to Model 1, Part A
[d] Relative to Model 2, Part A
[e] Relative to Model 1, Part B

more complete adjustment for oestrogen than was possible using the binary variable EST alone. The coefficients for these variables should be contrasted with those shown in Table 7.7. Gall-bladder disease continues to stand out as an important, independent risk factor with an estimated relative risk of exp(1.498) = 4.5 compared with the 3.6 found earlier (Model 6, Table 7.7). The interaction of gall-bladder disease with oestrogen use is no longer statistically significant when the dose and duration variables are included in the equation. While the coefficient for non-oestrogen drugs is little changed, obesity is now estimated to carry a relative risk of exp(1.059) = 2.9, which is significantly different from 1 at the p = 0.02 level. Part of these differences, of course, may result because slightly different data sets were used.

In conclusion, we can simply reiterate a point which is well illustrated by the preceding example: all the techniques of multivariate analysis which were once restricted to unmatched studies are now available for use with matched data as well.

7.5 Combining sets of 2×2 tables

Besides individual case-control matching, another situation in which the calculations based on the exact conditional likelihood may be quite feasible is when information is combined from a set of 2×2 tables. We noted earlier that the conditional likelihood in this case took the form of a product of non-central hypergeometric distributions (see § 4.4 for notation):

$$\prod_{i=1}^{I} \frac{\binom{n_{1i}}{a_i}\binom{n_{0i}}{m_{1i}-a_i}\psi_i^{a_i}}{\sum_u \binom{n_{1i}}{u}\binom{n_{0i}}{m_{1i}-u}\psi_i^{u}} . \qquad (7.4)$$

As usual, the summations in the denominator range over all possible values u which are consistent with the observed marginals in the i^th^ table, namely $\max(0, n_{1i}-m_{0i}) \leq u \leq \min(m_{1i}, n_{1i})$. Calculation of exact tail probabilities (4.6, 4.7) and confidence intervals (4.8, 4.9) based on this distribution requires that all possible sets of tables which are compatible with the given marginals are evaluated. Their number is

$$\prod_{i=1}^{I}\{\min(m_{1i},n_{1i})-\max(0,n_{1i}-m_{0i})\},$$

i.e., the *product* of the number of possible tables at each level, which can rapidly become prohibitively large (Thomas, 1975). On the other hand, evaluation of the log-likelihood function and its first and second derivatives requires calculations which increase only in proportion to the *sum*

$$\sum_{i=1}^{I} \{\min(m_{1i},n_{1i})-\max(0,n_{1i}-m_{0i})\}$$

of the number of possible tables at each level. Hence a conditional likelihood analysis, similar to those already developed in this chapter for matched designs, is often possible for problems involving sets of 2×2 tables, even where the completely exact analysis would be unfeasible. Only if the entries in some of the tables are very large will problems be encountered in the evaluation of the binomial coefficients appearing in (7.4).

Usually cases and controls will have been grouped into strata (tables) on the basis of covariables which are thought either to confound or to modify the effect of exposure on disease. Suppose that a vector $\mathbf{z}_i$ of such covariables is associated with the i^{th} table. Then there are several hypotheses about the odds ratios ψ_i which are of interest:

$$H_0:\ \psi_i \equiv 1$$

$$H_1:\ \psi_i \equiv \psi = \exp(\beta)$$

$$H_2:\ \psi_i = \exp(\beta + \Sigma_l \gamma_l z_{il})$$

$$H_3:\ \text{No restrictions on } \psi_i.$$

In Chapter 4 we concentrated on the estimation of ψ under H_1, tests of the null hypothesis H_0, and tests for constancy in the relative risk (H_1) against global alternatives (H_3). We have remarked on several occasions that these latter may be insensitive to particular patterns of interaction and that a preferred strategy is to model specific variations in the relative risk associated with the covariables using H_2. In § 6.12 several such models were fitted to the Oxford Childhood Survey data using unconditional logistic regression in which a separate α parameter was estimated for each stratum. As we saw in § 7.2, however, it is possible seriously to overestimate the relative risk with this procedure if the data are thin. Hence it will often be preferable to use instead the conditional likelihood, which may be written

