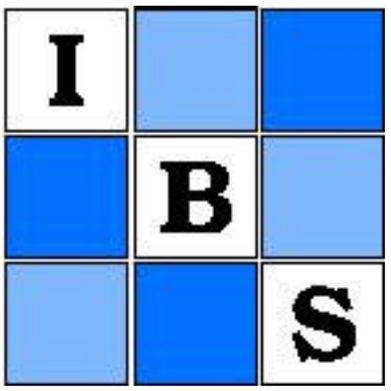

BIOMETRICS 42, 919–926
December 1986

Sample Size for Individually Matched Case–Control Studies

R. A. Parker

Division of Chronic Disease Control, Center for Environmental Health,
Centers for Disease Control, Atlanta, Georgia 30333, U.S.A.

and

D. J. Bregman

National Institute of Occupational Safety and Health,
Cincinnati, Ohio 45226, U.S.A.

SUMMARY

The standard formulas used to calculate sample size for an individually matched case–control study assume a constant probability of exposure throughout the pool of possible controls. We propose new formulas that allow for heterogeneity in the probability of exposure among controls in different matched sets. Since matching factors are suspected of being confounders, they are expected to divide the total population into subgroups with different proportions exposed. Thus, the assumption of homogeneity of exposure among controls, made by the currently used formulas, is inconsistent with the assumptions used to design a matched study. The proposed formulas avoid this inconsistency. We present an example to illustrate how heterogeneity can affect the required sample size.

1. Introduction

In a case–control study, controls are often selected to match cases on one or more characteristics. For example, a hospital control may be selected as the first admission after the case with the same sex and age as the case. Both individual matching (one or more controls matched to each specific case) and frequency matching (groups of controls matched to groups of cases) are used. There are standard formulas available to calculate approximate sample sizes (number of cases) for both individually matched and frequency matched studies (e.g., Schlesselman, 1982, Chap. 6).

The sample size for a 1:1 matched study is the number of discordant pairs required (calculated from the desired power for a specified odds ratio and a given level of statistical significance) divided by the estimated probability that a randomly selected pair is discordant (calculated from the average prevalence of exposure in the general population and the specified odds ratio). This approach, recognized as an approximation, does not take into account variability in the probability of discordance between matched sets.

The frequency matched formula was originally developed for prospective stratified studies (Gail, 1973) and extended by Muñoz and Rosner (1984). As these formulas calculate sample size for a stratified analysis, they are not intended to provide the sample size for an individually matched study. However, these formulas explicitly consider the heterogeneity of exposure prevalence among different strata of the controls. In addition to specification

Address for correspondence: R. A. Parker, CEH/CD/OD Building 5A (Chamblee), Centers for Disease Control, Atlanta, Georgia 30333, U.S.A.

Key words: Case–control; Matching; Sample size.

of odds ratio, significance level, and desired power, the formulas require an estimate of the size of each stratum and prevalence of exposure within each stratum.

For either design, if there are $M > 1$ controls for each case, then the approximate number of cases needed for the study is the fraction $(M + 1)/(2M)$ of the sample size for a study with a single control per case. This is the reciprocal of the asymptotic relative efficiency of a 1:M to 1:1 study (Ury, 1975). In addition, an extension that explicitly allows for the failure to recruit exactly M controls for each case is available for individually matched studies (Walter, 1980).

Miettinen (1970) discusses when individual matching is appropriate for a case–control study. He concludes that matching should be considered only if the matching factor is expected to be a true confounder, i.e., if the matching factor is expected to be related both to exposure and to disease. Otherwise, the matching is irrelevant to the validity of the study and may reduce the efficiency of the study. By extension, these conclusions apply to frequency matched studies as well. In order for a matching factor to be a confounder, the subpopulations identified by the matching factor must have different proportions exposed. Therefore, the calculation of sample size for an individually matched study should explicitly consider the heterogeneity of the subpopulations of controls.

In this paper we develop sample size formulas for individually matched studies that explicitly include the heterogeneity in the proportion exposed between matched groups. Section 2 presents the notation and assumptions underlying our approach, while sample size formulas for 1:1 and 1:M individually matched studies are developed in Section 3. The example in Section 4 illustrates how heterogeneity can affect the estimated sample size requirements for a study.

2. Notation and Assumptions

To calculate the number of cases needed for an individually matched case–control study (N), the investigator must specify (i) power of the study $(1 - \beta)$ for (ii) a specified odds ratio (ψ) at (iii) a specified Type I error (α), with (iv) a specified number of controls (M) per case. In addition, since the matching factor (which may consist of one or more characteristics) is a suspected confounder, the matching factor should identify subpopulations that differ in both prevalence of exposure (π) and incidence rate (P) among the unexposed.

