

ON THE NEED FOR THE RARE DISEASE ASSUMPTION IN CASE-CONTROL STUDIES

SANDER GREENLAND¹ AND DUNCAN C. THOMAS²

Greenland, S. (UCLA School of Public Health, Los Angeles, CA 90024), and D. C. Thomas. On the need for the rare disease assumption in case-control studies. *Am J Epidemiol* 1982;116:547-53.

The conditions under which matched and unmatched odds ratios are consistent estimators of the incidence-density ratio in case-control studies are examined. Under "incidence-density" sampling, in which controls are selected from those at risk at the time of onset of each case, the matched estimator is shown to be consistent. In contrast, the unmatched estimator is biased unless the proportion exposed in the population at risk is constant over the study period; the bias is, however, negligible unless there is very large variation in the proportion exposed. No assumption of rarity of the disease is needed for these results. On the other hand, when the risk ratio is the parameter of interest, the assumption of rarity is needed for the odds ratio to be a consistent estimator. In such situations, the odds ratio obtained under "incidence-density" sampling will in general provide a better approximation to the risk ratio than will the odds ratio obtained under "cumulative-incidence" sampling, in which controls are selected from those still unaffected at the end of the study period. Even if the disease is rare, however, the odds ratio obtained under cumulative sampling need not consistently estimate any parameter of interest unless the proportion exposed is approximately constant.

biometry; case-control studies; odds ratio; risk

One of the basic principles in the analysis of case-control studies is that the odds ratio can be used as an estimator of the ratio of incidence rates, provided the disease under study is "rare." The condition that the disease be rare, first suggested by Cornfield (1), has frequently

been considered necessary and appears in most textbooks (e.g., MacMahon and Pugh (2)). However, the justification for this assumption has often been obscured by use of the term "incidence rate" for two distinct measures of occurrence (3), one being the expected number of new cases per unit of person-time at risk (variously termed the incidence density, person-time incidence rate, instantaneous incidence rate, or hazard rate), the other being the expected proportion of a fixed population-at-risk that develops the disease over some specific period (sometimes termed the cumulative incidence rate). In an attempt to clarify matters, Miettinen (4) distinguished two types of ratio measures of effect: the "incidence-density ratio" (IDR), for the ratio of two incidence

Received for publication February 1, 1982, and in final form April 26, 1982.

Abbreviations: CIR, cumulative incidence rate; ID, incidence density; IDR, incidence density ratio; OR, odds ratio; RR, risk ratio.

¹ Division of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90024.

² Dept. of Epidemiology and Health, McGill U., 3775 University Street, Montreal, Quebec, H3A 2B4, Canada.

The authors thank Connie Brown and Marlene Dyck for assistance in manuscript preparation.

This research was partially supported by NCI grant no. R01-CA-16042.

densities; and the "risk ratio" (RR), for the ratio of two cumulative incidence rates. Cornfield was concerned with the latter measure in his discussion of the use of the odds ratio. With the widespread application of the proportional hazards model, the incidence-density ratio has become the parameter of primary interest in survival analysis and modeling of chronic disease incidence (5, 6), while the risk ratio is still of primary interest in studies of conditions with brief or variable risk periods, such as perinatal mortality or acute intoxication (7).

Miettinen's major conclusion was that the need for the rare disease assumption depends on the sampling design and the measure to be estimated. In both of the designs he described, cases are selected from those developing the disease in some population over a period of time (t_0, t_1). Under "cumulative-incidence sampling," controls are selected from those still unaffected at t_1 . This is the design for which use of the odds ratio was first proposed and which continues to be the model in most textbooks. As noted by Miettinen, the rare disease assumption is indeed necessary if one wishes to estimate the risk ratio using the sample odds ratio generated under this design.

In the other design described by Miettinen, "incidence-density sampling," one or more controls are selected for each case from those at risk at the time of onset of the case. For this design, Miettinen claimed that the unmatched odds ratio (i.e., the familiar "cross-products ratio" (2))

is a statistically consistent estimator of the incidence-density ratio (i.e., converges asymptotically to the desired parameter), with no assumption of rarity being required. In this paper, we show that this is not in general correct, though for most situations, the error in the approximation is negligible. We also provide a proof that the matched-pairs odds-ratio estimator (8) does consistently estimate the incidence-density ratio when the design is of the "incidence-density" type and the matching is random except for sampling time. This was first shown by Sheehe (9) and later generalized to arbitrary matching schemes by Prentice and Breslow (5). Finally, we show that the odds ratio obtained under "incidence-density" sampling is in general a better approximation to the risk ratio than is the odds ratio obtained under "cumulative-incidence" sampling.

