

# Design options in epidemiologic research

## An update

by Olli Miettinen, MD, PhD<sup>1</sup>

MIETTINEN O. Design options in epidemiologic research: An update. *Scand j work environ health* 8 (1982): suppl 1, 7—14.

I felt embarrassed about the prospect of giving a "lecture" to such a learned audience, especially an "honorary" lecture. I had to find something beyond the ordinary, and I entertained very seriously some topical problem areas in methodology, circumscribed and somewhat esoteric. At the same time, I continued to be very preoccupied with something more fundamental, the options in epidemiologic study design, which is an aspect of my current research interest. I had an urge to talk about this latter topic but felt insecure of my mastery of the issues. I finally felt confident enough to dare attempt an update — and indeed a revision — of my previous teachings in this area. Not only is the topic important in its own right, but its review gains added urgency from the fact that so many of you are familiar with my past approach in the International Advanced Course in Epidemiology that the Institute of Occupational Health in Helsinki has sponsored over the years.

### Preliminaries

Given the meaning of "design" in general, "study design" may be taken to mean a

vision of the end-product of a study on one hand and a scheme for carrying out a study on the other.

In epidemiologic research the concern is with the occurrence of events and states of illness and health in man. The magnitude of any parameter of such occurrence generally depends on various particulars of people's constitutions, behaviors, and/or environments. Therefore the quantification of any given occurrence parameter is, in general, a matter of relating its magnitude to the various determinants on which it depends. Such relationships, or *occurrence functions*, thus constitute the general formal object of epidemiologic research.

The function of concern in any given study may be abstract-general (divorced from time and place) or particularistic. Either way, the direct yield of the study is a particularistic function, one that is specific for the population experience that formed the base of the study, and thereby is the direct referent of its results. Such an *empirical* occurrence function — or qualitative information about it — can be thought of as the direct result of an epidemiologic study.

When the aforementioned general meaning of "study design" is applied to this formulation of the direct result of an epidemiologic study, the broadest aspects of epidemiologic study design may be said to include the stipulations of (i) the type of occurrence *function* (empirical) to be derived, (ii) the type (and size) of the population experience that is to form the empir-

<sup>1</sup> Departments of Epidemiology and Biostatistics, School of Public Health, Harvard University, Boston, Massachusetts, United States.

Reprint requests to: Dr O Miettinen, Departments of Epidemiology and Biostatistics, School of Public Health, Harvard University, 677 Huntington Avenue, Boston, MA 02115, USA.

ical base — and thereby the direct referent — of the function, and (iii) the type of *sampling* that is to be used in the ascertainment of the occurrence pattern in the base. Our concern, then, is with the options in each of these aspects of design.

## Occurrence functions

When the health state or outcome at issue is viewed as an all-or-none characteristic, the occurrence or outcome *parameter* may be taken as a rate of either *prevalence* or *incidence* as a matter of design options (rather than as simply different types of *Fragestellung*). In the context of a quantitative health characteristic (of individuals) the equivalent of a prevalence study is the assessment of (parameters of) the distribution of the characteristic (mean, etc) among people. The counterpart of incidence studies in such a case is the study of the distribution of changes in the characteristic over a period of time.

Whatever type of outcome parameter (P) is considered, its relationship to the determinants (D) considered is viewed in either *descriptive* or *causal* terms, and this duality of interpretation has bearing on the structure of the function as well. A descriptive function is simply of the form

$$P = f(\{D\}), \quad (\text{eq 1})$$

where *D* represents a set of determinants ( $D_1, D_2, \dots$ ) and *f* the functional relationship. By contrast, a causal relationship between the parameter and any given determinant depends on modifiers (M) of the effect, and it is expressed conditionally on confounders (C):

$$P = f(D, \{M\} \mid \{C\}). \quad (\text{eq 2})$$

It may be worthy of emphasis that even the latter function is a directly empirical one; the causal-inferential judgement comes to bear on it in the selection of the set of confounders on which it is conditioned — a question of study design (and data analysis).

