

PROPORTION OF DISEASE CAUSED OR PREVENTED BY A GIVEN EXPOSURE, TRAIT OR INTERVENTION¹

OLLI S. MIETTINEN

(Received for publication December 13, 1973)

Miettinen, O. S. (Harvard School of Public Health, Boston, Mass. 02115). Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am J Epidemiol* 99: 325-332, 1974.—The structures of two epidemiologic parameters are explored. One, the "etiologic fraction," relates to markers of increased risk, and it is the proportion of disease attributable to the marker and/or to factors associated with it. The other, the "prevented fraction," is the equivalent of this for a marker of reduced risk. It is shown that both parameters depend—in different ways—on the frequency of the marker among cases of the disease, and on the "standardized morbidity ratio" for those with the marker. Point estimation of these parameters is often straight-forward, particularly in case-control studies.

biometry; epidemiologic methods

The quantitation of the disease-producing role of an etiologic factor is pertinent not only from the scientific point of view but for the planning of intervention programs as well. One appropriate parameter for this is the proportion of the disease that is attributable to the factor at issue. Thus, when contemplating coronary heart disease (CHD) control through anti-smoking programs one wonders what fraction of CHD might be due to smoking. Introduced two decades ago (1), this parameter has—rather inexplicably—received little attention from epidemiologists (2, 3).

The preventive role of a protective factor, or the preventive success of an implemented program of intervention, could be

Abbreviations: CHD, coronary heart disease; *CF*, case fraction; *CR*, crude risk; *CRR*, crude risk ratio; *EF*, etiologic fraction; *ER*, expected risk; *MI*, myocardial infarction; *PF*, prevented fraction; *RR*, risk ratio; *SF*, source fraction; *SMR*, standardized morbidity ratio.

¹ From the Departments of Epidemiology and Biostatistics, Harvard School of Public Health, and the Department of Cardiology, Children's Hospital Medical Center, Boston, Massachusetts 02115.

Supported by grants 5 P01 CA 06373 and HE 10436 from the National Institutes of Health.

expressed in terms of an analogous parameter: the fraction of the disease prevented by the factor or program. As a population parameter, this measure, too, has remained without the attention it seems to warrant.

In this paper, previous results as to the structures of these parameters are reiterated and then adjusted to take account of the usual circumstances which involve confounding and also nonuniformity of risk ratio ("relative risk" (*RR*)) among strata of the confounding factor(s).

ETIOLOGIC FRACTION RELATED TO A MARKER OF INCREASED RISK

The proportion of the total load of a disease that is attributable to a given factor is that fraction of the disease which would not have occurred had the factor been absent from the population. This parameter may be regarded as the "fraction of etiology" or "etiologic fraction" attributable to the factor.

In nonexperimental epidemiologic research it would usually be pretentious to offer any specific estimate of the etiologic

fraction *attributable* to a given marker of increased risk—as if confounding by extraneous factors were well under control. Instead, it would be prudent to think in terms of the etiologic fraction *related* to the marker—with certain, though not all, confounding factors under control. Were it known that no material confounding is left, then the observable etiologic fraction would be attributable to that risk marker itself; otherwise that fraction is, at least in part, due to other factors associated with the risk marker at issue.

All-or-none marker of increased risk

Previous results. An expression for the structure of the etiologic fraction (*EF*) related to an all-or-none marker of increased risk was first developed by Levin (1). For those with the marker this fraction was taken as

$$EF_1 = (RR_1 - 1)/RR_1, \tag{1}$$

$$(RR_1 \geq 1)$$

where *RR*₁ is the risk ratio for those with the marker, with those who do not exhibit the marker as the referent (and characterized by *RR*₀ = 1). By the same token, for the referent category, *EF*₀ = 0. For the overall *EF*, Levin gave the expression

$$EF = SF_1(RR_1 - 1) / [1 + SF_1(RR_1 - 1)], \tag{2}$$

$$(RR_1 \geq 1),$$

where *SF*₁ is the source (population) fraction of people with the marker. These results were also derived by Cole and MacMahon (2).