$$\prod_{i=1}^{I} \frac{\binom{n_{1i}}{a_i}\binom{n_{0i}}{m_{1i}-a_i} \exp\{a_i(\beta + \Sigma_l \gamma_l z_{il})\}}{\sum_u \binom{n_{1i}}{u}\binom{n_{0i}}{m_{1i}-u} \exp\{u(\beta + \Sigma_l \gamma_l z_{il})\}} . \tag{7.5}$$

A listing of a computer programme for fitting models of the form H_2 to sets of 2 × 2 tables using the conditional likelihood is given in Appendix VI. This programme may be used as an alternative to that of Thomas (1975) for finding the exact MLE of the relative risk in H_1, provided of course that exact tests and confidence intervals are not also desired. Zelen (1971) develops exact tests for the constancy of the odds ratio against alternatives of the form H_2 with a single covariable, and also against the global alternative H_3. We presented in (4.31) the score statistic based on (7.5) for testing H_1 against H_2 with a single covariable.

If the data in each table are truly extensive it may be burdensome to evaluate the binomial coefficients in (7.5). In this case an asymptotic procedure is available. Rather than use the exact conditional means and variances of the table entries a_i under hypothesized values for the odds ratios ψ_i, which are required by the iterative likelihood fitting procedure, one can use instead the asymptotic means and variances defined by (4.11) and (4.13). This substitution yields likelihood equations and an information matrix which are identical to those obtained by applying a two-stage maximization procedure to the *unconditional* likelihood function whereby one first solves the equations for the α coefficients in terms of β and γ (Richards, 1961). The estimates $\hat{\beta}$ and $\hat{\gamma}$ so obtained, as well as their standard errors and covariances, are thus identical to those obtained using unconditional logistic regression (Breslow, 1976). The advantage is that the unconditional model is fitted without explicit estimation of all the nuisance parameters. This is a serious consideration if there are many tables, since the required number of parameters may exhaust the capacity of the available computer. Nevertheless, no matter how they are calculated, the unconditional estimates may be subject to bias in such circumstances and the conditional analysis is preferred whenever it is computationally feasible.

To illustrate the use of the conditional likelihood with a set of 2 × 2 tables we found new estimates of the parameters β and γ_1, representing the log relative risk of obstetric radiation and its linear decrease with calendar time, which we estimated earlier from the Oxford Childhood Cancer Survey Data using unconditional logistic regression (6.12). We recall that several estimates for these parameters were made depending on the degree of polynomial adjustment for the stratifying variables age and calendar year. In fact, for the last line in Table 6.17 where the confounding effects of age and year were completely saturated, we avoided explicit estimation of separate α parameters for each of the 120 2 × 2 tables by using the technique just discussed.

The parameter estimates and standard errors calculated directly from the conditional likelihood (7.5) were

$$\hat{\beta} = 0.5165 \pm 0.0564$$

and

$$\hat{\gamma}_1 = -0.0385 \pm 0.0144 .$$

It is of considerable theoretical interest that these quantities are closer to those obtained from the unconditional fifth degree polynomial model than to those obtained with the saturated model (see last two lines, Table 6.17). This suggests that the confounding effects of age and year are suitably accounted for by the polynomial regression, and that inclusion of additional nuisance parameters in the equation serves only to increase bias of the type considered in § 7.1. However, because of the exceptionally large sample (over 5 000 cases and controls) the inflation of the relative risk estimates due to the excess of nuisance parameters was not terribly serious.