The realizations of both the outcome and the determinant(s) tend to be functions of *time* (age, duration of follow-up, etc). In terms of the interrelationship of their time referents, one may opt for either a *cross-sectional* or a *longitudinal* function.

In a cross-sectional function these time referents are the same:

$$P_t = f(D_t), \quad (\text{eq 3})$$

ie, the value of the outcome parameter at any given point in time ( $T = t$ ) is related to the realization(s) of the determinant(s) at that same time. In a longitudinal function the time referent of the determinant value(s) is previous to that of the parameter:

$$P_t = f(D_{T < t}). \quad (\text{eq 4})$$

(It should be noted that even though causal functions are theoretically longitudinal, consideration of practicalities may lead to the pursuit of a cross-sectional empirical function.)

*Example 1.* In the Collaborative Perinatal Study (3) the main concern was with potential teratogenic effects of maternal drug use. Theoretically, *incidence* of malformations over the period of organogenesis could have been related to drug exposure at that time, ie, a cross-sectional incidence function could theoretically have been taken as the object of the study. The study actually addressed the *prevalence* of malformations (and other anomalies) in the postnatal period in relation to fetal drug exposure (and other factors), ie, a *longitudinal* prevalence function.

*Example 2.* The Framingham Heart Study (2) was mainly concerned with the occurrence of coronary heart disease in terms of both descriptive and causal-interpretative functions. It could, theoretically, have focused on prevalence, but it concentrated on incidence/risk. When, in that study or in any study, the incidence over a particular span of time (5-a incidence, say) is expressed as a function of age (at the beginning of the risk period) and the values of other determinants at that age, the incidence function is totally cross-sectional. When values before that age are taken into account, the function is longitudinal in terms of the given definition.

## Base

In a *prevalence* study the population experience (that constitutes its base) may be a *cross-section* of a population, ie, such that each member of the population is considered at one point in time (age, time since first exposure, ...) only. In such a study the concern is the health status at that time and the realization(s) for the determinant(s) at that time (cross-sectional function) or at a previous time (longitudi-

nal function). With the subjects distributed over time, a cross-sectional experience can provide for studying even time itself as a determinant of the occurrence parameter. An alternative is to consider the experience of a *cohort* as it moves over the time span under study.

*Example 3.* In the Collaborative Perinatal Study the health outcome was studied from birth to 7 a of age. One option would have been to study each member of the cohort of newborns only once in that span of age, with a suitable scatter of subjects within the range, ie, to examine only a cross-section (an oblique one) of the cohort. Actually, the cohort was followed, by means of periodic examinations, from birth to 7 a of age, ie, the full cohort experience was observed.

An *incidence* study cannot be based on a cross-section of a population; the observation of transitions (from health to illness, say) requires longitudinal population experience, as in the movement of a cohort over time. An alternative to a cohort base is the experience of a *dynamic population* over time.

*Example 4.* While, in the Framingham Heart Study, the base was taken as a cohort of 1948 residents of the town, an alternative would have been to follow the *dynamic* population of Framingham — by repeatedly surveying it for the determinants under study and maintaining a register of coronary events. (The experience of the 1948 cohort, now rapidly fading, would have been subsumed under such a dynamic base, potentially studiable in perpetuity.)

Whatever the dynamics of the base in the aforegiven terms, its *distribution* according to any given determinant and its respective modifiers and confounders in the function, the *design matrix*, is, in principle at least, for the investigator to decide on. In nonexperimental research the decisions are implemented by selectivity only, and thus the main options in this regard are *nonselective* and *selective* distribution or matrix. For a determinant under study, selectivity means pursuit of greater variability. This definition applies to modifiers as well, given that modification is actually studied; if it is not, the distribution of the modifier may be constrained to a narrow range. In the case of a confounder there is no point in maximizing variability, whereas restricting range is a means of control.

*Example 5.* In the Collaborative Perinatal Study, expectant mothers were enrolled, and their pregnancies and offspring followed, regardless of what their drug use was in early pregnancy. An alternative would have been to be selective according to drug use — taking, say, all users of drugs of interest (such that their use is reasonably common) and only a sample of those who did not use any drugs.

