AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 115, No 5

Printed in U.S.A.

SAMPLE SIZES FOR INDIVIDUALLY MATCHED CASE-CONTROL STUDIES

A GROUP SEQUENTIAL APPROACH

BERNARD S PASTERNACK AND ROY E SHORE

Pasternack, B. S. (New York University Medical Center, New York, NY 10016), and R. E. Shore. Sample sizes for individually matched case-control studies: a group sequential approach. *Am J Epidemiol* 1982;115:778-84.

This paper proposes the use of group sequential methods to calculate sample sizes for individually matched case-control study designs. A table is presented in which the average sample size required for a group sequential (i.e., multistage) matched pair design is compared to that of the conventional matched pair fixed sample size plan for the usual constant relative risk situation. The table shows that group sequential designs are in general more efficient than fixed sample size plans. Computer simulations showed that group sequential methods yield the appropriate type I and type II error rates not only for matching on a one-to-one basis, but also more generally with multiple matched controls per case. Further simulation studies indicated that there may be only a small loss of power when the matching variable(s) is associated with the probability of exposure but not with the disease. This is shown for both the multistage and fixed sample tests.

biometry; epidemiologic methods; risk; retrospective studies; statistics

O'Neill and Anello (1) have proposed the use of a sequential approach to the design of *matched pair* case-control studies (the Wald sequential probability ratio test for comparing two binomial populations) as an alternative to fixed sample size plans when information on case-control pairs can be acquired sequentially in time. They also indicated that group sequential plans might be of relevance in the design of such studies. In a subsequent paper (2), we proposed the use of group sequential methods (based upon the theory of repeated significance tests) for *unmatched* cohort and case-control studies, and later demonstrated that group sequential (i.e. "multistage") designs are more efficient than fixed sample size plans for such studies (3). The recent results obtained by Pike et al. (4) suggest that case-control studies with individual matching are to be recommended for routine use by epidemiologists, as the reduction in confounding bias produced by such matching may be considerable, whereas there is only a small loss in efficiency compared with unmatched case-control designs when the matching variables are not associated with both the disease variable and the exposure variable under consideration; see also Miettinen (5).

In this paper, we specifically compare sample size requirements for *matched*

Received for publication July 6, 1981, and in final form November 6, 1981

Abbreviation: SPRT, sequential probability ratio test.

From the Dept of Environmental Medicine, New York University Medical Center, 550 First Avenue, New York, NY 10016 (Reprint requests to Dr Pasternack.)

Supported in part by Center Program Grants ES 00260 from the National Institute of Environmental Health Sciences, and CA 13343 from the National Cancer Institute

pair (all-or-none response) group sequential designs with traditional *matched pair* fixed sample size plans. We assume, as did O'Neill and Anello (1), that the relative risk (R) is constant and independent of matching variables and/or other covariables. In addition, we indicate that the group sequential approach is applicable to the more general case of multiple matched controls per case. For further discussion of the assumptions underlying use of the sequential (or group sequential) approach and the possible advantages and disadvantages of this method for individually matched case-control studies, see O'Neill and Anello (1).

Sample Size Requirements

The fixed sample size approach

When cases and controls are individually matched the distribution of all pairs can be described schematically as follows:

	Controls	
Cases	With exposure	Without exposure
With exposure	a	b
Without exposure	c	d

It is well known that both the Mantel-Haenszel (6) and maximum likelihood (7) estimates of R (which under the rare disease assumption is approximately by ρ, the odds ratio) are given by $\hat{\rho} = b/c$, and that an appropriate large sample test of statistical significance ($\rho = 1$ vs. $\rho \neq 1$) is the *McNemar* (8) one-degree-of-freedom χ^2 test (corrected for continuity), where

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}.$$

This same test statistic can be obtained as a special case of the Mantel-Haenszel summary chi-square procedure (6) with the number of strata equal to the number of matched pairs. The effective sample size for estimation and testing of ρ is thus seen to be the number of pairs in which one member has, and the other does not have, the factor (exposure) under study; i.e., $n = b + c$ = the number of exposure discordant pairs.