The relationships discussed in the next section apply to case-control studies conducted on either of two types of study population. The first type is a "dynamic" population-at-risk, which is continually being depleted by incidence of the disease, death, emigration or other losses and replenished by immigration and births. Such a dynamic population need not necessarily be in a "steady state." The second type of population is a fixed cohort, which may be depleted in a similar way but is not replenished. In terms of theory, the fixed cohort is a special case of the dynamic population, one in which immigration is not allowed.

RELATIONSHIPS BETWEEN ODDS RATIOS, INCIDENCE-DENSITY RATIOS, AND RISK RATIOS

Define for the population under study over the time interval (t_0, t_1): $N(t)$ = total size of the population-at-risk at time t ; $P_1(t)$ = proportion of $N(t)$ in the exposed group at t ; $P_0(t) = 1 - P_1(t)$ = proportion of $N(t)$ in the unexposed group at t ; $ID_i(t)$ = incidence density in exposure group i at t ($i = 0$ for unexposed and $i = 1$ for exposed); $IDR(t) = ID_1(t) / ID_0(t)$ = incidence-density ratio at t .

Density sampling: constant incidence-density ratio

Unmatched estimator. First let us assume the $IDR(t)$ is equal to a constant, \overline{IDR} , over the study interval. Now in the subinterval $(t, t + dt)$ the expected number of incident cases in exposure group i is $a_i(t) dt = N(t) P_i(t) ID_i(t) dt$, leading to $A_i = \int a_i(t) dt$ for the number of incident cases expected over the interval (t_0, t_1) . (Throughout, the limits of integration are, implicitly, t_0 and t_1 .) Under density sampling, one control will be selected for each case at the occurrence time of the case. The expected number of controls selected from exposure group i over the subinterval $(t, t + dt)$ is then $b_i(t) dt = a_+(t) P_i(t) dt$, where $a_+(t) = a_0(t) + a_1(t)$, leading to $B_i = \int b_i(t) dt$ controls being expected in subgroup i over the total study interval. The unmatched odds ratio estimator OR_U would thus consistently estimate the parameter

$$OR_U = A_1 B_0 / A_0 B_1.$$

OR_U is a function of $P_i(t)$ and the $ID_i(t)$, and is not in general equal to \overline{IDR} . However, if $P_1(t)$ is constant over (t_0, t_1) , further simplification is possible. Substituting the definitions of $a_i(t)$, $b_i(t)$ and \overline{IDR} given above produces

$$\begin{aligned} OR_U &= \frac{\int a_1(t) dt \int b_0(t) dt}{\int a_0(t) dt \int b_1(t) dt} \\ &= \frac{\int N(t) P_1 \overline{IDR} ID_0(t) dt \int a_+(t) P_0 dt}{\int N(t) P_0 \overline{IDR} ID_0(t) dt \int a_+(t) P_1 dt} \\ &= \overline{IDR}, \end{aligned}$$

where $P_1(t) \equiv P_1$.

Thus the unmatched odds ratio is a consistent estimator of the incidence-density ratio if the exposed proportion does not vary over the study interval. (Several common exposure factors tend to have constant distributions within dynamic populations, e.g., gender, blood type, and, to a lesser extent, ethnicity, although in a fixed cohort, differential survival will lead to distributional changes.)

Matched estimator. Assuming sampling was random within each matching subinterval $(t, t + dt)$, the expected number of pairs over the subinterval for which the case is in exposure group i and the control is in exposure group j is $m_{ij}(t) dt = a_i(t) P_j(t) dt$, so that the total number of such pairs expected over (t_0, t_1) is $M_{ij} = \int m_{ij}(t) dt$. The matched odds ratio estimator OR_M (8) would thus estimate the parameter

$$OR_M = M_{10} / M_{01}.$$

Again, substituting the definitions of $m_{ij}(t)$, $a_i(t)$, and \overline{IDR} produces

$$\begin{aligned} OR_M &= \frac{\int N(t) P_1(t) \overline{IDR} ID_0(t) P_0(t) dt}{\int N(t) P_0(t) \overline{IDR} ID_0(t) P_1(t) dt} \\ &= \overline{IDR}. \end{aligned}$$

Thus the matched odds-ratio estimator is a consistent estimator of the incidence-density ratio, with no assumption of stability of the proportion exposed $P_1(t)$.