Let Z_x be the standard normal variate exceeded with probability x. For a one-sided test, Z_α is used to test for statistical significance, while for a two-sided test, $Z_{\alpha/2}$ would be used instead. Formulas are presented for a one-sided test. We make the usual assumption that the odds ratio ψ is constant for all matched sets and assume, without loss of generality, that $\psi > 1$. We also make the usual assumption that the odds ratio is approximately the same as the relative risk.

Since the matching factor is expected to be a confounder, the factor is expected to be associated with both the incidence rate (P) among the unexposed and the prevalence of exposure (π) in different subpopulations (Breslow and Day, 1980, p. 95). However, we assume that, when subpopulations are stratified by prevalence of exposure, the average incidence rate among the unexposed is constant across all strata. Note that this does not imply that incidence is constant across strata when strata are formed based on the matching factors. Let ρ denote this average incidence rate. Thus, the formulas derived below do not take account of any joint dependence of π and P.

Let $f_p(\pi)$ be the density function of π over the universe of subpopulations identified by the matching factor. For example, if the matching factor is medical practice, π might be .7 in one practice and .5 in another practice. The density $f_p(\cdot)$ may be either continuous or discrete, depending on whether the matching factor identifies a (theoretically) infinite

number of subpopulations (e.g., using neighborhood as the matching factor) or a (small) finite number of subpopulations (e.g., using sex as the matching factor). In the former case, individual matching is appropriate, while in the latter case either frequency matching or individual matching could be used. Let $\bar{\pi}_p$ equal the expectation of π under $f_p(\pi)$, i.e., the proportion of the total population exposed. Similarly, $f_c(\pi)$ is the density function of π over the population of cases. We assume that the sampled cases are a random sample of all cases in the population. Note that $f_c(\cdot) \equiv f_p(\cdot)$ if and only if $\psi = 1$. Since controls are selected from the subpopulations providing cases, the probability of selecting a subpopulation with proportion exposed between π and $\pi + \Delta\pi$ from which to sample a control is given by $F_c(\pi + \Delta\pi) - F_c(\pi)$, where $F_c(\cdot)$ is the cumulative distribution function corresponding to $f_c(\cdot)$.

Let $P_{m,M+1}(\pi)$ be the probability that exactly m individuals in a set of one case and M controls are exposed, when the case and matched controls come from a subpopulation with prevalence of exposure π, while $P_{m,M+1}$ is the expectation of $P_{m,M+1}(\pi)$ over the density function $f_c(\pi)$. For example, $P_{1,2}$ is the probability that a matched pair is discordant on exposure, given that the pair is sampled with probability density $f_c(\cdot)$ from the universe of subpopulations.

3. General Formulas for Sample Size

3.1 *Individually Matched Studies with* 1:1 *Matching*

In our notation, the standard formula for sample size (Schlesselman, 1982, eqs. 6-20 and 6-23) is

$$N = \{[Z_\alpha(1 + \psi) + 2Z_\beta(\psi)^{1/2}]/(\psi - 1)\}^2/[k(\psi + 1)\bar{\pi}_p(1 - \bar{\pi}_p)] \tag{3.1}$$

where $k = 1/[1 + (\psi - 1)\bar{\pi}_p]$.

For a subpopulation with proportion exposed π, cases arise in the unexposed segment (proportion $1 - \pi$ of the subpopulation) with incidence rate ρ and in the exposed segment (proportion π of the subpopulation) with incidence rate $\psi\rho$, using the odds ratio as an estimate of the relative risk. Thus, the number of cases (n_c) expected to come from subpopulations with proportion exposed π is proportional to

$$n_c(\pi) \propto [(1 - \pi)\rho + \pi\psi\rho]f_p(\pi) = \rho[1 + (\psi - 1)\pi]f_p(\pi)$$

and, thus, the density function of exposure among subpopulations providing cases is

$$f_c(\pi) = k[1 + (\psi - 1)\pi]f_p(\pi), \tag{3.2}$$

where k was defined above.

From standard results (e.g., Breslow and Day, 1980, eq. 5-16), the number of discordant pairs needed for a one-tailed test is given by

$$N_d = \{[Z_\alpha(1 + \psi) + 2Z_\beta(\psi)^{1/2}]/(\psi - 1)\}^2$$

independent of the distribution of π.