7.6 Effect of ignoring the matching

Prior to the advent of methods for the multivariate analysis of case-control studies, in particular those based on the conditional likelihood (7.2), it was common practice to ignore the matching in the analysis. In simple problems one often found that taking explicit account of the matched pairs or sets did not seriously alter the estimate of relative risk. With the Los Angeles study of endometrial cancer, for example, there were only slight differences between the unmatched (Table 7.5) and matched (Table 7.6) estimates for each risk variable considered individually. However, the agreement is not always as good, and there has been considerable confusion regarding the conditions under which incorporation of the matching in the analysis is necessary.

A sufficient and widely-quoted condition for the 'poolability' of data across matched sets or strata is that the *stratification variables are either:* (i) *conditionally independent of disease status given the risk factors; or* (ii) *conditionally independent of the risk factors given disease status.* If either of these conditions is satisfied, both pooled and matched analyses provide (asymptotically) unbiased estimates of the relative risk for a dichotomous exposure (Bishop, Fienberg & Holland, 1975). [Whittemore (1978) has shown that, contrary to popular belief, both types of analyses may sometimes yield equivalent results even if conditions (i) and (ii) are both violated.] In matched studies condition (i) is more relevant since the matching variables are guaranteed to be uncorrelated with disease in the sample as a whole. Of course this does not ensure that they have the same distributions among cases and controls conditionally, within categories defined by the risk factors. Therefore an unmatched analysis may give biased results.

One result of using an unmatched analysis with data collected in a matched design, however, is that the *direction of the bias tends towards conservatism.* Relative risk estimates from the pooled data tend on average to be closer to unity than those calculated from the matched sets. This phenomenon was noted in § 3.4 when pooling data from two 2×2 tables, where the ratio of cases to controls in each table was constant. Seigel and Greenhouse (1973) show that the same thing happens if matched pairs are formed at random from among the cases and controls within each of two strata, and the data are then pooled for analysis. Armitage (1975) gives a slightly more general formulation. He supposes that there are I matched sets with exposure probabilities $p_{1i} = 1-q_{1i}$ for the cases and $p_{0i} = 1-q_{0i}$ for the controls, and that the odds ratio $\psi = p_{1i}q_{0i}/(p_{0i}q_{1i})$ is constant across all sets. It follows that the estimate of relative risk calculated as the cross-products ratio from the 2×2 table formed by pooling all the data tends towards the value

$$\frac{\Sigma p_{1i}\Sigma q_{0i}}{\Sigma p_{0i}\Sigma q_{1i}}$$

$$= \psi\,\frac{\Sigma q_{1i}\vartheta_i\Sigma q_{0i}}{\Sigma q_{0i}\vartheta_i\Sigma q_{1i}} \tag{7.6}$$

where $\vartheta_i = p_{0i}/q_{0i}$. For $\psi>1$ the bias term multiplying ψ in (7.6) is less than one, unless the exposure probabilities p_{0i} are constant across sets (in which case there is no bias). Similarly, for $\psi<1$, the bias term exceeds unity. Thus, failure to account for the matching in the analysis can (and often does) result in conservatively biased estimates of the relative risk.

A related question is to consider the cost, in terms of a loss of efficiency in the analysis, of using a matched analysis when in fact the matching was unnecessary to avoid bias. Suppose that the exposure probabilities p_{0i} in the above model are all equal to the constant p_0, so that both matched and unmatched analyses tend to estimate correctly the true odds ratio ψ. According to (4.18), the large sample variance of the pooled estimate of log ψ is

$$\frac{1}{I}\left\{\frac{1}{p_1}+\frac{1}{q_1}+\frac{1}{p_0}+\frac{1}{q_0}\right\}=\frac{p_1q_1+p_0q_0}{Ip_1q_1p_0q_0}\,.$$

Standard calculations show that the large sample variance of the estimate of log ψ based on the matched pairs in this situation is