Following O'Neill and Anello (1), we define π_i (for the ith exposure discordant pair) as the probability that the case is exposed (P_{1i}) and the control is not ($1 - P_{2i}$), i.e.,

$$\pi_i = \frac{P_{1i}(1 - P_{2i})}{P_{1i}(1 - P_{2i}) + (1 - P_{1i})P_{2i}}. \quad (1)$$

The odds ratio (relative risk) is thus $\rho = \pi_i/(1 - \pi_i)$, assumed constant for all pairs. Since $\pi_i = \pi = \rho/(1 + \rho)$, the McNemar test of $\rho = 1$ vs. $\rho \neq 1$ is equivalent to testing $\pi = \frac{1}{2}$ vs. $\pi \neq \frac{1}{2}$. The square root of the McNemar statistic thus provides a large sample normal approximation for testing a binomial distribution in b based on n observations with probability $\pi = \frac{1}{2}$ vs. $\pi \neq \frac{1}{2}$. The number of exposure discordant pairs for the fixed sample size plan required for specified values of α (type I error rate), β (type II error rate) and ρ (for a two-sided test) can be obtained by application of a normal approximation to the binomial distribution using the arc sine transformation (9), in which case

$$n = \left\{\frac{z_\beta + z_{\alpha/2}}{2\arcsin\sqrt{\pi} - 2\arcsin\sqrt{\tfrac{1}{2}}}\right\}^2. \quad (2)$$

where z_β and $z_{\alpha/2}$ denote upper percentiles of the standard normal distribution.

The group sequential approach

Pocock (10) has described a group sequential procedure, based upon repeated significance testing, which can be easily

adapted to the matched pair case-control study. When α-level significance tests are repeated periodically on accumulating data, the overall significance level can be greatly increased. One way to overcome this problem is to choose the maximum number of tests to be performed in advance (N) and then to apply a lower, more stringent nominal significance level (α') for each repeated test, so that the overall significance level (α) is maintained. Table 1 shows the values of the nominal significance level for the case of a normal response variable. The chief advantage of the group sequential approach in the conduct of case-control studies is that fewer observations are needed on the average than a fixed sample size plan to obtain a statistically significant result when the "true" relative risk differs from unity, other specifications being the same, i.e., α and β (type II error rate). (This advantage becomes a disadvantage when the null-hypothesis is true, since the expected sample size is greater for sequential methods in this circumstance.)

The group sequential procedure we will describe is based on the sequential use of the *positive* square root of the McNemar one-degree-of-freedom χ^2 statistic (χ, without a correction for continuity—in accord with the known inappropriateness of the continuity correction when multiple test statistics are combined) for the matched pair case-control study,

$$\chi = \frac{|b - c|}{\sqrt{b + c}}.$$

The method calls for the computation of χ (which is asymptotically distributed as a standardized normal deviate, z) at each stage of the group sequential accumulation of exposure discordant pairs. The per-test significance level α' (two-sided) and its corresponding standardized normal deviate $z_{\alpha'}$ can be obtained from table 1 once N has been chosen. At each sequential test, the computed χ is compared to this $z_{\alpha'}$ and significance is achieved when $\chi > z_{\alpha'}$; the maximum number of tests is N whether or not significance is attained.

For illustrative purposes and to fix these ideas, consider the example given by Schlesselman (11) and used by Pasternack and Shore (3). Namely, an investigator is interested in studying whether or not there is an increased risk of giving birth to a child with a congenital heart defect among mothers who have oral contraceptive exposure in the period three months before or after conception. Assume that a group sequential matched pair case-control study is to be conducted with a preselected maximum of N stages. If the investigator specifies the value of R which he regards as important to detect, α and β, then the required increment in sample size (n = the number of exposure discordant pairs) at each stage of sequential testing is given by

$$n = \frac{\Delta^2}{(2 \arcsin\sqrt{\pi} - 2 \arcsin\sqrt{1/2})^2} \quad (3)$$

where $\pi = R/(1 + R)$ and Δ is obtained from table 2 in Pasternack and Shore (3). (Note that when $N = 1$, equation 3 reduces to equation 2, since in that case $\Delta = z_\beta + z_{\alpha/2}$). The *maximum* sample size (exposure discordant pairs) is: $n' = nN$ and the *average*

TABLE 1

*The per-test significance level (α') and its corresponding standardized normal deviate z for use in normal group-sequential testing for various values of number of groups (stages) N and overall significance level (α, two-sided)**

N	α = 0.05		α = 0.01	
	α'	z	α'	z
1†	0.0500	1.960	0.0100	2.576
2	0.0294	2.178	0.0056	2.772
3	0.0221	2.289	0.0041	2.873
4	0.0182	2.361	0.0033	2.939
5	0.0158	2.413	0.0028	2.986

* Extracted from table 1 in Pocock (10).