Given a subpopulation with proportion exposed π, a randomly selected pair is discordant if either the case is exposed (probability $\psi\pi/[1 + (\psi - 1)\pi]$) and the control is unexposed (probability $1 - \pi$) or vice versa. Thus, the probability that a matched pair is discordant for a given value of π is given by

$$P_{1,2}(\pi) = (\psi + 1)\pi(1 - \pi)/[1 + (\psi - 1)\pi].$$

Since the controls for matched sets are sampled from subpopulations according to density

function (3.2), the probability of a randomly sampled pair being discordant is

$$P_{1,2} = k(\psi + 1) \int_0^1 \pi(1 - \pi) f_p(\pi)\, d\pi,$$

which can be rewritten as

$$P_{1,2} = k(\psi + 1) I_{1,2}$$

where

$$I_{a,b} = \int_0^1 \pi^a (1 - \pi)^{b-a} f_p(\pi)\, d\pi.$$

Thus, the number of pairs needed for a study is

$$N = \{[Z_\alpha(1 + \psi) + 2Z_\beta(\psi)^{1/2}]/(\psi - 1)\}^2/[k(\psi + 1)I_{1,2}]. \tag{3.3}$$

If $f_p \propto$ Beta(p, q), then $I_{1,2} = B(p + 1, q + 1)/B(p, q)$ can be substituted into (3.3).

Equations (3.1) and (3.3) differ only in the denominator, where (3.1) has the term $\bar{\pi}_p(1 - \bar{\pi}_p)$, while (3.3) has the term $I_{1,2}$. Since $I_{1,2} \leq \bar{\pi}_p(1 - \bar{\pi}_p)$ (proof provided in the Appendix), the standard formula (3.1) normally underestimates the required sample size. When the proportion exposed varies in subpopulations, some cases will come from extreme subpopulations, that is, subpopulations with exposure prevalence near 0 or 1. However, matched sets drawn from extreme subpopulations are less likely to be discordant than are sets drawn from subpopulations with the average exposure prevalence. Thus, more pairs must be sampled to allow for heterogeneity.

3.2 *Individually Matched Studies with* 1:*M Matching*

Assume that M controls will be selected for each case. The standard formula (3.1) is multiplied by $(M + 1)/(2M)$, so the required number of cases is

$$N = (M + 1)\{[Z_\alpha(1 + \psi) + 2Z_\beta(\psi)^{1/2}]/(\psi - 1)\}^2/[2Mk(\psi + 1)\bar{\pi}_p(1 - \bar{\pi}_p)]. \tag{3.4}$$

However, it would be more accurate to allow for the actual number of discordant sets expected to occur. For a set selected from a subpopulation with proportion π exposed, the probability that exactly m individuals in the set are exposed is

$$P_{m,M+1}(\pi) = \binom{M}{m} \frac{\pi^m (1 - \pi)^{M-m+1}}{1 + (\psi - 1)\pi} \frac{M + (\psi - 1)m + 1}{M - m + 1}. \tag{3.5}$$

Combining (3.2) with (3.5) yields the probability that exactly m individuals are exposed for a randomly selected subpopulation providing a case:

$$P_{m,M+1} = \binom{M}{m} k I_{m,M+1} \frac{M + (\psi - 1)m + 1}{M - m + 1}. \tag{3.6}$$

The expectation and variance of the number of discordant sets with the case exposed can be calculated from formula 5-16 of Breslow and Day (1980), both for the given odds ratio (ψ) and for an odds ratio of 1. Since these numbers are conditional on the particular number of discordant sets observed, they involve f_p only through equation (3.6). In an

obvious notation, with $K_m = M + (\psi - 1)m + 1$, these values are

$$E_\psi = \psi kN \sum_{m=1}^{M} \binom{M}{m-1} I_{m,M+1};$$

$$V_\psi = \psi kN \sum_{m=1}^{M} \binom{M}{m} m K_m^{-1} I_{m,M+1};$$

$$E_1 = k(M+1)^{-1} N \sum_{m=1}^{M} \binom{M}{m-1} K_m I_{m,M+1};$$

$$V_1 = k(M+1)^{-2} N \sum_{m=1}^{M} \binom{M}{m} m K_m I_{m,M+1}.$$

Since the required sample size satisfies

$$Z_\alpha = [E_\psi - E_1 - Z_\beta (V_\psi)^{1/2}]/(V_1)^{1/2},$$

the number of cases required is given by

$$N = \left\{ Z_\alpha \left[\sum_{m=1}^{M} \binom{M}{m} m K_m I_{m,M+1} \right]^{1/2} + Z_\beta (M+1) \left[\psi \sum_{m=1}^{M} \binom{M}{m} m K_m^{-1} I_{m,M+1} \right]^{1/2} \right\}^2 \times \left\{ k \left[(\psi - 1) \sum_{m=1}^{M} \binom{M}{m} m I_{m,M+1} \right]^2 \right\}^{-1}. \tag{3.7}$$

Again, if $f_p \propto \text{Beta}(p, q)$, then $I_{m,M+1} = B(p+m, q+M-m+1)/B(p, q)$ can be substituted into (3.7).