*Example 6.* In the Framingham Heart Study, the screenees were enrolled without any selectivity according to the determinants of interest. Among the alternatives would have been to take people in the extremes of each determinant (“two-point design”), possibly supplemented by a sample from the middle of the distribution (“three-point design”). Similarly, within the broad age range of admissibility, age being a potential modifier of major interest, the cohort was totally nonselective. Again, the alternatives would have included the two-point design, etc.

In situations in which a single determinant is under study, *matching*, by modifier(s) and/or confounder(s), represents an added form of selectivity in the formation of the study base.

(In nonexperimental research, as in general, the choice of the design matrix has to do with study efficiency in the sense of amount of information per subject.)

The studied occurrence function generally involves but a few of a multitude of determinants of the parameter at issue. All the other determinants jointly determine the “*background*” level of the rate or other parameter, say the “*intercept*” of a rate function. It is a question of design to choose the preferred background level of the parameter. For example, in the evaluation of preventives (factors capable of neutralizing otherwise sufficient causes), it is commonly preferred to use a base with a high background rate for the outcome at issue.

The placement of the study base in time involves, in nonexperimental research, the choice between *retrospective* and *prospective* options — given that the research problem is scientific (abstract-general). If it is particularistic, as in the evaluation of health practices, then the problem itself determines the time (and place) for the experience to be studied.

*Example 7.* The Collaborative Perinatal Study could not have been based on any cohort experience (from birth to 7 a of age) in the *past* because information about drug exposure in early pregnancy could not have been obtained.

Even the use of a prospective cohort (of newborns) would not have been a solution *per se*. A partial solution would have been the use of a prospective cohort with the mothers interviewed in an appropriate manner in the early postpartum period. In point of fact, the prospective placement of the cohort was exploited even further; a setting was created in which newborn babies had recorded histories of drug exposure in early pregnancy. (Mothers were enrolled already during pregnancy, and drug uses were recorded forthwith on entry, with updates later in pregnancy.)

*Example 8.* For the Framingham Heart Study, information on the determinants of interest was not available retrospectively. It was, therefore, made available by means of examinations on the prospective cohort that formed the base of that study. [The same problem could have been solved by the use of a prospective dynamic population (cf example 4).]

As for the *place* in which the study base is located, options analogous to the options in time exist (on the same condition). However, these options do not reduce to a single dichotomy analogous to that for time.

(Evidently, the options in time and place have implications for quality of information and efficiency in the sense of cost per subject. In addition, selection of time and/or place can be used as a means of attaining a desired design matrix and/or level of "background" rate.)

## Representation

The base (including its size) having been defined, it remains to ascertain what the empirical occurrence function in it was (retrospective base) or will turn out to be (prospective base). To this end, one needs to learn about numerators and denominators of rates — how the cases and the base, respectively, are distributed over the determinant, modifiers, and confounders. One way to achieve this information is the use of a *census*: each subject in the base is examined as to all the pertinent facts — determinant(s), modifiers, confounders and outcome. An alternative is *outcome-selective* sampling, ie, the use of a case-referent (case-control) approach. (It is to be noted that sampling according to the determinant(s)/modifiers/confounders is not an alternative to the census approach in the context of abstract objectives; it is

tantamount to reducing the base of the sample).

Outcome-selective sampling is customarily thought of in terms of a census (or possibly sample) of the cases of illness together with a sample of the *noncases* (1, 4, 5, 8). Consider a base experience as laid out in panel A of table 1, ie, a base which is either a cross-section of a population (prevalence study) or a cohort experience (incidence study), with a binary determinant and outcome. For the base the rate ratio contrasting the index category ( $D = 1$ ) to the reference category ( $D = 0$ ) is

$$RR_1 = (C_1/B_1)/(C_0/B_0) = (C_1/C_0)/(B_1/B_0). \quad (\text{eq } 5)$$

The ratio  $C_1/C_0$  is estimable from the case series, and, if the illness is rare,  $B_1/B_0$  can be estimated from the series of noncases (1). Thus, using the notation in panel B of table 1,

$$\hat{RR} = (c_1/c_0)/(n_1/n_0). \quad (\text{eq } 6)$$

The aforementioned, customary type of outcome-selective study (in the case of a cross-sectional or cohort base) has an alternative which seems not to have received proper attention: replacement of the sample of noncases by a *sample of the base*. In terms of the notation in panel C of table 1, this design provides the estimate

$$\hat{RR} = (c_1/c_0)/(b_1/b_0). \quad (\text{eq } 7)$$

This estimate, in contrast to the one from the design with noncases (equation 6) does not depend on any rare disease assumption. Its statistical treatment is outlined in appendix 1.

