ESTIMABILITY AND ESTIMATION IN CASE-REFERENT STUDIES

OLLI MIETTINEN


The concepts that case-referent studies provide for the estimation of "relative risk" only if the illness is "rare," and that the rates and risks themselves are inestimable, are overly superficial and restrictive. The ratio of incidence densities (forces of morbidity)—and thereby the instantaneous risk-ratio—is estimable without any rarity-assumption. Long-term risk-ratio can be computed through the coupling of case-referent data on exposure rates for various age-categories with estimates, possibly from the study itself, of the corresponding age-specific incidence-densities for the exposed and nonexposed combined—but again, no rarity-assumption is involved. Such data also provide for the assessment of exposure-specific absolute incidence-rates and risks. Point estimation of the various parameters can be based on simple relationships among them, and in interval estimation it is sufficient simply to couple the point estimate with the value of the chi square statistic used in significance testing.

biometry; statistics

The principles that currently govern epidemiologic thinking as to the fundamentals of case-referent (case-“control”) studies do not apply to the most common type of such study in chronic-disease epidemiology. Here the principles are extended to encompass this kind of study. A simple, general-purpose statistical approach is also proposed. The results presented are generally self-evident, but some explanations are offered in appendix 1.

1. The classical principles

1.1. Essence. The prevailing principles concerning the estimability of parameters in case-referent studies derive from a classical paper by Cornfield (1). The principles might be expressed as follows (1, 2): First, the ratio of the odds of developing the illness for the exposed as compared to the non-exposed equals the ratio of the odds of having been exposed, contrasting cases of the illness to a reference series, and therefore the illness-odds ratio contrasting the exposed to the non-exposed is estimable from case-referent studies; and second, this parameter is approximately equal to the risk ratio when the illness is rare. The rationale is as follows (1, 2): Given risks of illness \( R_1 = A/(A + C) \) and \( R_0 = B/(B + D) \) for exposed and non-exposed people, respectively, the odds ratio for the illness is \( [R_1/(1 - R_1)]/[R_0/(1 - R_0)] = AD/BC = (A/B)/(C/D) \). The last formulation for the odds ratio for illness between the exposed and the non-exposed reveals the identity of this parameter with the odds ratio for past exposure between cases and non-cases. Obviously, the ratio \( A/B \) is estimable from a series of cases, and \( C/D \) can be estimated from a reference (comparison, “control”) series. Finally, the odds ratio parameter can be seen to equal the risk ratio \( R_1/R_0 \) itself on the condition that \( (1 - R_0)/(1 - R_1) = 1 \), and this condition obtains with

Received for publication April 3, 1975, and in final form July 23, 1975.

From the Departments of Epidemiology and Biostatistics, Harvard School of Public Health, and Department of Cardiology, Children's Hospital Medical Center, Boston, MA 02115.

Supported by Grants 5 PO1 CA 06373 and HE 10436 from the National Institutes of Health.
good approximation if the illness is rare.

1.2. Applicability. Upon careful appreciation of that rationale it is apparent that the classical principles of estimability apply, as such, to a particular type of case-referent study only. This special type is the one in which the subjects are ascertained at or after the end of the entire risk-period of interest. Such studies, though commonplace in acute-disease epidemiology, are rare in the chronic-disease field. (A conspicuous example is, however, the study of teratogenesis by means of ascertaining malformed and healthy newborns and comparing their exposure-histories in reference to the period of organogenesis.)

If formulated in terms of prevalence rather than risk, the classical rationale for estimability in case-referent studies also implies that studies based on prevalent cases provide for the estimation of prevalence-odds ratio; and when the prevalences are low, this parameter is practically interchangeable with the prevalence ratio itself.

The classical rationale does not, however, bear on the ordinary type of case-referent study in chronic-disease epidemiology—the type of study in which ascertainment occurs before the individual risk-periods are over, and in which incident rather than prevalent cases are enrolled.

2. The nature of the study

For a given exposure and illness, the objectives of a case-referent study are basically no different from those of a follow-up ("cohort") study. Thus, with reference to populations it is desired to learn about rates of occurrence of the illness in relation to the exposure (possibly in causal terms), within categories of age and other characteristics; and for individuals the concern is with risks (for various time periods) of the development of the illness in relation to the exposure, conditional on age and other characteristics.

The defining features of case-referent studies are that a series of people with and another without the illness are enrolled, and that their profiles with respect to the exposure, past or present, are ascertained and compared.