† $N = 1$ is included to enable comparison with the critical value for a fixed sample size design.

sample size is $\bar{n}' = \bar{A}n$, where the appropriate value of $\bar{A}$, the average number of tests, is taken from table 2 in Pasternack and Shore (3).

Table 2 contains the sample sizes (n = number of discordant pairs) required for fixed sample size plans ($N = 1$) as well as for group sequential designs ($N = 2, 3, 4, 5$) obtained by use of equation 3. Average sample sizes ($\bar{n}'$) are also included. The sample sizes apply to matched pair case-control study designs using a two-sided test. As an example of the use of table 2, if (in the example described above) the investigator wished to detect $R = 2$ with $N = 2$, $\alpha = 0.05$ and $\beta = 0.10$, he would perform a sequential test after each accumulation of 50 exposure discordant pairs until $\chi > 2.178$ (from table 1) or until two nonsignificant tests had been performed. On the average, only 71 exposure discordant pairs would need to be obtained, as compared with 91 for the fixed sample size plan ($N = 1$, table 2).

The "expected" total number of pairs n_t required to yield the desired number of exposure discordant pairs at each stage (n) can be approximated by dividing n by ψ as suggested by O'Neill and Anello (1), where

$$\psi = P_1 (1 - P_2) + P_2 (1 - P_1)$$

is the probability of obtaining a discordant pair and P_1 and P_2 are presumed to be "anticipated" or "average" probabilities of exposure over the matched pairs. Note that if a value of P_2 (the probability of a control being exposed) is selected, P_1 is automatically determined by the choice of R, since

TABLE 2

*Sample sizes for multistage and fixed sample matched pair case-control designs for selected values of relative risk (R) and probability of exposure (P_2) in the control group. $\alpha = 0.05$, $\beta = 0.10$ and two-sided significance tests are assumed**

R	N	n	$\bar{n}'$	$P_2 = 0.1$		$P_2 = 0.2$		$P_2 = 0.3$		$P_2 = 0.5$	
				n_t	$\bar{n}_t'$	n_t	$\bar{n}_t'$	n_t	$\bar{n}_t'$	n_t	$\bar{n}_t'$
1.7	1	153	153	673	673	403	403	326	326	306	306
	2	84	118	370	522	222	313	179	253	168	237
	3	59	110	258	485	155	291	125	235	117	220
	4	45	107	199	470	119	281	96	228	90	213
	5	37	105	162	461	97	276	79	223	74	209
2.0	1	91	91	371	371	228	228	188	188	182	182
	2	50	71	204	287	125	176	103	146	100	141
	3	35	66	142	267	87	164	72	135	70	131
	4	27	64	110	259	67	159	56	131	54	127
	5	22	62	89	254	55	156	45	129	44	125
2.5	1	54	54	196	196	124	124	106	106	107	107
	2	29	42	108	152	68	96	58	82	59	83
	3	21	39	75	141	48	90	41	76	41	77
	4	16	37	58	137	37	87	31	74	32	75
	5	13	37	47	134	30	85	25	72	26	73
3.0	1	38	38	128	128	84	84	73	73	77	77
	2	21	30	70	99	46	65	40	57	42	59
	3	15	28	49	92	32	60	28	53	29	55
	4	11	27	38	89	25	59	22	51	23	54
	5	9	26	31	88	20	57	18	50	18	53

* N = number of group sequential tests (stages) planned ($N = 1$ is for a fixed sample plan); n = number of discordant case-control pairs per stage; $\bar{n}'$ = average number of discordant pairs required per study; n_t = average total number of pairs per stage; $\bar{n}_t'$ = average total number of pairs per study.

$$R = \frac{P_1 (1 - P_2)}{(1 - P_1)P_2}$$

and *vice versa*.

Table 2 contains values for n_t and $\bar{n}'_t = \bar{n}'/\psi$ for a range of values of P_2 and R. For example, if we assume that $N = 3$, $P_2 = 0.10$ and $R = 2$, then we see from table 2 that the average number of pairs required *per test*, $n_t = 142$, and the average *total* number of pairs required, $\bar{n}'_t = 267$.