Formula (3.7) can also be used to calculate the sample size for an individually matched study with $M > 1$ controls per case, even when the controls are assumed to be homogeneous; this would avoid the asymptotic approximation involved in (3.4).

4. Example

In August 1984, the National Academy of Sciences Institute of Medicine discussed the design of the U.S. Public Health Service Study on the Association of Reye's Syndrome and Medication. Reye's syndrome is a rare disease of unknown etiology, typically seen in children apparently recovering from a mild respiratory infection, influenza, chicken pox, or other viral disease. Reye's syndrome normally occurs within a week of the original infection. Previous studies have raised the suspicion that Reye's syndrome is associated with a nonprescription medication used to treat the antecedent illness. These results have been published both in professional journals and in the news media. As a result, it is expected that the use of this nonprescription medication would differ between pediatric practices and between communities within the United States.

All sample sizes are based on a one-sided test with $\alpha = .05$, $\beta = .10$, $\psi = 4.0$, an expected proportion of exposure ($\bar{\pi}_p$) in the general population of .50, and 1:2 matching. Because this is a national study, individual matching on geographic area was required to identify controls for personal interviews. Based on the standard formula (3.4), the estimated sample size (number of cases required) is 29.7. If it is assumed that exposure was constant throughout the United States, the exact formula (3.7) gives an estimated sample size of 30.2, a slight increase.

During the planning meeting, several strategies to obtain controls were considered. One method considered would draw controls from the same pediatric practice as the case. An alternate approach would select controls by random-digit dialing within the same area code and telephone exchange as the case.

Medication is likely to be used differently in different pediatric practices. Moreover, it was considered likely that many practices would have a general policy concerning the use of the nonprescription medication in question. Thus, in some practices the medication would virtually never be recommended while in others it would virtually always be recommended. However, since not all patients consistently follow medical advice, we estimated that 5% of the individuals in a practice would ignore the advice given. Thus, a practice that universally recommended the medication would have a true exposure rate of 95%, while a practice that never recommended the medication would have an exposure rate of 5%. Several possible discrete distributions were considered as models for medication exposure (Table 1). These distributions lead to different sample sizes since the underlying distribution of proportion of controls exposed is not concentrated at $\pi = .5$. The most homogeneous of these discrete assumptions leads to about a 75% increase over the 29.7 cases estimated using the standard formula for a matched study, while the most extreme assumption leads to a sample size more than 5 times as large. In this latter case, we are in fact matching on a near-perfect predictor of exposure among the control population.

Since these are discrete distributions, it is possible to calculate the sample size for a stratified study [e.g., Schlesselman, 1982, eq. 6-19 with adjustment factor $(M + 1)/(2M)$]. The sample size calculated from this formula, shown in Table 1, is much closer to the

Table 1
Sample size required for various distributions of population exposure

	Number of cases
Standard matched pair formula [eq. (3.4)]	29.72
Proposed formula [eq. (3.7)]: homogeneous exposure	30.19

Heterogeneity: Discrete distributions

Prop. exposed in subpopulation	Prop. of population in subpopulation	Stratified study size[a]	Matched study size [eq. (3.7)]
.25	.643	62.68	54.02
.95	.357		
.05	.111		
.25	.500	74.24	63.34
.95	.389		
.05	.500	209.19	158.89
.95	.500		

Heterogeneity: Beta distribution for exposure

Parameters of beta distribution	Proportion of total population with exposure between .4 and .6	Matched study size [eq. (3.7)]
2.051	.30	37.55
5.816	.50	32.79
13.404	.70	31.32
33.387	.90	30.64

[a] Standard formulas are given by Schlesselman [1982, equations (6.19) with adjustment (6.25)], and are for a stratified analysis.

sample size calculated from (3.7) than the sample size calculated from (3.4). The formula for a stratified study overestimated the true study size required in these examples, but this formula is not intended to estimate the sample size for a study with a matched analysis.