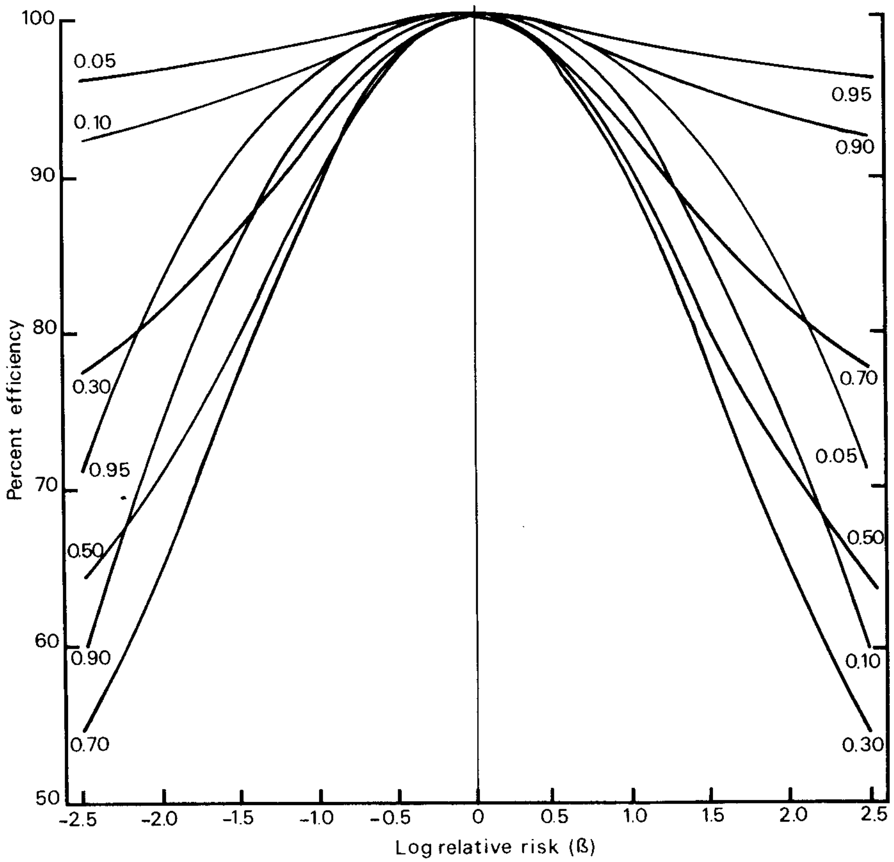
$$\frac{p_1q_0+q_1p_0}{Ip_1q_1p_0q_0}\,.$$

Consequently, using the ratio of variances to measure the relative precision of the two estimates, the efficiency of the matched pairs analysis when pairing at random is

$$\text{eff} = \frac{p_1q_1+p_0q_0}{p_1q_0+p_0q_1}\,. \tag{7.7}$$

When $\psi = 1$, i.e., $p_1 = p_0$, the matched pairs estimate is thus seen to be fully efficient. Otherwise eff < 1, reflecting the loss in information due to the random pairing. Nevertheless Figure 7.1 shows that the loss, which tends to be worse for intermediate values

Fig. 7.1 Loss in efficiency with a matched-pair design of using a matched statistical analysis, when the matching was unnecessary to avoid bias. Different curves correspond to different proportions exposed in the control population.



of p_0, is not terribly important unless the odds ratios being estimated are rather extreme. Pike, Hill and Smith (1979) reach similar conclusions on the basis of studies of the power of the chi-square test of the null hypothesis computed from the matched *versus* unmatched data.

While no additional theoretical studies have yet been made, it is likely that these same general conclusions regarding the bias and efficiency of matched *versus* unmatched *analyses* apply also to the estimation of multiple relative risk functions. Two numerical examples will serve to illustrate the basic points. The first contrasts the fitting of both conditional and unconditional logistic regression analyses to data from an IARC sponsored study of oesophageal cancer occurring among Singapore Chinese (de Jong et al., 1974). The analysis was based on 80 male cases and on 320 matched controls whose ages were within five years of the corresponding case. Two controls for each case were drawn from the same hospital ward as the case, while two others were selected from an orthopaedic unit. However, as there were no important differences in exposure histories between the two control groups, they were not separated in the analysis.