Density sampling: variable incidence-density ratio

In the general case of $IDR(t)$ varying over the interval (t_0, t_1) , neither \hat{OR}_U nor \hat{OR}_M appear to estimate any parameter of immediate appeal, although \hat{OR}_M does estimate a weighted average of the $IDR(t)$. Specifically, we have that \hat{OR}_M estimates $IDR_+ = \int W(t)IDR(t)dt / \int W(t)dt$, where $W(t) = N(t)P_1(t)P_0(t)ID_0(t)$. If the proportion exposed varies over time, these weights have no simple interpretation, but if it is constant, IDR_+ becomes a standardized morbidity ratio, $IDR_+ = SMR = A_1/E_1$, where E_1 denotes the expected number of cases in the exposed population if the unexposed rates applied. A calculation similar to the one for OR_U in the constant IDR case shows that OR_U estimates IDR_+ if the proportion exposed is constant.

We note that in all of the above derivations, no assumptions are required about either the rarity of disease or the constancy of the incidence densities. More generally, it is also unnecessary to assume that only one control be selected for each case or even that the matching ratio be fixed, but the proof for the matched estimator becomes more complicated. Similarly, the restriction to a single, time-invariant, binary exposure factor was unnecessary and was made only to simplify the presentation. Proofs applying to more general situations have been given by Prentice and Breslow (5).

Unmatched density sampling

Rather than time-matching controls to cases, it may be possible to select controls in a fashion so that the expected exposed proportion among them equals the exposure rate among the total person-time at risk in the source population (4). The expected number of exposed and unexposed cases are the same as in the matched case; and assuming constant IDR , OR_U becomes

$$\frac{\overline{IDR} \int ID_0(t)P_1(t)N(t)dt / \left[\int ID_0(t)P_0(t)N(t)dt \right]}{\int P_1(t)N(t)dt / \left[\int P_0(t)N(t)dt \right]}$$

this being the ratio of the expected exposure-odds for the cases to the expected exposure-odds for the controls. In general, this expression does not equal \overline{IDR} unless either $ID_0(t)$ or $P_1(t)$ is constant, and thus \hat{OR}_U will not in general estimate \overline{IDR} under unmatched density sampling. Analogously to OR_M , however, an odds ratio estimate based on sufficiently fine stratification on t will consistently estimate \overline{IDR} from such a sampling design. Thus in situations involving time-dependence of ID_0 and P_1 , t must be treated as a confounder in order to avoid bias in the estimation of \overline{IDR} , even if the controls provide an unbiased estimate of the total proportion of person-time exposed over the study interval.

Density sampling versus cumulative sampling

It is apparent from the preceding derivations that density sampling permits direct estimation of the incidence-density ratio, and thus should be preferred to cumulative sampling when the incidence-density ratio is the parameter of interest.

Consider now the situation in which the risk ratio is the parameter of interest.

The cumulative incidence rate (CIR) in exposure group i over t_0 to t_1 (given by references 3, 4 and 8) is

$$CIR_i = 1 - \exp(-\int ID_i(t)dt).$$

The risk ratio parameter is thus $RR = CIR_1/CIR_0$. If the disease is "rare," we have that

$$RR = \frac{CIR_1}{CIR_0} \approx \frac{\int ID_1(t)dt}{\int ID_0(t)dt} = \overline{IDR} = OR_M.$$

Thus, the matched odds ratio derived under density sampling is an approximation to the risk ratio if the disease is rare. This approximation should be compared to Cornfield's (1), which involves the use of the "risk-odds" ratio given by

$$OR_R = \frac{CIR_1(1 - CIR_0)}{CIR_0(1 - CIR_1)}$$

as an approximation to RR. Table 1 shows such a comparison for a range of values likely to be encountered in epidemiologic studies. It appears that the estimation of IDR as an approximation to RR will be better than the classical approach of estimating OR_R as an approximation to RR. Examples involving a variable IDR yield similar conclusions.

A serious difficulty arises with the classical approach because OR_R is *not* consistently estimated by the unmatched odds ratio estimator \hat{OR}_{CI} obtained under cumulative-incidence sampling, except in some very special cases (\hat{OR}_{CI} is the odds ratio discussed in most introductory texts, e.g., MacMahon and Pugh (2)). In fact, \hat{OR}_{CI} consistently estimates $OR_{CI} = A_1P_0(t_1)/A_0P(t_1)$, which may take on arbitrarily large or small values depending on the size of the exposed proportion at t_1 . The special case considered by Cornfield (1) involved a fixed population with no

exposure-related withdrawals or losses, in which case OR_{CI} does equal OR_R . It is also interesting to note that if both the exposed proportion and the IDR are constant, OR_{CI} equals \overline{IDR} rather than OR_R . In general, however, \hat{OR}_{CI} is not a consistent estimator for any parameter of interest. These results indicate that, even if the risk ratio is the parameter of interest, density sampling is preferable to cumulative sampling when the sample odds ratio is to be used as the estimate of RR.