The distinction between the presented two ways of defining the reference series in outcome-selective sampling is a nonissue in the context of incidence studies with a dynamic base; the noncases are a sample of the base (of candidates for incident case), and equation 6 (as well as equation 7) gives an estimate of the incidence-density ratio without any rare disease assumption (7).

*Example 9.* In the Collaborative Perinatal Study the census approach to the experience of the cohort was employed. All information on drug exposure, etc, was secured and proc-

essed for each baby in the study, and even a very detailed editing, referring back to the original data sheets, employed this census approach (3). An alternative would have been simply to file the prenatal records, then ascertain the health outcome on each baby/child, and finally process and analyze the data on all "cases" (representing problems frequent enough for meaningful study) and on a sample of the base cohort of newborns.

*Example 10.* Had the Framingham Heart Study been carried out in terms of a dynamic base as outlined in example 4, the register data would presumably have been processed in detail, while the survey results would ideally have been processed routinely to a minimal, necessary extent only. For example, electrocardiograms would have been filed away without any readings, etc. Any given analysis would have been based on a case series (census, from the register) together with a sample of the (dynamic) base, drawn on the basis of the rosters of screenees simultaneously with the appearances of the cases (time-matching).

Outcome-selective series may, of course, be drawn with or without *matching* on modifiers and/or confounders. [However, matching on factors that are not part of the occurrence function can be counterpro-

ductive in terms of efficiency in studies of this type (6)].

With or without matching, which means selectivity in the sampling of potential reference subjects only, it may be desirable to employ *selectivity for both series*, index (case) as well as reference (noncase or base) series, according to the determinant and/or the modifiers.

Consider first the added selectivity by the determinant in an already outcome-selective study. Commonly the interest is in a rare exposure, so that  $B_1$  is very small relative to  $B_0$ . In such a case, a two-stage sampling strategy may be attractive (in terms of efficiency). The first-stage sampling, nonselective as to the determinant, could be used to identify the exposure status (of cases and of reference subjects). In the second stage, only a sample of the nonexposed would be selected, randomly, from the nonexposed in the first stage sample. If the sampling fractions for the nonexposed in the case and noncase series are  $f_c$  and  $f_n$ , respectively, with second-stage sample sizes of  $c_0'' = f_c c_0$  and  $n_0'' =$

**Table 1.** Layout of numbers of subjects of different types in a cross-sectional or cohort base and also in outcome-selective samples.

A. Base experience

|          | Determinant (D) |         |       | Total    |
|----------|-----------------|---------|-------|----------|
|          | $D = 1$         | $D = 0$ | Other |          |
| Cases    | $C_1$           | $C_0$   | $C'$  | $C + C'$ |
| Noncases | $N_1$           | $N_0$   | $N'$  | $N + N'$ |
| Base     | $B_1$           | $B_0$   | $B'$  | $B + B'$ |

B. Samples of cases and noncases

|          | Determinant (D) |         |       | Total    |
|----------|-----------------|---------|-------|----------|
|          | $D = 1$         | $D = 0$ | Other |          |
| Cases    | $c_1$           | $c_0$   | $c'$  | $c + c'$ |
| Noncases | $n_1$           | $n_0$   | $n'$  | $n + n'$ |
| Total    | $t_1$           | $t_0$   | $t'$  | $t + t'$ |

