STATISTICS IN MEDICINE, VOL. 5, 29–36 (1986)

CHOOSING THE NUMBER OF CONTROLS IN A MATCHED CASE-CONTROL STUDY, SOME SAMPLE SIZE, POWER AND EFFICIENCY CONSIDERATIONS

JEREMY M. G. TAYLOR

Division of Biostatistics, School of Public Health, UCLA, Los Angeles, CA 90024, U.S.A.

SUMMARY

This paper investigates the efficiency of using multiple controls in a case-control study, when there is a single binary exposure variable. Specifically, we consider the asymptotic power of the Cochran[1] test statistic against non-local alternatives of interest. When it is desirable to take multiple controls per case, we show that the marginal return rapidly diminishes as the number of controls per case increases. The effect is as strong, if not stronger, for non-local alternatives as it is for local alternatives. Hence, it is rarely worth choosing more than three controls per case. We also provide a table of sample sizes necessary to achieve 80 per cent power for some odds ratios not equal to one. We extend the results to a special case when there are two binary exposure variables.

KEY WORDS Asymptotic power Case-control studies Contiguity Multiple controls Odds ratio

1. INTRODUCTION

The use of case-control studies with multiple controls is widespread in epidemiologic research. Frequently, one has only a limited number of cases available or they are expensive to obtain, whereas the controls are readily available. However, situations do exist where the controls are just as hard to obtain as are the cases. Under these circumstances it is, practically speaking, unwise to take a large number of controls per case. Gail *et al.*[2] refer to this as the 'rule of diminishing power gains'.

For example, in the Los Angeles area there is a well kept population registry of subjects with skin cancer, so it is relatively easy to obtain a sample from the population of cases. If, however, one chooses neighbourhood controls, then, in terms of subject co-operation and interviewing costs, the controls are probably as expensive as the cases.

We investigate what is a good choice for the number of controls when there is a single dichotomous exposure variable. We are usually interested in detection of odds ratios (ψ) in the range of 1·5 to 4. Values of ψ less than 1·5 have little interest from a practical or biological point of view, and values greater than 4 have little interest from a statistical point of view, because of their easy detection with reasonable sample sizes.

The statistic used to test the hypothesis of no case-control difference is the generalization to multiple controls of McNemar's test,[3] given by Cochran[1], and is also discussed by Miettinen[4] and Pike and Morrow.[5]

0277–6715/86/010029–08$01.00
© 1986 by John Wiley & Sons, Ltd.
Received January 1985
Revised May 1985

Ury[6] considered the asymptotic relative efficiency of two designs with a different number of controls per case. However, his results are largely local, i.e. for alternatives near $\psi = 1$. He showed that the efficiency of a design with J_1 controls relative to a design with J_2 controls is given by $J_1(J_2+1)/J_2(J_1+1)$. He also considered how the large sample practical efficiency differed from the asymptotic relative efficiency for non-local alternatives. Usually the asymptotic relative efficiency does provide a good approximation to the practical efficiency; for some alternatives of interest, however, the difference can be around 40 per cent.

Here we compare two designs by looking at the power of the test statistic against non-local alternatives. Consideration of the power is an alternative to examination of the efficiency. The advantage is that some researchers find probabilities easier to interpret than variances. Miettinen[4] gives an expression for the asymptotic unconditional power function; see also Reference 7. We derive a similar expression and evaluate it at specific alternatives. The expression given here is in the special case where the sample is homogeneous. i.e. we assume equal probability of exposure for all cases. We show further that the expression for the power is reasonably accurate for a more general model involving a non-homogeneous sample.

Miettinen[4] also performed a cost analysis. He looked at the cost of choosing J controls per case for n cases, where c_1 is the cost per case and c_2 is the cost per control, making the total cost $n(c_1 + Jc_2)$. For fixed J he found the sample size that gives a certain power against a local alternative; then, by minimizing the total cost with respect to J he showed that the best choice of J is $(c_1/c_2)^{1/2}$; Gail *et al.*[2] refer to this as the square root rule; see also Reference 7. So, for example, with equal expenses for sampling cases and controls, $J = 1$ is the best choice, and rarely could one justify values of J greater than 3.