Multiple sampling occurrences under density sampling

Under density sampling, it is possible for a person to be sampled as a control several times for several different cases, or to be sampled as a control and then develop the disease and be sampled as a case (5). Such individuals must be counted repeatedly in the data (once for each sampling occurrence) to allow application of the above theory regarding density sampling. Density sampling coupled with matched analysis corresponds to doing a failure time analysis of the population experience, but examining only a sample of the population-at-risk at each case occurrence time rather than the entire population (5). A study population member is part of the population-at-risk at all times before death, disease occurrence, emigration, or study termination; therefore, that person must remain part of the control sampling frame for all sampling times (case occurrence times) before one of the latter events, even if the person has already been sampled as a control or later becomes a case.

EXAMPLES

Case-control study within a dynamic population

As an example of the potential bias inherent in the use of OR_U as an estimator for a constant incidence-density ratio \overline{IDR} , consider the following example

TABLE 1

Comparison of odds ratio (OR) approximations to the risk ratio (RR) under different sampling designs and constant incidence-density ratio (IDR)

CIR ₀ *	CIR ₁	RR	OR _U (= \overline{IDR})	OR _R
0.02	0.05	2.5	2.539	2.579
0.01	0.05	5.0	5.104	5.211
0.005	0.05	10.0	10.233	10.474
0.04	0.10	2.5	2.581	2.667
0.02	0.10	5.0	5.215	5.444
0.01	0.10	10.0	10.483	11.000
0.08	0.20	2.5	2.676	2.875
0.04	0.20	5.0	5.466	6.000
0.02	0.20	10.0	11.045	12.250

* CIR, cumulative incidence rate.

based on a recent case-control study of estrogens and endometrial cancer (10). Over the two years of the study, the proportion of approximately 10,000 women, aged 50 to 64 years with intact uteri, who were currently using conjugated estrogens was estimated to decline from 10.5 per cent in 1975 to 4.4 per cent in 1977, probably because of the appearance of the first reports of an estrogen-cancer association in December 1975. The incidence densities also declined from 1.07 to 0.67 per 1000 woman-years in nonusers and from 31.2 to 16.7 per 1000 woman-years in users, but the incidence-density ratio was relatively stable (29.1 in 1975 versus 25.0 in 1977). To simplify the discussion, let us assume that $N(t) = 10,000$, $P_1(t) = 0.10 - 0.025t$, $ID_0(t) = (1 - 0.25t)/1000$ years, and $ID_1(t) = (30 - 7.5t)/1000$ years, where t runs from 0 to 2 years. Then computing directly from the formulae given earlier, $A_0 = 13.83$, $A_1 = 35.00$, $B_0 = 38.51$, $B_1 = 3.89$. Hence $OR_U = 25.1$, compared with $OR_M = 30.0 = \overline{IDR}$; thus the bias in the unmatched estimator is about -15 per cent. The cumulative incidence ratio in this example is $RR = 29.4$, much closer to OR_M than OR_U . If unmatched density sampling had been performed, OR_U would have equalled 31.2.

In this study, controls were selected from women hospitalized for acute illness or elective surgery at about the same time as the cancer cases (though not individually matched). It is worth noting, however, that had controls been drawn at the end of the study, the odds ratio estimator \hat{OR}_{CI} would have been a consistent estimator of $OR_{CI} = 35.0(0.95) / (13.8(0.05)) = 48$. OR_{CI} should be contrasted with the parameter that \hat{OR}_{CI} is usually assumed to estimate, i.e., the risk-odds ratio OR_R . OR_R is equal to 30.7 in this example; this illustrates the severe bias that can occur when cumulative-incidence sampling is used, even if the disease is rare. Note that the use of density sampling in the *design*

prevents most of the bias in estimating the incidence-density ratio or risk ratio; the use of the matched *estimator* is then called for, although the bias resulting from the use of the unmatched estimator on the matched data is rather small unless the proportion exposed varies considerably over the study period.

Case-control study within a fixed cohort

Another illustration of the error of the approximations is provided in table 2 for some situations in which a case-control study is conducted within a fixed cohort, the $ID_i(t)$ are constant, and $P_1(t_0) = P_0(t_0) = 0.5$. We see that \hat{OR}_U tends to underestimate \overline{IDR} and overestimate RR , but the bias in estimating \overline{IDR} is negligible unless the cumulative incidence rates are enormous (say, greater than 50 per cent in either subgroup). On the other hand, the odds ratio based on cumulative incidence sampling (\hat{OR}_{CI}) tends to overestimate \overline{IDR} and RR . Note that in all instances the odds ratios are closer to \overline{IDR} than RR , and that the odds ratios derived under density sampling are always better approximations to RR and \overline{IDR} than the odds ratios derived under cumulative sampling.