An alternate formulation of the *EF* was offered by Miettinen (4) and Panayotou et al. (5):

$$EF = CF_1 \times EF_1, \tag{3}$$

$$(SMR_1 \geq 1)$$

where *CF*₁ is the case fraction in the higher-risk category of the risk indicator (proportion of cases with the marker of increased risk). Substituting (*RR*₁ - 1)/

*RR*₁ for *EF*₁, yields, as an equivalent of formula 2,

$$EF = CF_1 \times (RR_1 - 1)/RR_1, \tag{4}$$

$$(RR_1 \geq 1)$$

Proposed modifications. These formulas require modification to encompass the usual case which involves not only confounding but also nonuniformity of the *RR* among strata of the confounding factor(s).

For those with the marker, the etiologic fraction (*EF*₁) expresses what proportion of the “crude” (observable) risk (*CR*₁) is in excess of the “expected” risk (*ER*₁) which would result from the referent risks (of those without the marker) obtaining, within categories of the confounding factor(s), among the higher-risk people (those with the marker) as well. Thus, for the higher-risk category of the risk indicator, *EF*₁ = (*CR*₁ - *ER*₁)/*CR*₁ = 1 - *ER*₁/*CR*₁ = 1 - 1/*SMR*₁, i.e.,

$$EF_1 = (SMR_1 - 1)/SMR_1, \tag{5}$$

$$(SMR_1 \geq 1)$$

where *SMR*₁ is the “standardized morbidity (mortality) ratio” in its customary definition (6).

By the same token, the generalization of formula 2 rests on the recognition that in general

$$EF = (CR - ER)/CR, \tag{6}$$

$$(CR \geq ER)$$

where *CR* is the “crude” (observable) risk in the population at issue, and *ER* is the “expected” risk for it—the “expectation” corresponding to the rates of the referent category of the risk indicator within each stratum of the confounding factor(s). Thus, *CR* = *SF*₀ × *CR*₀ + *SF*₁ × *CR*₁ = *CR*₀(*SF*₀ + *SF*₁ × *CRR*₁), where *CR*₀ and *CR*₁ are the “crude” risks in the lower- and higher-risk categories of the indicator of risk, and *CRR*₁ is the “crude” risk ratio *CR*₁/*CR*₀. Similarly, *ER* = *SF*₀ × *CR*₀ + *SF*₁ × *CR*₁/*SMR*₁ = *CR*₀(*SF*₀ + *SF*₁ ×

CRR_1/SMR_1). Consequently, by formula 6,

$$EF = SF_1(CRR_1 - CRR_0/SMR_1) / [1 + SF_1(CRR_1 - 1)]. \quad (7)$$

$$(SMR_1 \geq 1)$$

Comparison of formula 2 to this result shows that, in the previous formula, the RR in both the numerator and the denominator should be the "crude" one, and that in the numerator—which has to do with cases in excess of the "expectation"—the referent for the "crude" RR should not generally be taken as unity, but, naturally enough, as CRR_0/SMR_0 , which is the RR the controlled confounding alone would produce in the absence of control (6). For comparability with formulations involving the distribution of cases, rather than of the source population, according to the risk indicator (formulas 4 and 9) it is of interest to recast formula 7 as

$$EF = [SF_1 \times CRR_1 / (SF_0 + SF_1 \times CRR_1)] (SMR_1 - 1) / SMR_1. \quad (8)$$

$$(SMR_1 \geq 1)$$

The adjustment required for generality in the formula based on case distribution (formula 4) is simply to use SMR_1 in place of RR_1 . Thus, formula 4 should be superseded by

$$EF = CF_1 \times (SMR_1 - 1) / SMR_1. \quad (9)$$

$$(SMR_1 \geq 1)$$

This expression must be interchangeable with formulas 7 and 8. This implies that, in general,

$$CF_1 = SF_1 \times CRR_1 / (SF_0 + SF_1 \times CRR_1). \quad (10)$$

This equivalence is obvious as such. So is the relative simplicity of formula 9 as compared to formula 8.