C. Samples of cases and base

|          | Determinant (D) |         |        | Total        |
|----------|-----------------|---------|--------|--------------|
|          | $D = 1$         | $D = 0$ | Other  |              |
| Cases    | $c_1$           | $c_0$   | $c'$   | $c + c'$     |
| Base     | $b_1$           | $b_0$   | $b'$   | $b + b'$     |
| Cases    | $c_1^*$         | $c_0^*$ | $c'^*$ | $c^* + c'^*$ |
| Noncases | $n_1$           | $n_0$   | $n'$   | $n + n'$     |

$f_n n_0$ , then the estimate in equation 6 is replaced by

$$\hat{RR} = [(c_1/c_0'')/(n_1/n_0'')] (f_c/f_n). \quad (\text{eq } 8)$$

Similarly, if a sample of the base is used instead of a series of noncases and if the sampling fractions of the nonexposed index (case) and reference (base) subjects in the first-stage sample are  $f_c$  and  $f_b$ , respectively, then the estimate in equation 7 is replaced by

$$\hat{RR} = [(c_1/c_0'')/(b_1/b_0'')] (f_c/f_b), \quad (\text{eq } 9)$$

where  $b_0'' = f_b b_0$ . Statistical aspects of these two estimators are outlined in appendix 2.

(Added selectivity by determinant in an already outcome-selective study deserves consideration in situations in which, after the initial selection and ascertainment of exposure status, expensive data acquisition remains to be done. This situation may concern verification or details of diagnosis, or it may deal with modifiers and/or confounders. Also, if exposure is very common so that the exposed are sampled, the data acquisition of concern may deal with details of the exposure.)

Analogously with determinant selectivity, selectivity by *modifier* in an already outcome selective study is aimed at increasing the variability of the modifier in the final series so as to increase the amount of information (about modification) per subject in those series. Thus, in the modifier domain in which the base is scarce, all cases are enrolled (in the first stage of determinant selective sampling), while elsewhere only a fraction of the available cases are drawn into the index series. The size of the reference series in the different domains of the modifier in such a study would generally be proportional to that of the index series (matching).

*Example 11.* Suppose the Framingham Heart Study was carried out in terms of a dynamic base and case-base sampling, as outlined in examples 4 and 10. Suppose further that people with a history of coronary heart disease (CHD) were not excluded from the case register nor in the periodic surveys of the population. Somewhere along the way one might have wished to examine serum cholesterol level as a determinant of acute coronary events, with history of CHD as a modifier of interest. The cases would presumably have been quite nicely (evenly) distributed between the two categories

of the modifier (positive and negative history), so that cases would have been enrolled in the spirit of a census (without selectivity by history). On the other hand, the base sample would have had a very lopsided distribution by the modifier in the absence of matching by it. For some other potential modifiers, such as age, the case series too might have been formed in a selective fashion.

## Epilogue

From the preceding analysis it is evident that the core issue in epidemiologic study design is not the choice between cohort and case-referent studies, contrary to a prevalent belief. Indeed, cohort and case-referent studies are not even alternatives to each other. For a cohort experience, which is a type of study base, the alternatives are a dynamic population experience or a population cross-section, while for a case-referent approach to the ascertainment of the base experience the only alternative is the use of a census.

In the formation of the base, epidemiologists still have a lot to learn from laboratory experimenters, especially in the employment of an efficient design matrix. Even in clinical trials that are immensely expensive it is still customary to stipulate only the *ranges* of age and other modifiers, with no selectivity within the range. (Rarely do laboratory investigators purchase animals from a store simply stipulating a wide range of age or weight and then accept, within that range, a totally arbitrary distribution; it would be recognized as obviously inelegant and inefficient.)

Conversely, experimenters are very committed to the census approach to the ascertainment of the experience of any group of subjects, whether animals or humans, and they might learn from epidemiologists the efficient approach of outcome selectivity.

Even in epidemiology, the use of outcome-selective studies is still rather primitive. The reference series is routinely taken as a series of noncases, even when a sample of the base would be preferable. Moreover, the efficiency potential of further selectivities according to the determinant or modifiers under study seem not to have been realized.

It has been my purpose to draw attention to the various design alternatives that

are available in epidemiologic research. Mere awareness of them, I believe, will lead to more rational choices in study design — with occasionally very major savings through enhanced efficiency.