Table 7.10 Coefficients (± standard errors) of variables in the multiple relative risk function, estimated using linear logistic regression analyses appropriate for both matched and unmatched samples. IARC study of oesophageal cancer among Singapore Chinese[a]

Variables in equation[b]	Matched analysis Coefficient ± S.E.	Unmatched analysis Coefficient ± S.E.
A. Interaction term excluded		
x_0 Constant		–3.2062 ± 0.3650
x_1 Dialect	1.2570 ± 0.3273	1.4145 ± 0.3301
x_2 Samsu	0.5064 ± 0.2936	0.5352 ± 0.2766
x_3 Cigarettes	0.0122 ± 0.0099	0.0121 ± 0.0095
x_4 Beverage temperature	0.7846 ± 0.1640	0.7556 ± 0.1493
Goodness-of-fit statistic (G)	197.43	336.23
B. Interaction term included		
x_0 Constant		–3.2123 ± 0.3661
x_1 Dialect	1.2559 ± 0.3280	1.4200 ± 0.3312
x_2 Samsu	0.5072 ± 0.2941	0.5303 ± 0.2774
x_3 Cigarettes	0.0123 ± 0.0099	0.0124 ± 0.0096
x_4 Beverage temperature	0.7872 ± 0.1726	0.7447 ± 0.1563
$x_5 = x_4 \times$ (age-60)	–0.0009 ± 0.0179	0.0034 ± 0.0147
Goodness-of-fit statistic (G)	197.43	336.18

[a] de Jong et al. (1974)

[b] Coding of risk variables:

$x_1 =$ 1 Hokkien/Teochew
0 Cantonese/other

$x_3 =$ No. of cigarettes/day average

$x_2 =$ 1 Drinkers (Samsu)
0 Abstainers

$x_4 =$ No. of beverages (0–3) drunk "burning hot"

Information was obtained regarding diet, alcohol and tobacco usage, and on various social factors including dialect group, which indicates the patient's ancestral origin within China. Only four variables are considered here: dialect group, cigarettes, samsu (a distilled liquor made from a mixture of grains) and beverage temperature (the number of beverages among tea, coffee and barley wine that the patient reported drinking at "burning hot" temperatures). The coding of these variables has been simplified from that used in the original analysis, and an interaction term between beverage temperature and age (a matching variable) was introduced to see if the log relative risk for beverage temperature changed linearly with age.

Table 7.10 presents the estimated regression coefficients and standard errors obtained by fitting the unconditional logistic model with a single stratum parameter α to the pooled data. Shown for comparison are the same quantities estimated from the conditional likelihood. With the exception of that for dialect group, the standard errors of the matched analysis are slightly larger than those for the unmatched. Small changes are evident in the regression coefficients themselves, so that this is evidently a situation in which the matching variables either have little relationship to the exposures conditional on disease status or else have little relationship to disease status conditional on exposure. As a partial confirmation of the latter interpretation, Table 7.11 shows that cases and controls have roughly equivalent average ages even within the levels of each risk factor. This analysis is incomplete, since it involves only averages and ignores possible higher order interactions of age with risk factor combinations. Nevertheless, it is consistent with the notion that the matching variables are conditionally independent of disease status given the exposures, and thus that the requirements for 'poolability'of matched data are satisfied.

Table 7.11 Average ages ± standard errors for cases and controls within levels of each risk factor: IARC study of oesophageal cancer among Singapore Chinese[a]