The internal validity of the study involves the following components: a) validity of selection: the probability of ascertainment is uninfluenced by the exposure history or status itself; b) validity of observation: lack of misclassification between cases and non-cases (referents, comparands, "controls") and between exposed and nonexposed; and c) validity of comparison: the use of a reference entity (usually diagnostic category) unrelated to the exposure, and the control of confounding.

3. Incidence density

3.1. The parameters. Incidence density ("force of morbidity" or "force of mortality")—perhaps the most fundamental measure of the occurrence of illness—is the number of new cases divided by the population-time (person-years of observation) in which they occur. For scientific purposes this measure is more meaningful if the experience of only actual candidates for the illness are considered in defining the population-time, i.e., if prevalent cases are not counted as contributing to the follow-up experience. For example, the incidence density of death is more meaningfully—and routinely—expressed in reference to follow-up experience with the living rather than with the living and the dead combined. In these terms, then, for the exposed described in figure 1 the incidence density (ID$_1$) in the time interval from t' to t'' is defined as

$$ID_1 = a''/C(t'' - t')$$ \hspace{1cm} (1)

instead of $ID_1 = a''/(A + C) (t'' - t')$. For the nonexposed, similarly,

$$ID_0 = b''/D(t'' - t').$$ \hspace{1cm} (2)

It follows that the incidence density ratio (IDR) relating the exposed to the nonexposed is

$$IDR = (a''/b'')/(C/D),$$ \hspace{1cm} (3)
while the other relative measure, incidence density difference (IDD), is

\[
IDD = (a''/C - b''/D)/(t'' - t').
\]  
(4)

3.2. Estimability of ratio. In a case-referent study involving incident rather than prevalent cases, the cases (a exposed and b non-exposed) provide for the estimation of \(a''/b''\) (as \(a/b\)), and \(C/D\) is estimable from the reference series (as \(c/d\), the ratio of the sample numbers of exposed and non-exposed referents from the total pool of candidates for the illness). Consequently, the incidence density ratio is estimable from such studies; and in particular, no rarity-assumption is required for this. If the study is based on prevalent rather than incident cases, then it is necessary to assume that the duration of the illness is unrelated to the exposure.

3.3. Estimability of absolute parameters and difference. If the overall incidence density, for the exposed and the non-exposed combined, is known, then case-referent data provide for the estimation of the exposure-specific values and, thereby, for the estimation of the difference measure of relative occurrence (IDD). Even though this is well known, the prevailing principles of this estimation (1, 2) require added specificity as well as extension.

Sometimes a case-referent study involves complete ascertainment of new cases (over a particular time period) in a well-defined population of known size \(N = A + B + C + D\); cf. figure 1) (3, 4). If the reference series (of size \(n\)) is drawn from the total source populations, i.e., without excluding prevalent cases in the ascertainment, then it is always feasible to estimate \(C\) (as \(cN/n\)) and \(D\) (as \(dN/n\)) and therefore \(ID_a\) and \(ID_b\) themselves (cf. formulas 1 and 2) as well as their difference. No rarity-assumption is involved in this. In (the usual) instances in which the reference series is (unnecessarily) confined to non-cases, this type of estimation of \(ID_a\) and \(ID_b\) is feasible if it is realistic to assume low prevalence ((\(C + D\))/\(N\) ≈ 1).

When those conditions do not obtain, it still is commonplace to have a priori knowledge about the overall incidence density (ID) for the exposed and non-exposed combined (for various categories of age and sex). Ordinarily, however, the overall incidence density is not known in the proper terms, with prevalent cases excluded in the computation of the population-time of experience. When this is the case, the reference series should again be drawn without excluding prevalent cases, and the proper incidence density (ID) may then be estimated from the available, improper value (\(ID^*\)) as

\[
\widehat{ID} = (ID^*)n/(c + d),
\]  
(5)

where \(n\) is the size of the reference series from the age category at issue, and \(c + d\) is the size of the unaffected subgroup of it.