Discussion

Examination of table 2 shows that the greatest reduction in the average sample size is achieved by using a two-stage design rather than a one-stage (i.e., fixed sample) plan; there is little to be gained in using a design with more than four or five stages. It is important to note that basing the group sequential procedure (as well as the fixed sample plan) on the required number of discordant pairs, n, insures that the desired power of the McNemar test to detect a given value of $R \neq 1$ obtains regardless of the presumed value of P_2; whereas if the assumed value of P_2 is widely in error, the resulting calculation of n_t can lead to a sizeable underestimate or overestimate of the actual required number of discordant pairs.

A limitation of both the sequential probability ratio test (SPRT) and multistage sampling plans is that their stopping strategy is based on only a single variable, whereas case-control studies often examine multiple exposures or risk factors, and cohort studies often examine multiple diseases. Thus the SPRT or multistage designs will find use primarily when there is one main hypothesis of interest.

Schlesselman (12) has indicated two further limitations of sequential sampling plans in case-control studies: 1) the approach uses an analysis that is overly simplistic insofar as it emphasizes a test of significance; and 2) the choice of too large a value for R as the alternative of interest may lead to very early rejection of the null hypothesis. Thus, one might discontinue a study with some assurance that a difference exists but be unable to measure it with any precision. Armitage (13) has proposed a particular stopping rule to be applied to such problems of sequential estimation: continue sampling until the precision of an estimate reaches a prescribed level noting that the usual formula for measuring precision (e.g., the standard error or confidence limits) indicates (approximately) when the required precision has been reached, *even though the sampling is sequential;* see Schlesselman (12) for the appropriate fixed sample formulas that measure the precision of relative risk estimates from matched and unmatched case-control studies.

Nevertheless, the usual methods of estimation of relative risk applied to fixed sample size case-control studies are not *strictly* appropriate when termination is based on sequential methods. Many epidemiologists believe that estimation is more important than significance testing so this can be an important limitation of the proposed group sequential procedure as well as the SPRT. If estimation of R, rather than significance testing, is of primary concern, fixed sample size methods should be considered as an alternative to the use of sequential sampling plans.

It should be borne in mind that although matched follow up studies are infrequently conducted, hypothesis testing is essentially the same as for matched case-control studies. Colton and McPherson (14), for example, have obtained optimal two-stage plans for the comparison of two binomial proportions in paired samples with the constraint that the number of untied pairs per stage are equal. Their optimal designs were based on the use of the binomial distribution rather than a large sample approximation. Although their applications were given in a clinical trial setting, comparisons with matched pair case-control de-

signs can be made for equivalent values of α, β, and π. For the example we have considered, where $N = 2$, $\pi = 2/3$, $\alpha = 0.05$ and $\beta = 0.10$ (two-sided test), the optimum sample size (discordant pairs) per stage and the overall average sample size given by Colton and McPherson (14) were 48 and 68. These values are very close to those we obtained for the two-stage group sequential test using the square root of the McNemar χ^2 statistic; *viz.*, 50 and 71, respectively. Thus, the use of group-sequential matched pair designs can be considered for cohort studies, as well as clinical trials, when feasible.

Multistage analyses appear to permit more flexibility than currently available sequential analyses. For example, we have performed computer simulations that show multistage tests can be applied when there are multiple matched controls per case using the adaptation of the Mantel-Haenszel procedure reported by Pike and Morrow (15). Although we have not tested it by computer simulation or other means, in principle one would expect that the recent statistical methods for matched case-control data which permit multivariate analyses of covariates (16, 17) could be used in multistage fashion as well. In a similar vein, a previous paper (3) showed that the Mantel-Haenszel method with stratification for covariates could safely be used for unmatched data.

Computer simulations were performed for multistage tests and fixed sample tests of matched data to evaluate whether their type I and type II error levels were satisfactory. The McNemar test (without a continuity correction) performed appropriately in both the multistage and fixed sample mode (e.g., mean type I errors at a nominal $\alpha = 0.05$ were 0.0499 and 0.0484, respectively). However, the McNemar fixed sample test with continuity correction was consistently conservative (e.g., a mean of 0.0364 at a nominal $\alpha = 0.05$).

Simulations were also performed for the case in which one tests every n_t total pairs (rather than every n discordant pairs), in which the number of additional discordant pairs per stage will vary. Again, both the multistage and fixed sample tests displayed appropriate α and β levels.