Risk factor	Level	Cases		Controls		Totals	
		n	Mean ± S.E.	n	Mean ± S.E.	n	Mean ± S.E.
Dialect group	Hokkien/Teochew	66	61.3 ± 1.0	160	60.6 ± 0.8	226	60.8 ± 0.6
	Cantonese/other	14	65.4 ± 2.6	160	63.0 ± 0.7	174	63.2 ± 0.6
Samsu	Drinkers	40	63.6 ± 1.2	109	62.4 ± 0.8	149	62.7 ± 0.7
	Abstainers	40	60.5 ± 1.4	211	61.5 ± 0.6	251	61.4 ± 0.6
Cigarettes	None	8	63.6 ± 5.4	55	62.8 ± 1.3	63	62.9 ± 1.3
	1–10 per day	14	65.9 ± 1.9	81	63.7 ± 1.0	95	64.0 ± 0.9
	11–20 per day	35	61.7 ± 1.0	115	62.2 ± 0.8	150	62.1 ± 0.7
	21+ per day	23	59.6 ± 1.8	69	58.2 ± 1.0	92	58.5 ± 0.9
Beverage temperature (no. "burning hot")	0	41	60.8 ± 1.4	261	61.5 ± 0.6	302	61.4 ± 0.5
	1	13	62.2 ± 2.1	31	62.8 ± 1.6	44	62.6 ± 1.3
	2	18	65.3 ± 1.9	25	63.6 ± 1.9	43	64.3 ± 1.3
	3	8	60.5 ± 2.8	3	66.3 ± 3.2	11	62.1 ± 2.3
Totals	All	80	62.0 ± 0.9	320	61.8 ± 0.5	400	61.9 ± 0.4

[a] de Jong et al. (1974)

Table 7.12 Coefficients (± standard errors) of variables in the multiple relative risk function, using a variety of analyses: Iran/IARC case-control study of oesophageal cancer in the Caspian littoral of Iran[a]

Variables in equation	Type of analysis					
	Fully matched	Stratified into				Unmatched
		7 Regions, 4 Age groups	4 Regions, 4 Age groups	4 Regions	4 Age groups	
Social class	−1.125 ± 0.254	−0.808 ± 0.212	−0.782 ± 0.206	−0.745 ± 0.201	−0.684 ± 0.180	−0.682 ± 0.179
Ownership of garden	−0.815 ± 0.250	−0.614 ± 0.222	−0.602 ± 0.219	−0.592 ± 0.218	−0.326 ± 0.191	−0.307 ± 0.190
Consumption of raw green vegetables	−0.552 ± 0.220	−0.459 ± 0.203	−0.439 ± 0.199	−0.432 ± 0.198	−0.429 ± 0.188	−0.440 ± 0.187
Consumption of cucumbers	−0.640 ± 0.217	−0.539 ± 0.196	−0.548 ± 0.192	−0.562 ± 0.192	−0.466 ± 0.182	−0.449 ± 0.181
Goodness-of-fit (G)	375.38[b]	776.54	777.60	780.80	787.04	789.56

[a] Cook-Mozaffari et al. (1979)

[b] Based on the conditional model and hence not comparable to the others

In general one must anticipate that the degree to which the matching variables are incorporated in the analysis will affect the estimates of relative risk. An example which better illustrates this phenomenon is provided by the joint Iran/IARC study of oesophageal cancer on the Caspian littoral (Cook-Mozaffari et al., 1979). In that part of the world both cancer incidence and many environmental variables show marked geographical variation. Cases and controls were therefore individually matched according to village of residence, as well as for age. Just as in the preceding example, the data were analysed using both the conditional fully matched analysis based on (7.2) and the unconditional analysis based on (6.10) in which the entire sample was considered as a single stratum. Intermediate between these two extremes, additional analyses were performed which incorporated various levels of stratification by age and by geographical area, the latter grouping the villages into regions with roughly homogeneous incidence.

Table 7.12 presents the results for males for four risk variables which appeared to be the best indicators of socioeconomic and dietary status. Substantial bias of the regression coefficients towards the origin, indicating a lesser effect on risk, is evident with the coarsely stratified and unmatched analyses. This confirms the theoretical results regarding the direction of the bias which were noted above to hold for the univariate situation. While the standard errors of the estimates ~~increase as~~ greater account is taken of the matching, the changes are not great and seem a small price to pay for avoiding bias.