Given an estimate of the overall incidence density, whether from the study itself or from an outside source, the separate estimates for the exposed and nonex-
posed can usually be determined from the relations

$$ID_1 = (IDR) \cdot (ID) \cdot (1 - EF)$$  \hspace{1cm} (6)

and

$$ID_0 = (ID) \cdot (1 - EF),$$  \hspace{1cm} (7)

respectively, where EF is the etiologic fraction (proportion of all cases) related (or perhaps even attributable) to the exposure. The IDR is estimable as already discussed, and the EF has (5) the structure of

$$EF = [(IDR - 1)/(IDR)] \cdot (ER_1),$$  \hspace{1cm} (8)

where $ER_1$ is the exposure rate among incident cases (i.e., $ER_1 = a''/(a'' + b'')$). This approach is applicable when the association between the exposure and the illness is non-negative in the data ($\overline{IDR} \geq 1$). Otherwise one may use the relationships

$$ID_1 = (IDR) \cdot (ID)/(1 - PF)$$  \hspace{1cm} (9)

and

$$ID_0 = (ID)/(1 - PF),$$  \hspace{1cm} (10)

where PF is the preventive fraction related to the exposure. It may be estimated (5) through the expression

$$PF = (1 - IDR) \cdot (ER_1)/[(1 - ER_1)(IDR) + ER_1].$$  \hspace{1cm} (11)

4. Cumulative incidence-rate and risk

4.1. The parameters. Even though the source population of subjects tends to constitute a (dynamically) static group in each category of age (figure 1), with new people continually entering it at the lower bound (and within the range) of age and others exiting it (within the range and) at the upper bound, the data from successive age categories may be used to make inferences about an aging cohort of fixed membership and homogeneous, though continually increasing, age.

With regard to such an age-cohort, the interest is, firstly in the cumulative incidence of the group as it passes from one age to another, and, secondly, in the corresponding risks of its individual members. The cumulative incidence-rate for a span of age is the proportion of the group developing the illness in that period, while the risk for an individual is the probability of his developing the illness in the particular span of age.

As a function of incidence density (ID) of first episodes of the illness (among those who never had it), the cumulative incidence-rate (CIR) for the age span $a'$ to $a''$ is

$$CIR_{a', a''} = 1 - \exp \left[-\int_{a'}^{a''} (ID_a) \, da \right].$$  \hspace{1cm} (12)

If age is discretized, then the cumulative incidence-rate (conditional on survival) over successive categories $j'$ through $j''$ has the approximate expression of

$$CIR_{j', j''} \approx 1 - \exp \left[-\sum_{j=j'}^{j''} (ID_j)w_j \right],$$  \hspace{1cm} (13)

where $w_j$ is the width of the $j^{th}$ category. When the cumulative incidence-rate is small, say less than 10 per cent, it may be reasonably approximated as

$$CIR_{a', a''} \approx \int_{a'}^{a''} (ID_a) \, da$$  \hspace{1cm} (14)

or as

$$CIR_{j', j''} \approx \sum_{j=j'}^{j''} (ID_j)w_j.$$  \hspace{1cm} (15)

Risk ($R$, for an individual) is the expected value of the cumulative incidence-rate (for a group):

$$R_{a', a''} = \mathbb{E}(CIR_{a', a''})$$  \hspace{1cm} (16)

4.2. Estimability of ratio. The estimability of cumulative incidence and risk from case-referent data is dependent, through the above relationships, on the estimability of incidence density.

The ratio of instantaneous risks is identical to the ratio of the corresponding incidence densities (cf. formula 15), so that the instantaneous risk ratio is estimable (through the exposure-odds ratio of incident cases) without any rarity-assumption in reference to either incidence density or prevalence.

For a longer span of age, from the beginning of category $j'$ to the end of category $j''$, the risk ratio (assuming survival from other
illnesses) is somewhat complicated: Even if the risk over that age span for both the exposed and the non-exposed is low enough to justify the use of formula 15, the corresponding risk ratio \((RR)\) approximation is

\[
RR_{j'} = \sum_{j''} (ID_{j'})w_{j''} \sum_{j''} (ID_{j''})w_{j'}
\]

where \(W_j = (ID_{j''})w_{j''}\), and where the subscripts 1 and 0 refer to the exposed and the non-exposed, respectively, as before. Thus, the (point) estimation of the risk ratio over several categories of age, even when the risks themselves over that age span are small, involves the computation of a weighted average of the age-specific density ratios; and what is more, the weights involve the actual incidence densities of the non-exposed or at least numbers proportional to these. The determination of the weights can pose a serious problem, although the relationships in equations 7 and 10 tend to be very helpful. If the risks are not "low," then it is necessary to compute the ratio directly from the exposure-specific estimates of absolute risk.

4.3. Estimability of absolute parameters and difference. The exposure-specific risks over several age-categories, together with the corresponding risk differences, are in principle estimable through the application of formula 13. This requires that the overall age-specific incidence densities, for the exposed and non-exposed combined, are estimable from the data or known a priori, so that exposure-specific incidence densities are estimable within the categories of age (see section 3).