We then generalized to the situation with multiple matched controls per case. It was found that the simple formula for determining sample size with multiple unmatched controls (formula 2 in reference 3) produced satisfactory results for multiple matched controls as well. Tests were performed using the adaptation of the Mantel-Haenszel test for multiple matched controls (15). Both the multistage and fixed sample tests yielded satisfactory α and β levels.

The simulations described above assumed that the exposure probability for controls remained constant from stratum (pair) to stratum; an assumption also made by Breslow (18) and Pike et al. (4) in their numerical evaluations of asymptotic means and variances of odds ratio estimators.

Finally, we simulated the case-control study situation in which the covariate(s) on which one is matching is in fact associated with the probability of exposure, but is unrelated to disease. This situation has been discussed by several authors recently (19–21), and has been termed "overmatching" by Miettinen (5). It should be noted that the simulations performed also correspond to the situation in which the matching variable is associated with disease as well as exposure conditional on disease. This would not be thought of as overmatching since the inclusion of the matching variable is required for the purpose of study validity (lack of bias); discussion of study efficiency (power) in this circumstance, however, is inappropriate since matching is motivated by the need to avoid confounding (5). We simulated two situations, one in which P_2 (the probability of exposure

for the control member of a pair) varied randomly over the range 0.075−0.225 with a small sample size, and the other with a range 0.15−0.45 and a relatively large sample size. We found that in both situations there was a small loss of power for both the multistage and fixed sample tests (i.e., for a nominal power of $1 - \beta = 0.75$, the mean obtained values without the continuity correction were 0.714 and 0.730, respectively; and 0.671 for the fixed sample test with the correction for continuity).

Additional material containing tables and a description of the computer simulations performed is available from the authors by written request.

References

1 O'Neill RT, Anello C. Case-control studies. a sequential approach Am J Epidemiol 1978;108: 415-24.
2. Pasternack BS, Shore RE. Group sequential methods for cohort and case-control studies J Chronic Dis 1980,33:365-73
3 Pasternack BS, Shore RE Sample sizes for group sequential cohort and case-control study designs Am J Epidemiol 1981;113·182-91.
4. Pike M, Hill A, Smith PG. Bias and efficiency in logistic analyses of stratified case-control studies. Intl J Epidemiol 1980,9.89-95
5 Miettinen O Matching and design efficiency in retrospective studies Am J Epidemiol 1970, 91:11-18
6. Mantel N, Haenszel W Statistical aspects of the analysis of data from retrospective studies of disease JNCI 1959,22:719-48
7 Cornfield J, Haenszel W. Some aspects of retrospective studies J Chronic Dis 1960;11:523-34.
8. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153-7.
9. Brownlee KA. Statistical theory and methodology in science and engineering. New York: John Wiley & Sons, 1960.
10 Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64.191-9.
11 Schlesselman JJ. Sample size requirements in cohort and case-control studies of disease. Am J Epidemiol 1974,99:381-4.
12 Schlesselman JJ. Case-control studies design, conduct, analysis. New York: Oxford University Press, 1982.
13. Armitage P. Sequential medical trials. 2nd ed New York: John Wiley & Sons, 1975.
14 Colton T, McPherson K Two-stage plans compared with fixed sample-size and Wald SPRT plans J Am Statis Assoc 1976;71 80-6.
15. Pike M, Morrow R. Statistical analysis of patient-control studies in epidemiology: factor under investigation an all-or-none variable Br J Soc Prev Med 1970;24:42-3
16 Breslow N, Day N, Halvorsen K, et al Estimation of multiple relative risk functions in matched case-control studies Am J Epidemiol 1978,108 299-307
17. Holford T, White C, Kelsey J Multivariate analysis for matched case-control studies. Am J Epidemiol 1978;107:245-56.
18 Breslow N. Odds ratio estimators when the data are sparse. Biometrika 1981;68:73-84
19 Kupper L, Karon J, Kleinbaum D, et al Matching in epidemiologic studies: validity and efficiency considerations. Biometrics 1981;37: 271-91.
20 Day N, Byar D, Green S Overadjustment in case-control studies Am J Epidemiol 1980; 122:696-706
21 Whittemore AW Efficiency of synthetic retrospective studies Biom J 1981;23.73-8.