In Section 2 we define the test statistics and provide a brief justification; see References 8 and 9 for a more complete discussion. Section 3 contains the power calculations and the tables of the results, including some Monte Carlo results. One can use these tables to help decide on the sample size for a specific alternative. Section 4 gives a mild extension of the work of Breslow and Patton[10] to the case of two binary exposure variables. For this likelihood analysis, it is more convenient to consider efficiency of estimates rather than power of test statistics.

2. TEST STATISTIC

Denote the exposure response by $Y_{ij}, i = 1, \ldots, n; j = 0, \ldots, J$. Y_{i0} refers to the case responses and $Y_{ij}, j \geqslant 1$, refers to the matched control responses.

The null hypothesis, H_0, is that there is no case control difference. Ignoring the continuity correction, a statistic for testing this hypothesis

$$X^2 = \frac{\left\{\sum_{i=1}^{n} (J+1)\, Y_{i0} - n_i\right\}^2}{\sum_{i=1}^{n} n_i(J+1-n_i)} \tag{1}$$

where

$$n_i = \sum_{j=1}^{J} Y_{ij}.$$

Under H_0, X^2 is distributed approximately as χ_1^2.

Cochran proposed this statistic; discussion of it appears elsewhere.[4, 5] It is a generalization of McNemar's test.[3] One can justify the use of this statistic with the assumption of an additive logistic

model for the exposures. Then, conditioning on the sufficient statistics for the nuisance parameters gives the uniformly most powerful unbiased test statistic. Calculation of the conditional mean and variance and use of the normal approximation gives equation (1).

Another justification for this test statistic derives from conditioning on the number of matched sets with exactly m subjects exposed; for example see Reference 11. Under this derivation X^2 appears slightly different, but in fact one can easily show equivalence of the two forms.

3. POWER CALCULATIONS

For simplicity of algebra we parametrize the probabilities as

$$\Pr(\text{case responds}) = \Pr(Y_{i0} = 1) = p + \delta$$
$$\Pr(\text{control responds}) = \Pr(Y_{ij} = 1) = p, j = 1, \ldots, J.$$

Note that we have assumed homogeneity of the sample. Later, we discuss an alternative model in which the probabilities that controls respond differ among controls. We shall see that a reasonable assumption concerning the variability of this probability has little effect on the power of the test statistic. The odds ratio $\psi = (p+\delta)(1-p)/p(1-p-\delta)$, so we can write $\delta = (\psi - 1)p(1-p)/(1-p+p\psi)$.

The hypothesis of interest is $\delta = 0$ or $\psi = 1$. We feel that the parameter ψ has more use and that it is important to express the null and alternative hypothesis in terms of the pair (ψ, p) rather than (δ, p). The algebra, however, is easier with use of (δ, p). Let $X = U/\sqrt{V}$, where

$$U = (1/\sqrt{n}) \sum_{i=1}^{n} [(J+1) Y_{i0} - n_i]$$

$$V = (1/n) \sum_{i=1}^{n} n_i (J + 1 - n_i)$$

and X^2 is the test statistic, as given in equation (1). By simple algebra we can show that unconditional expectation and variance of U are

$$E(U) = J\delta\sqrt{n} \quad \text{and} \quad \text{Var}(U) = J(J+1)p(1-p) + \delta J^2(1-2p) - \delta^2 J^2.$$

Similarly, the expectation of V is

$$E(V) = J(J+1)p(1-p) + \delta J(1-2p).$$

These expressions are similar to those given by Miettinen[4] and Ury;[6] however, we retain all terms of small order in δ.