5. Prevalence rate

As was already noted, a case-referent study based on prevalent cases provides for the estimation of prevalence-odds ratio; and if the prevalence rate is low among both the exposed and the non-exposed, then the prevalence-odds ratio is approximately equal to the prevalence ratio itself. This principle allows the estimation of (age-specific) ratio measures of relative prevalence with great ease, both conceptual and procedural.

But when a case-referent study is based on incident cases, as it usually is, inference about the relative prevalence among the exposed and the non-exposed involves some subtlety. Consider first a closed population with a static profile over time as to age-distribution etc. People make transitions from the candidate pool to the prevalence pool, and from the prevalence pool either back to the candidate pool or out of the population through death (figure 1). The static state is characterized by an equilibrium between the inceptions of new cases and the terminations of prevalent ones. Specifically, in a population of size \(N\), the equilibrium prevalence rate \(PR\), a fraction, satisfies the relationship \(N(1 - PR)(ID) = N(PR)(TD)\) or

\[
(1 - PR)(ID) = (PR)(TD),
\]

where \(TD\) stands for the termination density in the prevalence pool, i.e., for the number of case terminations (by cure or death) divided by the case-time of experience in which they occurred. In the equilibrium state \(TD = 1/\bar{D}\), the inverse of the mean duration of the illness. Substitution of this into equation 18 yields, as the static-state relationship of prevalence to incidence density,

\[
PR = (ID)\bar{D}/[1 + (ID)\bar{D}].
\]

As a deduction, then, the prevalence-odds are simply

\[
(PR)/(1 - PR) = (ID)\bar{D};
\]

and furthermore, for a comparison of the exposed (subscript 1) to the non-exposed (subscript 0), the prevalence-odds ratio \((POR)\) is

\[
POR = [(PR_1)/(1 - PR_1)]/[(PR_0)/(1 - PR_0)]
\]

\[
= (ID_1)\bar{D}_1/(ID_0)\bar{D}_0.
\]

Thus, in a static population, the prevalence-odds ratio is estimable in the same
terms as the incidence density ratio—through the exposure-odds ratio between
the cases and the referents, without any rarity-assumption (cf. section 3). And
again, if both of the exposure-specific prevalences are low, the prevalence-odds ratio
is approximately equal to the prevalence ratio itself.

Now consider a limited category of age. Even if the population of interest in such a
category can be thought of as static (with turnover) over time, and usually this is the
case, the prevalence for it is not represented by formula 19 with the ordinary
meaning for incidence density and duration. Instead, if that formula were to be
used, the tally of incident cases in any given period of time would have to include
the ones that enter the age-specific prevalence pool as carryover cases from the
previous category of age; also, the mean duration of the illness should be adjusted
downward to reflect those terminations within the category which result from cases
reaching the upper bound of the category. The implementation of these considera-
tions, in terms of formula 19 or otherwise, does not lead to any simple expression of
wide applicability.

6. Example

Cole et al. identified all newly-diagnosed cases of bladder cancer in a (static) popula-
tion (eastern Massachusetts) of known size over an 18-month period, drew a reference
series from the source-population of the cases, and inquired (inter alia) into the
subjects’ histories with respect to cigarette-smoking (4). Some of the data (7) are
presented in table 1.

The data allow the computation of age-specific overall incidence densities. For
example, for the 50- to 54-years category the value is 35/(77,400) (1.5 years) = 30/10^5
years (cf. formulas 1 and 2). Actually this result ought to be corrected by allowing for
prevalent cases (formula 5), but such a correction, which would be negligible in
magnitude, is not feasible, because prevai-
lent cases were excluded without tally in
the selection of the reference series. The
age-specific values obtained within the
study are in close conformity with those
derived (without prevalence correction)
from the cancer registry of a neighboring
region (Connecticut) a few years earlier (8).
The latter data, too, are shown in table 1.

The samples of cases and noncases in
each category of age allow the estimation of
the corresponding incidence density ratio
(without any rarity-assumption). For ex-
ample, for the 50- to 54-years age category,
the incidence density ratio (IDR) —i.e., the
incidence density for smokers (ID_s) divided
by that for nonsmokers (ID_n)—is estimated
to be \( \frac{\overline{IDR}}{ID_{n}} = \frac{24/1}{22/4} = 4.36 \) (formula 3).