Polytomous marker of increased risk

Previous result. A generalization of formula 4 to the simplest case of a multicate-

gory indicator of risk has been given by Miettinen (7). This simple case is the one where there is a reasonably nonarbitrary specification of the referent category, and where the overall risk in all other categories is at least as high as would be "expected" on the basis of the confounding-factor-specific risks in the referent. An example of this is smoking as an indicator of CHD risk, with the referent category taken as those who never smoked. The formula that was employed for this case is

$$EF = \sum_i CF_i \times (RR_i - 1) / RR_i, \quad (11)$$

$$(RR_i \geq 1 \text{ for all } i)$$

with the summation ranging over all categories of the indicator. (From the referent category the contribution is zero.)

Proposed modification. Formula 11 is an expression for the principle that $EF = \sum_i CF_i \times EF_i$, with EF_i taken as $(RR_i - 1) / RR_i$. However, in the general case $EF_i = (SMR_i - 1) / SMR_i$ (cf. formula 5) and therefore formula 11 should be superseded by the general formula

$$EF = \sum_i CF_i \times (SMR_i - 1) / SMR_i. \quad (12)$$

$$(SMR_i \geq 1 \text{ for all } i)$$

Alternative. An expression which involves the distribution of the entire source population (rather than of cases only) may be derived from formula 11 by the use of the relationship in equation 10. This alternative to formula 12 is

$$EF = \sum_i [SF_i \times CRR_i / \sum_i SF_i \times CRR_i] (SMR_i - 1) / SMR_i, \quad (13)$$

$$(SMR_i \geq 1 \text{ for all } i)$$

where the summations include the referent category. (The inclusion of the referent is actually pertinent only with respect to the denominator $\sum_i SF_i \times CRR_i$, for which $CRR_0 = 1$ by definition.)

Generalization. Since the above expressions for EF reflect the principle that $EF = \sum_i CF_i \times EF_i$, and since EF_i is defined (as $EF_i = (SMR_i - 1) / SMR_i$) only if $SMR_i \geq$

1, there arises the question of the applicability of formulas 12 and 13 in the more general case where there still is a clear choice of the referent category but for one or more of the other categories the overall risk is lower than that in the referent. In this case, by formula 6, $EF = 1 - ER/CR = 1 - (\sum_i CF_i/SMR_i)/(\sum_i CF_i)$, with the referent included in the summations; but since $\sum_i CF_i = 1$, we have

$$EF = 1 - \sum_i CF_i/SMR_i, \quad (14)$$

(if positive)

still including the referent in the range of the summation. This expression is algebraically identical to formula 12. Thus, in the general case (where possibly some $SMR_i < 1$) EF may be defined even if EF_i cannot be defined for some categories of the risk indicator, and when the EF is defined, the expression of its structure in terms of SMR 's need not be modified when having $SMR_i < 1$ (EF_i undefined) for some of the categories.

A further problem arises in the case of an arbitrary referent. Consider, as an example, the hypertension-related fraction of the etiology of CHD. The risk of CHD is a monotone increasing function of diastolic blood pressure. Thus, whereas diastolic hypertension presumably means diastolic pressure above "normal" values, "normalcy" eludes rational definition. Suppose one arbitrarily chooses the level of 75–84 mm Hg as the referent category. It would then be rather meaningless to construct the parameter as in formulas 12 and 14, including in the summation the relatively protected category of under 75 mm Hg. The summation would, instead, be confined to blood pressure categories in the range of 85 mm Hg or more. The interpretation of this parameter, under the assumption of no confounding, is not simply that of the fraction of CHD etiology attributable to diastolic blood pressure being 85 mm Hg or more; it is to be borne in mind that the implied alternative for those high pressures

is the 75–84 mm Hg range and not the broad and relatively nonhomogeneous range of under 85 mm Hg.

Composite marker of increased risk

The EF related to an aggregate of markers of increased risk does not differ in principle from that related to a single one, because the various combinations of the levels of the component indicators can be construed as categories of a single—though composite—indicator of risk. For example, when dealing with the aggregate of cigarette-smoking, hypertension and hypercholesterolemia in relation to CHD, the referent category of a (single) composite indicator of risk might most naturally be taken as that characterized jointly by no cigarette-smoking, normotension and normocholesterolemia (whatever the definitions).