Consider a contiguous sequence of alternatives,[12] i.e. a sequence of δ such that $\delta \to 0$ but that $\delta\sqrt{n}$ tends to a constant. Under this sequence, U is asymptotically normal and $V \xrightarrow{P} E(V)$ as $n \to \infty$, hence

$$X \xrightarrow{L} N(R, S^2) \quad \text{as} \quad n \to \infty \tag{2}$$

where $R = J\delta\sqrt{n}/\sqrt{E(V)}$ and $S^2 = \text{Var}(U)/E(V)$. The approximation of X by a normal distribution is more accurate for small δ, but by the nature of the asymptotics it is still valid for other values of δ. Also the Monte Carlo results confirm that the approximation seems accurate for non-local alternatives.

The two-sided α-level test of the hypothesis $\psi = 1$ is to reject if $|X| > u_\alpha$, where $\int_{-u_\alpha}^{u_\alpha} \phi(x)\,dx = 1-\alpha$. Using equation (2) we can obtain an expression for the asymptotic power, β, of the test, against a specific alternative for fixed sample size n. The power is

$$\beta = 1 - \Pr\left(\frac{-u_\alpha - R}{S} < Z < \frac{u_\alpha - R}{S}\right)$$

where Z has a standard normal distribution.

Table I gives the values of β for some specific alternatives, when $\alpha = 0{\cdot}05$. For each alternative we chose the value of n such that the power is 0·95 at $p = 0{\cdot}5$, $J = \infty$. So $n = (1{\cdot}96W + 1{\cdot}645(W^2-1)^{1/2})^2$, where $W = (\psi+1)/(\psi-1)$.

The values in the table show the gains achieved by taking more than one control per case. It appears that the gain in power rapidly diminishes as the number of controls increases; this holds true for large values of the odds ratio as well as for values near 1. In most cases there seems little reason to go beyond three controls per case.

Note also that the powers are not symmetric in p, i.e. the power of the test differs for $p = 0{\cdot}2$ compared to $p = 0{\cdot}8$. For odds ratios less than one, we can obtain the power by noting that power (ψ, p) = power $(1/\psi, 1-p)$.

As a check on the asymptotics, we calculated the Monte Carlo power for the configurations in Table I. We performed the Monte Carlo simulations with use of the SAS program on an IBM 4043 computer. For each configuration, we generated 2000 samples. A portion of the results appears in Table II for comparison. We found that the two values of the power closely agreed in nearly every case, the difference being less than 0·03. This held true for small sample sizes and for values of the odds ratio far away from one.

As mentioned previously, the model assumes homogeneity of the sample. Miettinen[4] and Walter[7] considered an alternative model in which Pr(response) = p_{1i} and p_{2i} for cases and controls respectively ($i = 1, \ldots, n$), and where p_{1i} and p_{2i} have a joint distribution with $\theta_1 = E(p_{1i})$, $\theta_2 = E(p_{2i})$ and $\delta = \theta_1 - \theta_2$. Let $\lambda = \theta_1 + \theta_2 - 2\theta_1\theta_2 - 2\mathrm{Cov}(P_1, P_2)$. Miettinen (Reference 4, equation (5.5)) showed that for local alternatives the power is larger for smaller values of λ. Intuitively, this means that the test statistic will have higher power if there is effective matching. Omission of the term $2\mathrm{Cov}(P_1, P_2)$ corresponds to the unmatched analysis presented above. In practice, $2\mathrm{Cov}(P_1, P_2)$ is usually negligible in comparison to $\theta_1 + \theta_2 - 2\theta_1\theta_2$, so, with use of the alternative model, we would expect little difference in the power estimates given in Tables I and II. As a check on this conclusion we performed a second Monte Carlo study whose results appear in parentheses in Table II. In this simulation, for each case/control set, we chose p_{2i} randomly from a uniform distribution, i.e. $p_{2i} \sim U(p-a, p+a)$, where $a = 0{\cdot}5\min(p, 1-(p+\delta))$, and p_{1i} was set equal to $p_{2i} + \delta$. We felt that this would represent a fairly extreme case in the sense that $2\mathrm{Cov}(P_1, P_2)$ was as large as it could reasonably be. This is because the correlation coefficient between P_1 and P_2 is one, and the variance of P_2 is reasonably large. Note that the odds ratio is not constant for every case/control set; we will still take it to be $(p+\delta)(1-p)/p(1-p-\delta)$. We see from the values in Table II that, as expected, there is a marginal but not large increase in power over the unmatched model, at most an increase of 0·03. Consequently, it seems reasonable to use the power calculations from the unmatched model.