In order to provide for the estimation of absolute incidence density separately for
smokers and nonsmokers, and as a matter of interest in its own right, the age-specific
estimates for the etiologic fraction (with
IDR \( \geq 1 \)) or the preventive fraction (IDR < 1) are computed next. Thus, for the 50-
to 54-years category age, for which \( \overline{IDR} (= 4.36) > 1 \), the etiologic fraction is
estimated as follows: \( E_F = \frac{(4.36 - 1)}{4.36}24/25 = 0.74 \) (formula 8). For the 60-
to 64-years category \( \overline{IDR} (= 0.49) < 1 \), and therefore the preventive fraction is cal-
culated (without inferring prevention): \( P_F = (1 - 0.49) (31/36)/(1 - 31/36) (0.49) +
31/36 = 0.47 \) (formula 11).

The incidence density estimates specific
for the exposed and the non-exposed are
then computed by the use of either for-
ulas 6 and 7 (if \( \overline{IDR} \geq 1 \)) or formulas 9
and 10 (if \( \overline{IDR} < 1 \)). For the 50- to 54-years
category the estimate for smokers is (4.36)
(30/10^5 years) (1 - 0.74) = 34/10^5 years
(formula 6), while the corresponding result
for nonsmokers is (30/10^5 years) (1 - 0.74)
= 8/10^4 years (formula 7). In the 60- to
64-years category the estimate for smokers
is 0.49 (56/10^5 years)/(1 - 0.47) = 52/10^5
years (formula 9), while for nonsmokers it
is (56/10^5 years)/(1 - 0.47) = 110/10^5 years.

Turning to the assessment of risk, con-
Two-digital accuracy.

Risk difference estimate: 
\[ \text{Risk ratio estimate: } H = \frac{\text{ratio of cases in smokers}}{\text{ratio of cases in non-smokers}} \]

Non-smokers: \[ H = \frac{0.00}{0.00} = \frac{0}{0} = 0 \text{ per cent} \]

Smokers: \[ H = \frac{0.00}{0.00} = \frac{0}{0} = 0 \text{ per cent} \]

30-year risk at age 50 years from other causes.

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Study connection</th>
<th>Case (cases)</th>
<th>No. of source cases of source</th>
<th>Case (cases)</th>
<th>No. of source cases of source</th>
<th>Overall incidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>70-79</td>
<td>100</td>
<td>38.0</td>
<td>76</td>
<td>23.2</td>
<td>76</td>
<td>0.00</td>
</tr>
<tr>
<td>70-74</td>
<td>100</td>
<td>47.4</td>
<td>70</td>
<td>23.2</td>
<td>70</td>
<td>0.00</td>
</tr>
<tr>
<td>65-69</td>
<td>100</td>
<td>47.1</td>
<td>64</td>
<td>22.4</td>
<td>64</td>
<td>0.00</td>
</tr>
<tr>
<td>60-64</td>
<td>100</td>
<td>47.0</td>
<td>60</td>
<td>22.4</td>
<td>60</td>
<td>0.00</td>
</tr>
<tr>
<td>55-59</td>
<td>100</td>
<td>47.0</td>
<td>56</td>
<td>22.4</td>
<td>56</td>
<td>0.00</td>
</tr>
<tr>
<td>50-54</td>
<td>100</td>
<td>47.0</td>
<td>52</td>
<td>22.4</td>
<td>52</td>
<td>0.00</td>
</tr>
<tr>
<td>45-49</td>
<td>100</td>
<td>47.0</td>
<td>48</td>
<td>22.4</td>
<td>48</td>
<td>0.00</td>
</tr>
<tr>
<td>40-44</td>
<td>100</td>
<td>47.0</td>
<td>44</td>
<td>22.4</td>
<td>44</td>
<td>0.00</td>
</tr>
<tr>
<td>35-39</td>
<td>100</td>
<td>47.0</td>
<td>40</td>
<td>22.4</td>
<td>40</td>
<td>0.00</td>
</tr>
<tr>
<td>30-34</td>
<td>100</td>
<td>47.0</td>
<td>36</td>
<td>22.4</td>
<td>36</td>
<td>0.00</td>
</tr>
<tr>
<td>25-29</td>
<td>100</td>
<td>47.0</td>
<td>32</td>
<td>22.4</td>
<td>32</td>
<td>0.00</td>
</tr>
<tr>
<td>20-24</td>
<td>100</td>
<td>47.0</td>
<td>28</td>
<td>22.4</td>
<td>28</td>
<td>0.00</td>
</tr>
<tr>
<td>15-19</td>
<td>100</td>
<td>47.0</td>
<td>24</td>
<td>22.4</td>
<td>24</td>
<td>0.00</td>
</tr>
<tr>
<td>10-14</td>
<td>100</td>
<td>47.0</td>
<td>20</td>
<td>22.4</td>
<td>20</td>
<td>0.00</td>
</tr>
<tr>
<td>5-9</td>
<td>100</td>
<td>47.0</td>
<td>16</td>
<td>22.4</td>
<td>16</td>
<td>0.00</td>
</tr>
<tr>
<td>0-4</td>
<td>100</td>
<td>47.0</td>
<td>12</td>
<td>22.4</td>
<td>12</td>
<td>0.00</td>
</tr>
<tr>
<td>0-1</td>
<td>100</td>
<td>47.0</td>
<td>8</td>
<td>22.4</td>
<td>8</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Cases as well as non-cases. See Section 6.