In summary, both theoretical and numerical studies confirm that the pooling of matched or stratified samples for analysis will result in relative risk estimates which are conservatively biased in comparison with those which would be obtained using the appropriate matched analysis. In some situations, where the matching was not essential to avoid bias, the pooled and matched estimates may scarcely differ at all. Even then, however, the additional information gained from the pooled data, as reflected in the variances of the estimates, is not great. Consequently, now that appropriate and flexible methods are available for doing so, the *matching should be accounted for in the analysis whenever it has been incorporated in the design.*

While the availability of methods for multivariate analysis of matched samples certainly makes such designs more attractive, it does not follow that they should always be used. Close pair matching may result in a number of cases being lost from the study for want of an appropriate match. It may also impose severe administrative costs which could be avoided with a less restrictive design. Increasing use is being made of "population controls" obtained as an age-stratified random sample of the population from which the cases were diagnosed. Many epidemiologists believe that this is the best way to avoid the selection bias inherent in other choices of the control population. The confounding effects of other factors which are causally related to disease may be accounted for by post-hoc stratification of the sample, or by modelling them in the analysis. Such designs and analyses accomplish many of the aims intended by the use of matching, and constitute a practical alternative which may be preferred in many situations.

REFERENCES

Andersen, E.B. (1973) *Conditional Inference and Models for Measuring*, Copenhagen, Mental Hygienisk Forlag., p. 69

Armitage, P. (1975) *The use of the cross-ratio in aetiological surveys*. In: Gani, J., ed., *Perspectives in Probability and Statistics*, London, Academic Press, pp. 349–355

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, Mass., MIT Press

Breslow, N.E. (1976) Regression analysis of the log odds ratio: a method for retrospective studies. *Biometrics*, *32*, 409–416

Breslow, N.E., Day, N.E., Halvorsen, K.T., Prentice, R.L. & Sabai, C. (1978) Estimation of multiple relative risk functions in matched case-control studies. *Am. J. Epidemiol.*, *108*, 299–307

Cook-Mozaffari, P.J., Azordegan, F., Day, N.E., Ressicaud, A., Sabai, C. & Aramesh, B. (1979) Oesophageal cancer studies in the Caspian littoral of Iran: results of a case-control study. *Br. J. Cancer*, *39*, 293–309

Cox, D.R. & Hinkley, D.V. (1974) *Theoretical Statistics*, London, Chapman & Hall

Day, N.E. & Byar, D. (1979) Testing hypotheses in case-control studies: equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics*, *35*, 623–630

de Jong, U.W., Breslow, N.E., Goh Ewe Hong, J., Sridharan, M. & Shanmugaratnam, K. (1974) Aetiological factors in oesophageal cancer in Singapore Chinese. *Int. J. Cancer*, *13*, 291–303

Fienberg, S.E. (1977) *The Analysis of Cross-Classified Categorical Data*, Cambridge, Mass., MIT Press

Holford, T.R., White, C. & Kelsey, J.L. (1978) Multivariate analysis for matched case-control studies. *Am. J. Epidemiol.*, *107*, 245–256

Jick, H., Watkins, R.N., Hunter, J.R., Dinan, B.J., Madsen, S., Rothman, K.J. & Walker, A.M. (1979) Replacement estrogens and endometrial cancer. *New Engl. J. Med.*, *300*, 218–222

Liddell, F.D.K., McDonald, J.C. & Thomas, D.C. (1977) Methods of cohort analysis: appraisal by application to asbestos mining. *J. R. stat. Soc. Ser. A*, *140*, 469–491

Mack, T.M., Pike, M.C., Henderson, B.E., Pfeffer, R. I., Gerkins, V.R., Arthur, B.S. & Brown, S.E. (1976) Estrogens and endometrial cancer in a retirement community. *New Engl. J. Med.*, *294*, 1262–1267

Miettinen, O.S. (1974) Confounding and effect modification. *Am. J. Epidemiol.*, *100*, 350–353