Table III gives the number of cases necessary to obtain a specific (80 per cent) power for a specified odds ratio, probability p and number of controls per case = J. The table is useful for designing a study when a particular alternative is of interest. Note that these values are based on the

Table I. Power of X^2 against specific alternatives

		$\psi = 1{\cdot}5$, $n = 319$			$\psi = 2{\cdot}5$, $n = 65$			$\psi = 4{\cdot}0$, $n = 30$		
		p			p			p		
		0·2	0·5	0·8	0·2	0·5	0·8	0·2	0·5	0·8
J	1	0·58	0·72	0·48	0·63	0·69	0·41	0·66	0·66	0·35
	2	0·71	0·83	0·58	0·77	0·82	0·49	0·82	0·80	0·40
	3	0·77	0·88	0·63	0·83	0·87	0·52	0·87	0·85	0·42
	4	0·79	0·90	0·66	0·85	0·89	0·55	0·89	0·88	0·43
	5	0·81	0·91	0·67	0·87	0·90	0·56	0·91	0·89	0·44
	∞	0·88	0·95	0·75	0·93	0·95	0·63	0·96	0·95	0·48

ψ = odds ratio
n = number of cases
p = exposure rate of controls
J = number of controls per case

Table II. Monte Carlo power of X^2 against specific alternatives; P fixed (P random); 2000 simulations

		$\psi = 2{\cdot}5$, $n = 65$			$\psi = 4{\cdot}0$, $n = 30$		
		p			p		
		0·2	0·5	0·8	0·2	0·5	0·8
J	1	0·65 (0·67)	0·70 (0·72)	0·41 (0·42)	0·69 (0·70)	0·68 (0·67)	0·33 (0·34)
	2	0·77 (0·78)	0·82 (0·83)	0·50 (0·51)	0·83 (0·82)	0·81 (0·81)	0·41 (0·42)
	3	0·82 (0·85)	0·85 (0·88)	0·55 (0·53)	0·88 (0·88)	0·87 (0·87)	0·45 (0·43)
	4	0·85 (0·87)	0·90 (0·91)	0·56 (0·55)	0·89 (0·90)	0·87 (0·89)	0·45 (0·45)
	5	0·87 (0·87)	0·89 (0·92)	0·58 (0·58)	0·92 (0·92)	0·89 (0·90)	0·47 (0·47)

ψ = odds ratio
J = number of controls per case
p = exposure rate of controls
n = number of cases

unmatched model, and so are slight overestimates of the necessary sample sizes. Based on the previous analysis, however, use of a matched model would make very little difference to these estimates.

The table shows clearly that when one has the choice, one gains more from an increase in the number of cases, rather than from an increase in the number of controls per case.

4. EXTENSIONS TO TWO COVARIATES

In practice the situation may be more complicated. For example, the exposure variable may not be dichotomous, or there may be more than one exposure variable. For these situations an appropriate method of analysis is maximization of the conditional likelihood from a logistic model (see Reference 10, p. 226).

Table III. Number of cases needed to obtain 80 per cent power

		$\psi = 1{\cdot}5$			$\psi = 2{\cdot}0$		
		p			p		
		0·2	0·5	0·8	0·2	0·5	0·8
	1	537	390	685	174	139	263
	2	396	292	521	127	104	201
J	3	349	259	466	111	92	181
	4	325	243	439	103	86	170
	5	311	233	422	98	82	164
		$\psi = 2{\cdot}5$			$\psi = 3{\cdot}0$		
		p			p		
		0·2	0·5	0·8	0·2	0·5	0·8
	1	97	83	166	66	60	125
	2	70	62	128	47	45	97
J	3	60	55	115	41	39	87
	4	56	51	108	37	37	82
	5	53	49	104	35	35	79