If two or more indicators of risk are statistically independent and have associated with them etiologic influences which are biologically independent, the overall EF for the composite indicator can be expressed in terms of the component EF 's:

$$EF_{a,b,\dots} = 1 - (1 - EF_a) \cdot (1 - EF_b) \dots \quad (15)$$

(statistical and biological independence)

where EF_a is the EF related to marker "A," etc.

By the same token, under statistical independence, departures from that equality imply either synergy or antagonism – synergy if

$$EF_{a,b,\dots} > 1 - (1 - EF_a) \cdot (1 - EF_b) \dots, \quad (16)$$

(synergy)

and antagonism if

$$EF_{a,b,\dots} < 1 - (1 - EF_a) \cdot (1 - EF_b) \dots \quad (17)$$

(antagonism).

The use of these criteria for the detection of

synergy or antagonism presupposes, in addition to near-independence statistically, direct assessment of the joint *EF* associated with the aggregate of indicators, as well as of the individual *EF*'s, with comparable definitions of referents, etc., in the composite and individual calculations, respectively.

PREVENTED FRACTION RELATED TO A MARKER OF DECREASED RISK

When dealing with a preventive intervention, or with any (person) characteristic or (environmental) exposure which protects against the disease at issue, the equivalent of the *EF* is, as already pointed out, the prevented fraction (of the disease) related to the particular indicator of protection. It is the fraction of the potential total load of the disease which is prevented by the marker of protection and/or factors associated with it.

It appears that no expression for the structure of this parameter has been offered.

All-or-none marker of decreased risk

In the case of a dichotomous indicator of protection, an expression for the prevented fraction (*PF*) specific for the lower-risk category of the risk indicator is

$$PF_1 = 1 - SMR_1 \tag{18}$$

($SMR_1 \leq 1$)

(cf. formula 5), and the overall *PF* has the definition

$$PF = (ER - CR)/ER \tag{19}$$

($ER \geq CR$)

(cf. formula 6).

In terms of the distribution of the risk indicator in the source population, we have, as before, $CR = CR_0(SF_0 + SF_1 \times CRR_1)$ and $ER = CR_0(SF_0 + SF_1 \times CRR_1/SMR_1)$. Substitution to formula 19 yields

$$PF = [SF_1 \times CRR_1 / (SF_0 \times SMR_1 + SF_1 \times CRR_1)] (1 - SMR_1) \tag{20}$$

($SMR_1 \leq 1$)

(cf. formula 8).

An expression for the *PF* in terms of the risk indicator distribution of cases (rather than of the source population) may be taken as

$$PF = CF_1^* \times PF_1, \tag{21}$$

($SMR_1 \leq 1$)

where CF_1^* is the case fraction falling in the lower-risk category among the potential totality of both actual and prevented cases. This hypothetical quantity is $CF_1^* = (CF_1/SMR_1)/(CF_0 + CF_1/SMR_1) = CF_1/(CF_0 \times SMR_1 + CF_1)$. Substitution of this expression for CF_1^* and the PF_1 expression from formula 18 to formula 21 gives

$$PF = [CF_1 / (CF_0 \times SMR_1 + CF_1)] \cdot (1 - SMR_1) \tag{22}$$

($SMR_1 \leq 1$)

(cf. formula 9).

Formulas 20 and 22 must be interchangeable just as the respective formulas for *EF* were (formulas 8 and 9). And again, this interchangeability implies the relationship between CF_1 and SF_1 given in equation 10.

Polytomous and composite markers of decreased risk

The issues relating to the structure of the *PF* in the cases of polytomous and composite indicators of protection are quite analogous to those discussed in the context of *EF*.

For a polytomous indicator of protection, the key relationship is the counterpart of equation 12, namely

$$PF = \sum_i CF_i^* \times (1 - SMR_i), \tag{23}$$

(if positive),

where CF_i^* is the equivalent, for the i^{th} category of the protection indicator, of CF_1^* in formula 21.

A set of indicators whose related overall effect is preventive can again be treated as a single—though composite—indicator of protection analogously with a composite

indicator of etiologic influence. With statistically independent components and (biologically) independent influences, the joint PF is

$$PF_{a,b,\dots} = 1 - (1 - PF_a)(1 - PF_b) \dots \quad (24)$$

(statistical and biological independence).