complete ascertainment of non-smokers' cases in a population (estimates of ascertainment of cases) of non-smokers is by age. Inferences were confounded in a sample of

cases referred to table of (a) and (b) referring the incidence of bladder cancer in a sample of non-smokers. The studies included

Table 1
sider the 30-year risk of bladder cancer for a 50-year-old man, assuming that without bladder cancer he would survive that period. If the man is a smoker, then the estimate is \( \hat{R}_{50, 50} = 1 - \exp \left| - \frac{(34 + 54 + 52 + 140 + 230 + 260)}{10^3} \right| \) years \( \approx 1 - \exp (-0.0385) = 1 - \text{antilog}(0.0385) = 3.8 \) per cent (formulas 16 and 13). Almost the same result can be obtained more simply from the approximate expression in formula 15. For a nonsmoker the corresponding estimate is 2.2 per cent. The estimate of the 30-year risk ratio at age 50 years is, then, 3.8/2.2 = 1.7, and the corresponding risk difference estimate is (3.8 - 2.2) per cent = 1.6 per cent.

7. Statistical aspects

7.1. Point estimation. As was illustrated in the above example, the various parametric relationships that were set forth provide for straightforward point estimation of the various parameters of most direct interest. In the expressions for those parameters, the component parameters were simply replaced by their "sample values." This is essentially tantamount to maximum-likelihood estimation, with all its desirable properties, in large samples in particular.

7.2. Interval estimation. For incidence density ratio at any given age, large-sample 100(1 - \( \alpha \)) per cent two-sided confidence limits (\( \text{IDR} \) and \( \text{IDR}^- \)) may be set simply as

\[
\text{IDR}, \text{IDR}^- = (\text{IDR})^{1 + g_{a/2}/\chi},
\]

where \( g_{a/2} \) is the 100(1 - \( \alpha/2 \)) percentile of the standard Gaussian distribution, and where \( \chi \) is the positive square root of the 1 d.f. chi square statistic for significance testing of the association (9). The chi may derive from the ordinary test for a single two-by-two table or from the Mantel-Haenszel procedure (10). Correspondingly, the point estimate (\( \text{IDR} \)) is either the "cross-product ratio" from a single two-by-two table (1) or an appropriate estimate based on multiple tables (11). The limits for the incidence density ratio are also the limits for the instantaneous risk ratio.

For the risk ratio from age \( a' \) to age \( a'' \), the limits may be set in an analogous manner:

\[
RR_{a', a''}, \overline{RR}_{a', a''} = (\overline{RR}_{a', a''})^{1 + g_{a/2}/\chi},
\]

The chi still derives from the overall significance test. For example, if limits were to be set for the risk ratio for which the point estimate was derived in table 1 and section 6, the chi value would be computed in terms of the Mantel-Haenszel test statistic (10) for the totality of age-specific two-by-two tables for which the data are given in table 1.