ψ = odds ratio
J = number controls per case
p = exposure rate of controls

The conditional likelihood takes the form

$$\prod_{i=1}^{n} \frac{1}{1 + \sum_{j=1}^{J} \exp\left\{\sum_{k=1}^{K} \beta_k (x_{ijk} - x_{i0k})\right\}}$$

where x_{ijk} is the value of the kth covariate for the case ($j = 0$) or jth control, $j = 1, \ldots, J$; in the ith stratum, $i = 1, \ldots, n$. For this method of analysis it is more convenient to consider changes in efficiency as J changes, rather than changes in power. Breslow and Patton[10] consider a specialization where there is a single binary covariate ($K = 1$), and, by considering the conditional information, they show that the local ($\beta_1 = 0$) efficiency of $1{:}J$ matching compared to $1{:}\infty$ matching is $J/(J+1)$, the same result as Ury's.[6] Note here that the odds ratio is $\exp(\beta_1)$.

We consider a special case where there are two binary covariates ($K = 2$), X_1 and X_2, with a multinomial distribution. Under the null hypothesis that $\beta_1 = \beta_2 = 0$

$$\left.\begin{aligned} &\Pr(X_{ij1} = 1,\ X_{ij2} = 1) = p_i \\ &\Pr(X_{ij1} = 1,\ X_{ij2} = 0) = q_i \\ &\Pr(X_{ij1} = 0,\ X_{ij2} = 1) = r_i \\ &\Pr(X_{ij1} = 0,\ X_{ij2} = 0) = 1 - p_i - q_i - r_i \end{aligned}\right\} \quad i = 1, \ldots, n;\ j = 0, \ldots, J.$$

By employing arguments similar to those of Breslow and Patton[10] (see appendix) and using the 2×2 conditional information matrix, we can show that

$$\text{Asympt. Var}(\hat{\beta}_1 | \beta_1 = 0, \beta_2 = 0) = \frac{J+1}{J}\left[\sum_{i=1}^{n}(p_i+q_i)(1-p_i-q_i) - \frac{\left[\sum_{i=1}^{n} p_i(1-p_i-q_i-r_i)-q_i r_i\right]^2}{\sum_{i=1}^{n}(r_i+p_i)(1-r_i-p_i)}\right]^{-1} \quad (3)$$

For the case where only one covariate is included then

$$\text{Asympt. Var}(\hat{\beta}_1 | \beta_1 = 0) = \frac{J+1}{J}\left[\sum_{i=1}^{n}(p_i+q_i)(1-p_i-q_i)\right]^{-1}. \quad (4)$$

Again the factor $(J+1)/J$ appears in equation (3). Hence in this more complicated situation the same qualitative statements concerning the utility of many controls still apply, namely, the return diminishes rather quickly as J increases. We note that this result is local in two senses; not only is the covariate of interest, X_1, local ($\beta_1 = 0$), but also the second covariate, X_2, is local ($\beta_2 = 0$).

From equations (3) and (4), we note in addition that if the second covariate is unimportant (i.e. $\beta_2 = 0$), then it is better to exclude it from the analysis; the loss in efficiency by its inclusion is

$$1 - \frac{[\sum(p_i(1-p_i-q_i-r_i)-q_i r_i)]^2}{[\sum(r_i+p_i)(1-r_i-p_i)][\sum(p_i+q_i)(1-p_i-q_i)]} = 1 - [\text{Correlation } (X_1, X_2)]^2.$$

5. CONCLUSIONS AND FURTHER DISCUSSION

We have looked at a simple example of a matched case-control study, where there is a single binary exposure variable. By looking at the power of the usual test against a specific alternative we showed that there is a rapidly diminishing return with an increase in the number of controls per case. This is true for local[4,6] and non-local alternatives. Rarely is it worth having more than three controls per case; five or more would seem futile.

With equal difficulty in obtaining cases and controls, it appears that $J = 1$ is the best choice. The results in Tables I, II and III agree with earlier findings that, one gains more from an increase in the number of cases, rather than from an increase in the number of controls per case. We also provide a table of sample sizes of use to an experimenter in designing a study.