ESTIMABILITY

Point estimation

From the above exposition of the structures of the etiologic and prevented fractions related to markers of increased and decreased risks, respectively, it is apparent that the point estimation of each parameter involves the estimation of two sets of component parameters, the case fractions (CF_i 's) in the various categories of the risk indicator and the "standardized morbidity ratios" (SMR_i 's) for those categories.

For the CF_i 's one can use the respective sample values as long as the cases included in the study can be considered representative of cases in the source population. In case-control studies this is generally the aim and, more or less, the attained reality. Cohort studies may include electiveness as to the levels of the indicator of risk and, in particular, there tends to be health-related restrictions for inclusion in the study.

Point estimation of SMR_i 's is, in princi-

ple, feasible not only in cohort studies but in case-control studies as well (6, 8). In case-control studies, however, it is necessary to have reasonably large numbers of subjects from each stratum of the confounding factor(s). Thus, the computation of SMR estimates is not feasible in case-control studies with individual matching, and the practical alternative is simply to assume uniformity of the risk ratio and to use an appropriate estimate of that parameter (9) as a substitute for the SMR estimate.

Example. Data from a case-control study relating coffee consumption and myocardial infarction (MI) are shown in table 1 (10). The estimation of the coffee-related etiologic fraction for MI depends only on the distribution of the cases by level of coffee consumption (case fractions at the different levels) together with the standardized morbidity ratios for those categories of exposure (cf. formulas 9 and 12). For the referent category (zero cups per day) the estimate of case fraction is $\widehat{CF}_0 = 71/440 = 16$ per cent. Similarly, $\widehat{CF}_1 = 302/440 = 69$ per cent, and $\widehat{CF}_2 = 67/440 = 15$ per cent. The morbidity ratios standardized for age, sex and history of MI are obtained as follows (6, 8). For the referent, $SMR_0 = 1$ by definition. For those drinking

TABLE 1
Coffee-drinking and/or its related factors in the etiology of myocardial infarction (MI) (10)

| Stratum | | | Series | Level of coffee consumption (cups/day) | | | |
|------------------------------|-----|---------------|----------|--|-------|-------|--------|
| Age (years) | Sex | History of MI | | 0 | 1-5 | 6+ | Total |
| 40-49 | M | No | Cases | 8 | 36 | 11 | |
| | | | Controls | 308 | 911 | 287 | |
| 40-49 | M | Yes | Cases | 1 | 13 | 6 | |
| | | | Controls | 35 | 65 | 25 | |
| All strata | | | Cases | 71 | 302 | 67 | 440 |
| | | | Controls | 2,864 | 8,039 | 1,416 | 12,319 |
| Case fraction (%) | | | | 16 | 69 | 15 | |
| Standardized morbidity ratio | | | | (1.00) | 1.63 | 2.19 | |
| Etiologic fraction (%) | | | | (0) | 39 | 54 | 35 |

one to five cups per day and for those drinking at least six cups per day the estimates \overline{SMR}_1 and \overline{SMR}_2 are the ratios of the observed numbers of cases in these categories, 302 and 67, respectively, to the estimates of their respective "expectations." The latter are accumulated from the strata, as follows. The stratum of 40- to 49-year-old males without history of MI contributes $8(911)/308 = 23.7$ and $8(287)/308 = 7.5$ to the two estimates of "expectations," respectively. The next stratum gives $1(65)/35 = 1.9$ and $1(25)/35 = 0.7$, respectively. Then, $\overline{SMR}_1 = 302/(23.7 + 1.9 + \dots) = 302/185.1 = 1.63$, and $\overline{SMR}_2 = 67/(7.5 + 0.7 + \dots) = 67/30.6 = 2.19$. Thus, were one to consider the cases to be representative of those in the source, and were one to ignore possible sources of confounding other than age, sex and history of MI, one would estimate that among those drinking one to five and those drinking at least six cups of coffee per day the proportions of MI's attributable to coffee-drinking are $(1.63 - 1)/1.63 = 39$ per cent, and $(2.19 - 1)/2.19 = 54$ per cent, respectively (cf. formula 5). Moreover, recalling the case fractions and continuing to entertain the assumptions, it could be estimated that, overall, a fraction of $0.69(0.39) + 0.15(0.54) = 35$ per cent of all MI's in the source population are attributable to coffee-drinking (cf. formula 12). If additional confounding remains, these are estimates of the fractions of MI's related rather than attributable to coffee-drinking.