SUMMARY

The "cumulative incidence" sampling design for case-control studies has been the model used in most textbooks because of its conceptual simplicity. As is well known, the risk ratio interpretation of the odds ratio arising from such studies strictly requires an assumption of disease rarity; less commonly emphasized is that if the exposed proportion varies over the study period, the study odds ratio need not consistently estimate any parameter of interest. The rarity requirement is usually not a practical issue; indeed, the bias from lack of rarity becomes substantial (greater than 10 per cent) only when the cumulative incidence over the study

TABLE 2
*Expectations of estimators under cumulative incidence and incidence density sampling plans for case-control studies**

Incidence densities (per 1000 person-years)			Proportion exposed		Five-year cumulative incidence rate			Sampling plan		
Un- exposed, ID ₀	Ex- posed, ID ₁	IDR	At start, P ₁ (0)	At five years, P ₁ (6)	Unexposed	Exposed	RR	Cumulative incidence		Incidence density
								OR _{CI}	OR _U	OR _M
4	5	1.250	0.500	0.499	0.01980	0.02469	1.247	1.253	1.250	1.250
2	5	2.000	0.500	0.497	0.01242	0.02469	1.988	2.013	2.000	2.000
1	5	5.000	0.500	0.495	0.00499	0.02469	4.950	5.050	5.000	5.000
40	50	1.250	0.500	0.488	0.18127	0.22120	1.220	1.283	1.250	1.250
25	50	2.000	0.500	0.463	0.11750	0.22120	1.883	2.133	1.999	2.000
10	50	5.000	0.500	0.450	0.04877	0.22120	4.536	5.540	4.994	5.000
400	500	1.250	0.500	0.378	0.86466	0.91792	1.062	1.750	1.249	1.250
250	500	2.000	0.500	0.223	0.71350	0.91792	1.287	4.490	1.964	2.000
100	500	5.000	0.500	0.119	0.39347	0.91792	2.333	17.238	4.512	5.000

* RR, OR_{CI}, OR_U, and OR_M are the asymptotic expectations of the estimators \hat{RR} , \hat{OR}_{CI} , \hat{OR}_U , and \hat{OR}_M , respectively.

interval is greater than about 10 per cent, which is larger than the lifetime risk of most diseases. On the other hand, moderate changes in the exposed proportion over the study period can lead to considerable bias in results of a cumulative-incidence case-control study.

Incidence-density sampling is a more appropriate model of the prevailing practice in chronic disease epidemiology (11): first, it is now widely recognized that only incident cases should be used; second, in perhaps the majority of studies, controls are matched to cases at least on age and time of diagnosis. When density sampling is coupled with the matched odds ratio estimator, no assumptions about the exposed proportion are necessary; and if the incidence-density ratio is of primary interest, the rare disease assumption can be eliminated as well. If the risk ratio is of primary interest the rarity assumption must still be invoked, but density sampling generally provides a better odds ratio approximation to the risk ratio than the cumulative design.

REFERENCES

1. Cornfield J. A method of estimating comparative rates from clinical data. *JNCI* 1951;11:1269-75.
2. MacMahon B, Pugh JF. *Epidemiology—principles and methods*. Boston: Little, Brown and Co., 1970.
3. Morgenstern H, Kleinbaum DG, Kupper LL. Measures of disease incidence used in epidemiologic studies. *Int J Epidemiol* 1980;9:97-104.
4. Miettinen OS. Estimability and estimation in case-referent studies. *Am J Epidemiol* 1976;103:226-35.
5. Prentice RL, Breslow NE. Retrospective studies and failure time models. *Biometrika* 1978;65:153-8.
6. Thomas DC. General relative risk models for survival time and matched case-control analysis. *Biometrics* 1981; 37:673-86.
7. Neutra RR, Drolette ME. Estimating exposure-specific disease rates from case-control studies using Bayes' theorem. *Am J Epidemiol* 1978;108:214-22.
8. Schlesselman JJ. *Case-control studies: design, conduct, analysis*. New York: Oxford University Press, 1982.
9. Sheehe PR. Dynamic risk analysis in retrospective matched pair studies of disease. *Biometrics* 1962;18:323-41.
10. Jick H, Watkins RN, Hunter JR, et al. Replacement estrogens and endometrial cancer. *N Engl J Med* 1979;300:218-22.
11. Cole P. The evolving case-control study. *J Chronic Dis* 1979;32:15-27.