In practice the situation may be more complicated. For example, the exposure variable may not be dichotomous, or there may be more than one exposure variable. The result in section 4 gives an indication that, even with more than one exposure variable, one would probably gain little from choosing a large number of controls per case. This is the same conclusion as that for the situation of a single non-dichotomous or continuous variable.[6]

Other difficulties include the inability to match cases, or the situation where the number of cases is fixed and small. But, whatever the situation, the investigator should give careful consideration to the gains, if any, that he would derive from a study with many controls per case.

ACKNOWLEDGEMENTS

The author thanks the referees for their constructive suggestions.

APPENDIX

An outline of derivation of equation (3), is as follows.

Let X_1 and X_2 denote the two binary exposure variables. Each stratum corresponds to a 2×4 table of the form

exposure $(X_1\ X_2)$

	+ +	+ −	− +	− −	
Case					1
Control					R
	t	u	v	w	

Let $S_{t,u,v}$ equal the number of strata where the marginal totals are as given in the 2×4 table. Lengthy algebra shows that the second derivatives with respect to β_1 and β_2 of the log conditional likelihood, at $\beta_1 = \beta_2 = 0$, are

$$\frac{\partial^2 \log(L)}{\partial \beta_1^2} = -\sum_{(t,u,v)} S_{t,u,v} \frac{(t+u)[R+1-(t+u)]}{(R+1)^2}$$

$$\frac{\partial^2 \log(L)}{\partial \beta_1 \partial \beta_2} = \sum_{(t,u,v)} S_{t,u,v} \frac{(wt-uv)}{(R+1)^2}$$

$$\frac{\partial^2 \log(L)}{\partial \beta_2^2} = -\sum_{(t,u,v)} S_{t,u,v} \frac{(t+v)[R+1-(t+v)]}{(R+1)^2}.$$

It is necessary to calculate $E(S_{t,u,v})$ under the null hypothesis; to do this note that the ith stratum contributes a term

$$\binom{R+1}{t,u,v} p_i^t q_i^u r_i^v (1-p_i-q_i-r_i)^w$$

to $E(S_{t,u,v})$. Then inversion of the 2×2 second derivative matrix and selection of the (1,1) term gives equation (3).

REFERENCES

1. Cochran, W. G. 'The comparison of percentages in matched samples', *Biometrika*, **37**, 256–266 (1950).
2. Gail, M., Williams, R., Byar, D. P. and Brown, C. 'How many controls?', *Journal of Chronic Diseases*, **29**, 723–731 (1976).
3. McNemar, Q. 'Note on the sampling error of the difference between corrected proportions or percentages', *Psychometrika*, **12**, 153–157 (1947).
4. Miettinen, O. S. 'Individual matching with multiple controls in the case of all-or-none responses', *Biometrics*, **25**, 339–355 (1969).
5. Pike, W. R. and Morrow, R. H. 'Statistical analysis of patient-control studies in epidemiology. Factor under investigation on all-or-none variable', *British Journal of Preventive Social Medicine*, **24**, 42–44 (1970).
6. Ury, H. K. 'Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data', *Biometrics*, **31**, 643–649 (1975).
7. Walter, S. D. 'Matched case-control studies with a variable number of controls per case', *Applied Statistics*, **29**, 172–179 (1980).
8. Cox, D. R. 'A simple example of a comparison involving quantal data', *Biometrika*, **53**, 215–220 (1966).
9. Gart, J. J. and Tarone, R. E. 'The relation between score tests and approximate UMPU tests in exponential models common in Biometry', *Biometrics*, **39**, 781–786 (1983).
10. Breslow, N. E. and Patton, J. 'Case-control analysis of cohort studies', *Energy and Health*, Society for Industrial and Applied Mathematics, 1979.
11. Breslow, N. E. and Day, N. E. *Statistical Methods in Cancer Research*, Vol. 1, International Agency for Research on Cancer Scientific Publications, 1980.
12. Hajek, J. and Sidak, Z. *Theory of Rank Tests*, Academic Press, 1967.