For the corresponding risk difference (\( RD \)) the limits may be taken as

\[
RD_{a', a''}, \overline{RD}_{a', a''} = (\overline{RD})(1 + g_{a/2}/\chi).
\]

For the etiologic and preventive fractions the upper confidence bound cannot exceed unity, while the lower bound must be zero when \( g_{a/2} = \chi \) (and also when \( g_{a/2} > \chi \)). Those constraints suggest the use of the limits

\[\begin{align*}
\text{EF}, \overline{\text{EF}} &= 1 - (1 - \overline{EF})^{1 + g_{a/2}/\chi} \\
\text{(with } \overline{EF} \geq 0) \end{align*}\]

and

\[\begin{align*}
\text{PF}, \overline{\text{PF}} &= 1 - (1 - \overline{PF})^{1 + g_{a/2}/\chi} \\
\text{(with } \overline{PF} \geq 0). \end{align*}\]

As to the incidence density among the exposed (formulas 6 and 9) or the non-exposed (formulas 7 and 10) the limits may be set as

\[
ID_i, \overline{ID_i} = (\overline{ID_i}) \exp (\pm g_{a/2} \overline{\nu_i}^{1/4}),
\]

\( i = 1, 0 \), where \( \overline{\nu_i} \) is the variance estimate of the natural logarithm of \( \overline{ID_i} \). This variance estimate may be taken as

\[
\overline{\nu_i} = \overline{\nu} + (\ln(\overline{ID_i}) - \ln(\overline{ID}))^2/\chi^2,
\]

where \( \overline{\nu} \) is the variance estimate for the natural logarithm of the estimated overall incidence density (\( \overline{ID} \)), computable as

\[\begin{align*}
\overline{\nu} &= 1/(a'' + b''), \end{align*}\]

i.e., as the inverse of the total number of cases involved in the estimate.
Finally consider confidence limits for the exposure-specific risks, such as the ones examined in table 1 and section 6. In the usual situation, in which the risks are quite low, the limits may be taken as

\[ R^{-1} = \exp (\pm s_{\alpha/2}/\hat{V}) \]

where \( \hat{V} \) is an estimate of the variance of the natural logarithm of the point estimate. For the overall risk (formulas 13 and 15) one may use

\[ \hat{V} = \sum_i [(\hat{a}_{ij} + b_{ij})/w_i^2] \]

where \( a''_j + b''_j \) is the number of cases on which \( ID_j \) is based. For the exposed and the non-exposed, the corresponding variances may be taken as

\[ \hat{V} = \hat{V} + (\ln \hat{R} - \ln \hat{R})^2/\chi^2, \]

where \( \chi^2 \) is still the 1 d.f. chi square statistic (10) for testing the hypothesis of no association.

Example. As an illustration of interval estimation of risk, consider again the data in table 1 and section 6, specifically the 30-year risk at age 50 years. The point estimate of the overall risk according to formula 13 (and 16) is \( \hat{R} = 1 - \exp \left\{-[(30 + 51 + 56 + 120 + 180 + 220)/10^4 \text{ years}] \right\} = 1 - \exp (-0.0329) = 3.2 \) per cent. I or the variance of its logarithm the estimate (formula 31) is \( \hat{V} = [(30^3)/35 + \ldots + (220^2)/76]/(5 \text{ years})^2/(0.0323)^2 = 0.0030 \). The corresponding 95 per cent confidence limits (formula 30) are, then, \( \hat{R} \), \( \hat{R} = (0.032) \exp \left\{ \pm 1.96(0.0030)^{1/2} \right\} = 2.9 \) per cent, 3.6 per cent. For the Mantel-Haenszel chi square, consider the exposed cases; the observed number is 24 + \ldots + 39 = 235, while the expectation (10) and variance (10) are 220.4 and 22.12, respectively, giving \( \chi^2 = (235 - 220.4)^2/22.1 = 9.6 \). Thus, with risk estimates \( \hat{R} = 3.8 \) per cent and \( \hat{R} = 2.2 \) per cent (section 6), the variance estimate for the logarithm of the risk for smokers (formula 32) is \( \hat{V} = 0.0030 + (\ln 0.038 - \ln 0.032)^2/9.6 = 0.0061 \), so that \( \hat{V} = 0.078 \). With this, the 95 per cent confidence limits (formula 30) are (3.8 per cent) \( \exp \{ \pm 1.96(0.078) \} = 3.3 \) per cent, 4.4 per cent.

References

7. Cole P: Personal communication, 1975

Appendix 1

Test-based confidence limits

Ordinarily, large-sample confidence limits for a parameter (\( \pi \)) are set as

\[ \pi = f^{-1}[f(\hat{\pi}) \pm s_{\alpha/2}(SE_f(\hat{\pi}))]. \]

The transformation function \( f \) is chosen with the aim of attaining a Gaussian and stable-variance sampling distribution for the metameter \( f(\hat{\pi}) \) of the point estimate (\( \hat{\pi} \)); and the standard error, \( SE_f \), is usually computed as a first-order Taylor series approximation, i.e., as \( (SE_f)^{f(\hat{\pi})} \).

