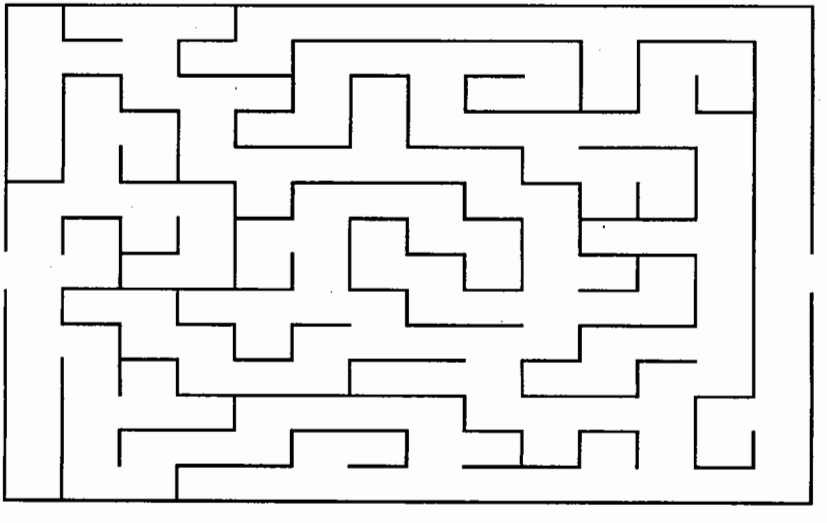
2

Comparisons



"A causes B." Scientists may mean that there is a regular temporal sequence of A then B, mediated by a physical mechanism that leads from one to the other. Even such a simple definition leaves room for ambiguity. A researcher who had studied the fates of insulation workers concluded that "All workers who have been heavily exposed to asbestos are certain to die of asbestos-related disease, unless they die of something else first." In striving to describe the effect of asbestos, the scientist had arrived at a tautology.

Estimates of "effect" in epidemiology are comparisons of measures such as those presented in the preceding chapter. Interpretation of the comparative measure as a statement about nature, however, demands further assumptions. The following three ideas provide a starting point.

Generalizability. *The physical processes that give rise to disease are largely similar in different individuals.*

Individuals determine groups. *Similar processes operating in different individuals give rise to predictable patterns of disease occurrence in groups.*[11]

Groups reflect individuals. *Patterns of disease occurrence in groups provide information about common physical processes operating within members of the group.*

With these ideas as their justification, epidemiologists derive mathematical expressions to summarize the differences between groups. The numerical relations among measured features of populations are the "measures of effect." They are a metaphor for disease processes within individuals. As with any metaphor, the image is powerful when it is precise and compact, and when it predicts that which was previously unseen.

Although there are infinitely many ways of joining even a pair of numbers, there are two that together have been used for almost all purposes of public health applications and etiologic research. These are *difference* measures and *ratio* measures.

Difference Measures

The *incidence rate difference* is a direct measure of the impact of an exposure.

Incidence rate difference. *The incidence rate difference is the difference between the incidence in an exposed population and that in an unexposed population.*

The incidence rate difference is also called simply the *rate difference.*[12] The term "incidence rate difference" is abbreviated as "*IRD*" or "*RD.*"

11. Students of infectious disease and of population dynamics will point out that there are simple processes whose long term behavior is so unpredictable as to call this assumption into question. Repeated iteration of a recursive formula for the prevalence of a disease as a function of the prevalence at an earlier date,

$$Pr_{i+1} = kPr_i(1-Pr_i)$$

for k greater than about 2.72, yields this sort of long-run instability. The literature of disease transmission is replete with intractable mathematics that will be ignored here.

12. Or even the "risk difference" or the "attributable risk," though the latter two terms are misnomers.

Table 2.1 gives annual incidence rates of death from coronary heart disease (CHD) in men in two age categories.[13] The incidence rate difference between the heavier smokers and the nonsmoking men aged 55-64 is 468 deaths per 100,000 man years. For men aged 65-74 years, the difference is 695 deaths per 100,000 man years. The incidence rate difference associated with smoking is greater in the older age group.

Table 2.1 Deaths from coronary heart disease per 100,000 man years

		Current smokers Cigarettes / day	
Age	Nonsmokers	10 - 20	21 - 39
55-64	501	798	969
65-74	1,015	1,501	1,710

Comparable figures for lung cancer mortality are shown in Table 2.2. The incidence rate differences for lung cancer deaths are 360 and 640 cases per 100,000 man years for men aged 55-64 and 65-74 years respectively for the heavier smokers. The relations of cigarette smoking to coronary heart disease mortality and to lung cancer mortality are roughly the same, in the sense that nearly equal numbers of deaths from each cause would be avoided if smokers could be given the same mortality rates as nonsmokers.