Interval estimation

The sampling variability of the above estimators poses a rather challenging problem. No results are available.

DISCUSSION

The proportion of a given disease attributable to a particular etiologic factor is obviously dependent not only on the risk of the disease among people with this factor

relative to those without it (RR_1) but also on the frequency (SF_1) of this factor in the population from which the cases arise. With the total number of cases proportional to $SF_0 + SF_1 \times RR_1$, the number attributable to the factor is proportional to $SF_1 \times (RR_1 - 1)$. Rewriting $SF_0 + SF_1 \times RR_1$ as $1 + SF_1 \times (RR_1 - 1)$ leads to the formula for etiologic fraction offered by Levin (1) and by Cole and MacMahon (2) (formula 2).

That the (natural) preoccupation with the frequency of the factor in the source population of cases may be replaced by its frequency among the cases themselves may not be immediately obvious. However, if among those in the source population who have the factor a fraction EF_1 of cases is attributable to the factor, then this fraction EF_1 of cases with the factor are attributable to the factor, while none of the cases without the factor are. This equivalence implies formulas 3 and 4.

As to the preventive fraction, intuition might suggest that this is nothing but a negative etiologic fraction, and that it could be thought of in terms of, say, formula 4 (with $RR_1 < 1$). However, the preventive fraction is the proportion of cases prevented by the factor among the totality of cases that would have developed in the absence of the protective factor, and it is thus necessary to replace the observable fraction of cases with the factor by that which it would have been in the hypothetical totality of cases (prevented and unprevented). Moreover, among those with the factor, the prevented fraction (formula 18) is not the absolute value for the etiologic fraction (formula 1).

In the generalization of these formulations to the usual case where confounding is to be allowed for and where the risk ratio varies over the categories of the confounding factor(s), it might be instructive to write the defining formulas for the etiologic and preventive fractions (formulas 6 and 19, respectively) as

$$EF = (SMR - 1)/SMR$$

and

$$PF = 1 - SMR,$$

respectively, where

$$\begin{aligned} SMR &= CR/ER \\ &= \frac{SF_0 + SF_1 \times CRR_1}{SF_0 + SF_1 \times CRR_1 / SMR_1} \\ &= 1/(CF_0 + CF_1/SMR_1). \end{aligned}$$

In these terms the overall etiologic and preventive fractions have structures similar to those of the corresponding parameters for those with the marker (formulas 5 and 18). However, this approach involves a certain generalization in the *SMR* concept: whereas it still is the "observed-to-expected" ratio, the referent is not necessarily confined to those exhibiting the risk marker at issue.

REFERENCES

1. Levin ML: The occurrence of lung cancer in man. *Acta Un Intern Cancer* 9:531-541, 1953
2. Cole PT, MacMahon B: Attributable risk percent in case-control studies. *Brit J Prev Soc Med* 25:242-244, 1971
3. Lilienfeld A: Epidemiology of infectious and non-infectious disease: some comparisons. *Am J Epidemiol* 97:135-147, 1973
4. Miettinen OS: Comprehensive monitoring of hospitalized patients in the evaluation of side-effects of oral contraceptives. Presented to the International Symposium on Statistical Problems in Population Research, the East-West Population Institute, Honolulu, August 1971
5. Panayotou PP, Kaskarelis DB, Miettinen OS, et al: Induced abortion and ectopic pregnancy. *Am J Obstet Gynecol* 114:507-510, 1972
6. Miettinen OS: Components of the crude risk ratio. *Am J Epidemiol* 96:168-172, 1972
7. Miettinen OS: Risk indicators for coronary heart disease. *Hart Bull* 4:64-70, 1973
8. Miettinen OS: Standardization of risk ratios. *Am J Epidemiol* 96:383-388, 1972
9. Miettinen OS: Estimation of relative risk from individually matched series. *Biometrics* 26:75-86, 1970
10. Jick H, Miettinen OS, Neff RL, et al: Coffee and myocardial infarction. *N Engl J Med* 289:63-67, 1973