In the context of the estimates dealt with in the above, the formulation of the standard error according to the ordinary principles would tend to involve substantial complexity. At the same time, point esti-
mation and significance-testing are quite simple.

This suggests the computation of the standard error from the point estimate and the test statistic. The rationale of this approach may not be completely transparent in the results offered, and some explanatory remarks may therefore be in order.

Consider first a parameter \( \pi \) with an expressly known null value \( \pi_0 \) corresponding to the absence of any association between the exposure and the illness, i.e., a parameter such as rate ratio \( \pi_0 = 1 \), rate difference \( \pi_0 = 0 \) or etiologic fraction \( \pi_0 = 0 \). Given that the metaterm is successfully chosen (vide supra),

\[
[f(\hat{\pi}) - f(\pi_0)]^2 / [SE_{f(\hat{\pi})}]^2 = \chi^2, \tag{A.2}
\]

if \( \pi = \pi_0 \), the chi square having one degree of freedom. Solving this for the standard error and substituting the result to formula A.1 yields

\[
\pi, \pi = f^{-1}[f(\hat{\pi}) \pm g_{\alpha/2} [f(\pi) - f(\pi_0)] / \chi], \tag{A.3}
\]

where \( \chi \) is a square root (positive or negative) of \( \chi^2 \). Finally, the chi value in this formulation may in fact be obtained from the Mantel-Haenszel statistic (10), which bears on the same null hypothesis. This principle underlies formulas 22-26, with no transformation in formula 24, and with the transformation \( f(\cdot) = \ln[1 - (\cdot)] \) in formulas 25 and 26, the inverse function being \( f^{-1}(\cdot) = 1 - \exp(-\cdot) \).

When the null value is not firmly known, as when dealing with absolute exposure-specific rates or risks (formulas 27 and 30), formula A.1 is still used. Here the computation of the standard error is somewhat more complicated. We have, analogously with formula A.2,

\[
[f(\hat{\pi}_i) - f(\pi_0)]^2 / [SE_{f(\hat{\pi}_i) - f(\pi_0)}]^2 = \chi^2, \tag{A.4}
\]

with the subscripts referring to the exposed and non-exposed respectively, as before. But equivalently,

\[
[f(\hat{\pi}_i) - f(\hat{\pi})]^2 / [SE_{f(\hat{\pi}_i) - f(\hat{\pi})}]^2 = \chi^2, \tag{A.5}
\]

\( i = 1, 0 \), where \( \hat{\pi} \) is the overall estimate for the exposed \( (i=1) \) and non-exposed \( (i=0) \) combined. As a further modification,

\[
[f(\hat{\pi}_i) - f(\hat{\pi})]^2 / [SE_{f(\hat{\pi}_i)}]^2 - [SE_{f(\hat{\pi})}]^2 = \chi^2, \tag{A.6}
\]

since \([SE_{f(\hat{\pi}_i)}]^2 = [SE_{f(\hat{\pi}_i) - f(\pi)}]^2 + [SE_{f(\hat{\pi})}]^2\). This implies that

\[
SE_{f(\hat{\pi}_i)} = [SE_{f(\hat{\pi})}]^2 + [f(\hat{\pi}_i) - f(\hat{\pi})]^2 / \chi^2 \tag{A.7}
\]

as in formulas 28 and 32.

As to the choice of the metaterm, the square root transformation might be preferred to the logarithmic one in formulas 27 and 30. This would imply

\[
\pi, \pi = \{\hat{\pi}_i \pm g_{\alpha/2} [(SE_{\hat{\pi}})^2/4\hat{\pi} + (\hat{\pi} - \pi_0)^2/\chi^2])^{1/2}\} \tag{A.8}
\]
**Figure 1.** Static population (e.g. a particular age-
Figure 1. Static population (e.g. a particular age-group) over the time-span of a case-referent study based on incident cases. The sizes of the different component populations remain static, but there is turnover of membership in each compartment. The arrows indicate occurrences of new cases, i.e., transitions from the candidate pools to the prevalence pools. Note that the incidence-densities are zero in each of the prevalence pools, and that the incident cases are referred to the follow-up experiences in the candidate pools only. Also note that the incidence-density ratio is \((a''/b'')/(C/D)\), with \(a''/b''\) and \(C/D\) estimable from the (incident) cases and referents, respectively, regardless of the levels of incidence or prevalence.