Table 2.2 Deaths from lung cancer per 100,000 man years

		Current smokers Cigarettes / day	
Age	Nonsmokers	10 - 20	21 - 39
55-64	40	250	400
65-74	80	500	720

13. U.S. Public Health Service. The Health Consequences of Smoking. A Reference Edition: 1976. U.S. Department of Health, Education, and Welfare, Public Health Service, Centers for Disease Control, HEW Publication No. (CDC) 78-8357, 1976, pp 657 ff

Incidence rate differences give a measure of the burden of disease for exposed individuals. Since smoking principally affects those who smoke, the health implications of smoking for the general public are related to the numbers of smokers and nonsmokers in the population. The general population incidence associated with smoking is the incidence rate difference that would be obtained by comparing the mixed population of smokers and nonsmokers, as it naturally occurs, with a population composed purely of nonsmokers. This difference is the *population rate difference* (*PRD*).[14]

Population rate difference. *The difference between an incidence rate in a population comprising both exposed and unexposed persons and the rate in a population comprising unexposed persons alone is the population rate difference.*

Algebraically,[15] the population rate difference can be shown to equal the incidence rate difference between exposed and unexposed, multiplied by the prevalence of the exposure in the population. Denote this last quantity by f_1, and let IR_1 and IR_0 be the incidence rates in exposed and unexposed, respectively. Then

$$PRD = f_1 (IR_1 - IR_0)$$

If forty percent of a male population aged 65-74 smoked 21-39 cigarettes a day, and nobody smoked more or less, then the *PRD* for coronary heart disease mortality would be calculable as follows:

$f_1 = 0.4$
$IR_0 = 1{,}015$

14. The PRD has been known as the "population attributable rate" and the "population attributable risk," both abbreviated PAR. The term "risk" is wrong in the present context. As discussed later, the word "attributable" is much stronger than warranted. Taken as technical jargon, the phrase was innocuous in an earlier time, but epidemiology so often enters into judicial and regulatory matters now that the misleading term needs to be corrected.

15. On the assumption that the members of a population have no individual risk that results from the smoking habits of their fellows, the incidence rate in the general population is the number of cases arising from exposed person time plus the number of cases arising from unexposed person time minus the number of cases that would have arisen had all the person time been subject to the rate of the unexposed, all divided by the total person time. Recognizing that the amount of exposed person time is f_1P and that the amount of unexposed person time is $(1-f_1)P$, the definition of PRD becomes

$$PRD = \frac{IR_1 f_1 P + IR_0(1-f_1)P - IR_0 P}{P}$$

which simplifies to the expression given.

$IR_1 = 1{,}710$
$PRD = 0.4 (1{,}710 - 1{,}015)$
$= 278$ cases per 100,000 man years

When there is more than one level of the exposure of interest, the defining equation is generalized to accommodate the prevalence of each exposure level, which is multiplied by the incidence rate difference appropriate to that level, and summed. Define f_i as the prevalence of exposure level i in the population, and let there be N exposure levels. Then[16]

$$PRD = \sum_{i=1}^{N} f_i(IR_i - IR_0)$$

If thirty percent of a male population aged 65-74 smoked 10-20 cigarettes per day, and ten percent smoked 21-39 per day, then the *PRD* for coronary heart disease mortality would be calculable as follows:

$f_1 = 0.3$
$f_2 = 0.1$
$IR_0 = 1{,}015$
$IR_1 = 1{,}501$
$IR_2 = 1{,}710$
$PRD = 0.3 (1{,}501 - 1{,}015) + 0.1 (1{,}710 - 1{,}015)$
$= 215$ cases per 100,000 man years

Since the *PRD* is a function of exposure prevalence as well as of the incidence of disease in exposed and unexposed, any presentation of the *PRD* should be accompanied by a specification of the population for which it was calculated. The *PRD* is zero when either the prevalence of exposure in the population is zero (i.e. when there is no exposure) or when the incidence rate difference is zero (when

16. The symbol immediately following the equality sign is a capital Greek letter sigma. It indicates that there should be a summation of terms whose form is given to the right of the sigma and that the terms should differ from one another by successive substitution of different values of one of the variables in the expression to the right. The i=1 below the sigma indicates that i is the variable to be successively changed, and that its first value should be 1. By convention, i should be successively increased by units of 1. The N above the sigma is the maximum value for i in the series; for unit increments in i, N is also the number of terms to be summed.

there is no effect of exposure). In the extreme case of all the population being exposed, the *PRD* equals the incidence rate difference.[17]

Ratio Measures

Cigarette smoking is widely believed to be more strongly associated with lung cancer than it is with heart disease. The basis for this impression lies with the second major way of comparing incidence rates. The ratios of the mortality rates of coronary heart disease in smokers of 21-39 cigarettes per day to those in nonsmokers are given in Table 2.3, together with the corresponding ratios for mortality from lung cancer. Cigarette smoking raises lung cancer and CHD mortality rates by about the same amount, but the increase in lung cancer is over a much lower baseline. The result is that the relative mortality is much higher for lung cancer than for CHD.