Pike, M.C., Hill, A.P. & Smith, P.G. (1980) Bias and efficiency in logistic analyses of stratified case-control studies. *Int. J. Epidemiol.*, *9*, 89–95

Richards, F.S.G. (1961) A method of maximum likelihood estimation. *J. R. stat. Soc. B.*, *23*, 469–475

Seigel, D.G. & Greenhouse, S.W. (1973) Multiple relative risk functions in case-control studies. *Am. J. Epidemiol.*, *97*, 324–331

Smith, D.C., Prentice, R., Thompson, D.J. & Herrmann, W.L. (1975) Association of exogenous estrogen and endometrial carcinoma. *New Engl. J. Med.*, *293*, 1164–1167

Smith, P.G., Pike, M.C., Hill, A.P., Breslow, N.E. & Day, N.E. (1981) Multivariate conditional logistic analysis of stratum-matched case-control studies (submitted for publication)

Thomas, D.G. (1975) Exact and asymptotic methods for the combination of 2×2 tables. *Comput. biomed. Res.*, *8*, 423–446

Whittemore, A.S. (1978) Collapsibility of multidimensional contingency tables. *J. R. stat. Soc. B.*, *40*, 328–340

Zelen, M. (1971) The analysis of several 2×2 tables. *Biometrika*, *58*, 129–137

LIST OF SYMBOLS – CHAPTER 7 (in order of appearance)

β_k	log relative risk associated with unit change in kth risk variable
$\mathbf{x}_j$	vector of risk variables for jth study subject; $\mathbf{x}_j = (x_{j1}, \ldots, x_{jk})$
n_1	number of cases
n_0	number of controls
n	total number of study subjects
$\mathbf{l}$	denotes a partition of the integers from 1 to n into two groups, one of size n_1 and the other of size $n_0 = n-n_1$; e.g., if $n_1' = 2$ and $n_0 = 3$ a possible partition is $l_1 = 3, l_2 = 4, l_3 = 1, l_4 = 2, l_5 = 5$ or $\mathbf{l} = (3,4,1,2,5)$
α_i	logit of disease probability for an individual with a standard ($\mathbf{x} = \mathbf{0}$) set of risk variables in the ith stratum
$pr_i(y = 1/x)$	disease probability in the ith stratum for an individual with value x for the risk variable
ψ	odds ratio
β	log relative risk (binary exposure)
n_{00}	number of matched pairs with neither case nor control exposed
n_{01}	number of matched pairs with case unexposed and control exposed
n_{10}	number of matched pairs with case exposed and control unexposed
n_{11}	number of matched pairs with both case and control exposed
μ	in discordant matched pairs with a binary exposure variable, denotes the fitted number of exposed cases under the unconditional model
π	conditional probability that in a discordant matched pair it is the case which is exposed
M	number of controls per case (fixed)
M_i	number of controls per case in the ith matched set
I	number of matched sets
x_{ijk}	value of kth exposure variable ($k = 1, \ldots, K$) for case ($j = 0$) or jth control ($j = 1, \ldots, M_i$) in the ith matched set
$\mathbf{x}_{ij}$	$(x_{ij1}, \ldots, x_{ijK})$ exposure vector for jth subject in ith set
G	goodness-of-fit statistie based on the (conditional) log likelihood
a_i	number of exposed cases in ith of I 2×2 tables
n_{1i}	number of cases in ith table
n_{0i}	number of controls in ith table
ψ_i	(expected) odds ratio associated with ith of I 2×2 tables
z_{il}	value of lth covariable for ith 2×2 table

$\mathbf{z}_i$	vector of covariable values for ith table
γ	vector of interaction parameters in logistic model for a series of 2×2 tables
p_{1i}	exposure probability for cases in the ith stratum
q_{1i}	$1-p_{1i}$
p_{0i}	exposure probability for controls in the ith stratum
q_{0i}	$1-p_{0i}$
ϑ_i	p_{0i}/q_{0i}