## Acknowledgment

This work has been supported by grant number 5-Pol-CA06373 from the National Institutes of Health.

## References

1. Cornfield J. A method of estimating comparative rates from clinical data: Applications to cancer of the lung, breast and cervix. *J natl cancer inst* 11 (1951) 1269—1275.
2. Dawber TR, Meadors GF, Moore FE. Epidemiological approaches to heart disease: The Framingham study. *Am j publ health* 41 (1951) 279—286.
3. Heinonen OP, Slone D, Shapiro S. Birth defects and drugs in pregnancy. Ed. David W. Kaufman (3rd printing). PSG Publishing Company Inc, Littleton, MA 1977. 510 p.
4. Lilienfeld AM. Foundations of epidemiology. Oxford University Press, New York, NY 1976. 283 p.
5. MacMahon B, Pugh TF. Epidemiology: Principles and methods. Little, Brown and Co, Boston, MA 1970.
6. Miettinen OS. Matching and design efficiency in retrospective studies. *Am j epidemiol* 91 (1970) 111—118.
7. Miettinen OS. Estimability and estimation in case-referent studies. *Am j epidemiol* 103 (1976) 226—235.
8. Prentice RL. Logistic disease incidence models and case-control studies. *Biometrika* 66 (1979) 403—411.

## Appendix 1

### Analysis under case-base selectivity

For the case-base strategy of sampling, consider the data layout in panel C of table 1 (p 11), with the following refinement of definitions: If the base sample brings up cases that were not included in the original case series itself, such added cases will be included in the final case series, ie, in the first row of the data layout. Consequently, the cases appearing in the base sample can be thought of as a subset of the final case series, ie, as "redundant cases."

In significance testing, the redundant cases are to be omitted, ie, the final case series is to be compared with the noncase subset of the base sample. Thus, the concern is with a layout of the form in panel B of table 1 (p 11). In those terms, the large-sample test is based on Gaussian approximation to a hypergeometric model for the distribution of  $c_1$  — following Mantel & Haenszel (1). The chi-square statistic, one degree of freedom, is

$$X^2 = (c_1 - ct_1/t)^2 / [cnt_1t_0/t^3]. \quad (\text{eq A})$$

For the logarithm of the point estimator of the rate ratio (RR) in equation 7 (p 10), the variance may be derived in terms of a first-order Taylor series approximation (with allowance for the correlation be-

tween  $c_1$  and  $b_1$  conditionally on  $c$  and  $b$ ). The result is

$$\hat{V}_{\ln(RR)} = 1/c_1 + 1/c_0 + (1 - 2c^*/c) (1/b_1 + 1/b_0). \quad (\text{eq B})$$

Thus,  $100(1 - \alpha)\%$  confidence limits for RR may be set as

$$\underline{RR}, \overline{RR} = \exp [\ln(RR) \pm \chi_\alpha (\hat{V}_{\ln(RR)})^{1/2}], \quad (\text{eq C})$$

where  $\chi_\alpha$  is the (positive square root of the  $100(1 - \alpha)\%$  centile of the chi-square distribution with one degree of freedom.

Alternatively, the limits may be computed by the use of the test-based method (2):

$$\underline{RR}, \overline{RR} = \hat{RR}^{1 \pm \chi_\alpha / X}, \quad (\text{eq D})$$

where  $X$  is the square root of the test statistic in equation A.

*Example.* Suppose the final case series, with some cases found only on the basis of the base sample, included 10 in the index category ( $D = 1$ ) of the determinant (D) under study, and 50 in the reference category ( $D = 0$ ), ie, that  $c_1 = 10$ ,  $c_0 = 50$ , and  $c = 60$ . Suppose, too, that the sample of the base included 10 with  $D = 1$  and 90 with  $D = 0$ , ie, that  $b_1 = 10$ ,  $b_0 = 90$ , and  $b = 100$ . Suppose, finally, that the corresponding numbers of cases in the base