Since community controls are potentially drawn from a number of different pediatric practices, the proportion of controls exposed in any community would be a weighted average of several practices. Thus, it is expected that controls from different communities would be more homogeneous than controls drawn from individual practices. However, there would still be some heterogeneity between different communities because of differences in the amount of publicity on the suspected association between medication and disease. To illustrate how heterogeneity affects sample size, the proportion of controls exposed was modeled as a beta variable, with mean of .5 and 30%, 50%, 70%, and 90% of the control population having a probability of exposure between .4 and .6. As shown in Table 1, the number of cases has to be increased to 37.5 for the most heterogeneous assumption, about a 25% increase over the sample size estimated from the standard formula (3.4).

As mentioned in Section 3.1, allowing for heterogeneity increases the calculated sample size requirements because as subpopulations become more heterogeneous for exposure, a higher proportion of cases will come from subpopulations with a high exposure. Since the probability that a matched set will be discordant decreases as the population exposure becomes more extreme, more sets are needed to obtain the necessary number of discordant sets. This is most noticeable when the distribution of π is discrete with mass at .05 and .95 only. In this case, 77% of cases come from the population group with prevalence of .95. However, from (3.5) only 10.9% of these sets will be discordant, compared to the expected 75% of all sets if the prevalence was actually .50.

Since properly selected matching factors will identify subpopulations heterogeneous on proportion exposed, it follows that the formula used to calculate the sample size needed for an individually matched case–control study should take this phenomenon into account.

Acknowledgements

We wish to thank the referees of an earlier version for suggestions on improving the presentation of the material.

Résumé

Les formules usuelles pour calculer des tailles d'échantillons pour une étude cas–témoins avec appariement individuel supposent une probabilité constante d'exposition parmi l'ensemble des témoins possibles. Nous proposons de nouvelles formules permettant de tenir compte d'une hétérogénéité de cette probabilité parmi les différents ensembles appariés. Puisque les facteurs d'appariement sont suspectés être confondants, ils doivent séparer la population en sous-groupes de taux d'exposition différents. De ce fait, l'hypothèse d'homogénéité d'exposition parmi les témoins, implicite dans les formules usuelles, est inconsistante avec les hypothèses à la base des études appariées. Les formules proposées évitent cette inconsistance. Nous présentons un exemple afin d'illustrer l'effet de cette hétérogénéité sur la taille d'échantillon requise.

References

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research: Volume* I—*The Analysis of Case–Control Studies.* Lyon: International Agency for Research on Cancer.

Gail, M. (1973). The determination of sample sizes for tables involving several independent 2 × 2 tables. *Journal of Chronic Diseases* **26,** 669–673.

Miettinen, O. (1970). Matching and design efficiency in retrospective studies. *American Journal of Epidemiology* **91,** 111–118.

Muñoz, A. and Rosner, B. (1984). Power and sample size for a collection of 2 × 2 tables. *Biometrics* **40,** 995–1004.

Schlesselman, J. J. (1982). *Case-Control Studies.* New York: Oxford University Press.
Ury, H. K. (1975). Efficiency of case-control studies with multiple controls per case: Continuous or dichotomous data. *Biometrics* **31,** 643–649.
Walter, S. D. (1980). Matched case-control studies with a variable number of controls per case. *Applied Statistics* **29,** 172–179.

Received October 1984; *revised July* 1985 *and April and August* 1986.

Appendix

Proof that $I_{1,2} \leq \bar{\pi}_p(1 - \bar{\pi}_p)$

$$
\begin{aligned}
I_{1,2} &= \int_0^1 \pi(1 - \pi) f_p(\pi)\, d\pi \\
&= \int_0^1 (\pi - \pi^2) f_p(\pi)\, d\pi \\
&= E_p(\pi) - E_p(\pi^2) \\
&= E_p(\pi) - E_p^2(\pi) - E_p(\pi^2) + E_p^2(\pi) \\
&= E_p(\pi) - E_p^2(\pi) - [E_p(\pi^2) - E_p^2(\pi)] \\
&= E_p(\pi) - E_p^2(\pi) - \mathrm{var}_p(\pi)
\end{aligned}
$$

where E_p means expectation and var_p means variance with respect to density $f_p(\cdot)$. Since $E_p(\pi) = \bar{\pi}_p$ is fixed, $I_{1,2}$ is a maximum when $\mathrm{var}_p(\pi)$ is a minimum. The minimum, zero, occurs when $f_p(\pi)$ is a point mass at $\bar{\pi}_p$, in which case

$$I_{1,2} = \bar{\pi}_p(1 - \bar{\pi}_p).$$

Since sample size is proportional to $(I_{1,2})^{-1}$ or $[\bar{\pi}_p(1 - \bar{\pi}_p)]^{-1}$, the sample size calculated from the standard formula (3.1) normally underestimates the required sample size.