sample were 5 from  $D = 1$  and 15 from  $D = 0$ , so that  $c_1 = 5$  and  $c_0 = 15$ , leaving 5 noncases from  $D = 1$  ( $n_1 = 5$ ) and 75 noncases from  $D = 0$  ( $n_0 = 75$ ,  $n = 80$ ). Thus, by equation A,

$$X^2 = [10 - 60(10 + 5)/(60 + 80)]^2 / [60(80)(10 + 5)(50 + 75)/(60 + 80)^3] = 3.89.$$

The point estimate of  $RR$  is, by equation 7 (p 10),  $\hat{RR} = (10/50)/(10/90) = 1.80$ , so that  $\ln(\hat{RR}) = 0.588$ . The variance of this log-metameter is, by equation B,

$$\hat{V}_{\ln(\hat{RR})} = 1/10 + 1/50 + [1 - 2(20)/60] / (1/10 + 1/90) = 0.157.$$

Thus, 95 % approximate confidence limits for  $RR$  are (equation C)

$$\overline{RR}, \underline{RR} = \exp [0.588 \pm 1.96 (0.157)^{1/2}] = 0.8, 3.9.$$

The corresponding test-based limits (equation D) are

$$\overline{RR}, \underline{RR} = (1.80)^{1 \pm 1.96/(3.89)^{1/2}} = 1.0, 3.2.$$

## References

1. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J natl cancer inst* 22 (1959) 719-748.
2. Miettinen OS. Estimability and estimation in case-referent studies. *Am j epidemiol* 103 (1976) 226-235.

## Appendix 2

### Analysis under selectivity by outcome and determinant

When the sampling fractions in the two series are the same ( $f_c = f_n$  or  $f_c = f_b$ ), the second-stage sampling according to the determinant does not influence the analytic procedures at all; equation 8 (p 12), becomes analogous to equation 6, etc. However, the general case ( $f_c \neq f_n$  or  $f_c \neq f_b$ ) involves some subtlety beyond that in the estimators in equations 8 and 9 (p 12).

In the context of the general case, consider first the estimator in equation 8 (p 12), based on a case-noncase sampling and valid only if the illness is rare (in exposure as well as in nonexposure). The underlying data may be laid out in the form of a  $2 \times 2$  table with, say,  $c_1$  and  $c_0$  constituting the first row and  $n_1$  and  $n_0$ , respectively, the second row. Suppose this layout is viewed as an ordinary  $2 \times 2$  table, with  $100(1-\alpha)$  confidence limits for the odds ratio computed (conditionally on the marginal frequencies) either exactly (2) or by the Cornfield asymptotic method (1). Let these limits be  $\overline{OR}$ ,  $\underline{OR}$ . Significance (two-sided) at the  $\alpha$  level corresponds to this interval not covering  $f_n/f_c$ . Similarly the confidence limits for  $RR$  (on the rare-disease assumption) may be taken as

$$\overline{RR}, \underline{RR} = \overline{OR} (f_c/f_n), \underline{OR} (f_c/f_n).$$

With case-base sampling for which the point estimator is given in equation 9

(p 12), the significance testing can be carried out on the basis of cases and noncases as already outlined, given that the proportion of cases which come up only in the base sample (cf appendix 1) is very small. (This situation is guaranteed by the use of a census for the cases.) For use in obtaining the actual p-value and in the computation of the test-based limits on this same condition, the chi-square statistic, one degree of freedom, may be computed as

$$X^2 = (c_1 - E_0)^2/V_0,$$

where  $E_0$  is the null expectation of  $c_1$  and  $V_0$  its null variance, both computed conditionally on the marginal frequencies.  $E_0$  is computed on the basis of the property that its associated odds ratio equals  $f_b/f_c$ . The inverse of  $V_0$  is the sum of the inverse of  $E_0$  and the inverses of its associated other cell frequencies.

## References

1. Cornfield J. A statistical problem arising from retrospective studies. In: Neuman J, ed. *Proceedings of the third Berkeley symposium on mathematical statistics and probability*. Volume 4. University of California Press, Berkeley, CA 1956, pp 135-148.
2. Thomas DG. Exact confidence limits for an odds ratio in a  $2 \times 2$  table. *Appl stat* 20 (1971) 105-110.