Table 2.3 Incidence rate ratios for deaths from lung cancer and coronary heart disease, comparing smokers (21-39 cigarettes/day) to nonsmokers

Age	CHD	Lung cancer
55-64	1.9	10
65-74	1.7	9

Incidence rate ratio. *The incidence rate ratio is the ratio of the incidence rate in an exposed population to that in an unexposed population.*

Incidence rate ratios (often called just *rate ratios*, and abbreviated as "*IRR*" or "*RR*") are heavily used in etiologic research. A large rate ratio is taken to indicate that the population characteristic being examined (here, smoking) is related to an important fraction of the disease in the exposed. The logic of this convention derives from the central role that the rate ratio plays in the answer to the question, "What fraction of the incidence rate (or cumulative incidence) in

17. The PRD could not be derived solely from the experience of a uniformly exposed population, since there would be no data on IR_0. It could nonetheless be estimated if there were an external estimate for IR_0, such as might be available from tables of vital statistics.

exposed persons could be eliminated, if the exposed could be given the same incidence rates as the unexposed?" The answer to this question is called the *relative excess incidence* (*REI*).[18]

Relative excess incidence. *The relative excess incidence is the fraction of the disease burden among exposed that would not have occurred if the exposed had experienced the same incidence rate as the unexposed.*

The *REI* can be calculated from incidence rates or from cumulative incidences. The defining equation for the *REI* (written here in terms of incidence rates) is

$$REI = \frac{IR_1 - IR_0}{IR_1} = \frac{RR - 1}{RR}$$

Since the incidence rate ratio for lung cancer presented in Table 2.3 is 9.0 among older men, the fraction of the lung cancer disease burden among older, heavier smokers that is associated with smoking is

$$REI = \frac{9.0 - 1}{9.0} = 0.89$$

By contrast, the relative excess incidence for CHD mortality is

$$REI = \frac{1.7 - 1}{1.7} = 0.41$$

Nearly 90 percent of the lung cancer mortality rate in heavy smokers is associated with smoking, whereas less than half of the CHD mortality is so associated.

If an exposure affects the time course of onset of disease, the *REI* calculated on the basis of incidence rates may vary with time since exposure. By contrast, the *REI* based on cumulative incidences presents a single summary figure that integrates the full time of observation, and is insensitive to changes in times of onset. Note

18. The relative excess incidence has been referred to as the "etiologic fraction" (abbreviated EF) in a number of textbooks. Like "attributable," "etiologic," when prepended to "fraction," is a technical term whose meaning differs both from the common meaning and from that which epidemiologists themselves generally intend when they discuss etiology. Presentation of an "etiologic fraction" does not imply any special standard of proof of causation.

that an *REI* calculated on the basis of cumulative incidence becomes arbitrarily small as the reference cumulative incidence approaches one, whereas the *REI* based on incidence rates is unaffected. The choice between incidence rate and cumulative incidence as the basis of the *REI* calculation depends on the perspective of the user of the data. As a rule of thumb, use cumulative incidences when the time period over which disease occurs is such that there is no useful distinction between early and late onsets. Many studies of infectious disease have this property, as may studies of fetal loss.

Mathematical Models of Incidence

Re-expressed as an equation or "model," the relation between the incidence rates in smokers and nonsmokers could be written as a function of the rate difference:

$$IR_1 = IR_0 + RD$$

or in terms of the rate ratio as

$$IR_1 = IR_0\, RR$$

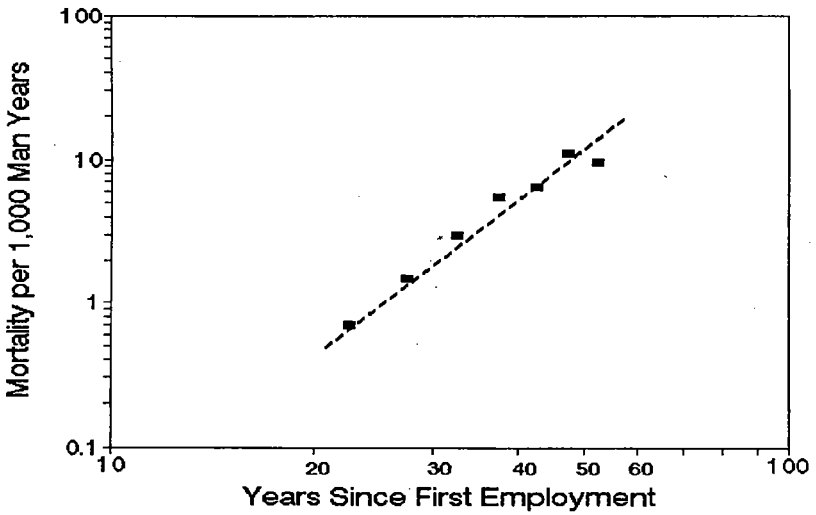


Figure 2.1 Annual mortality from mesothelioma among North American insulation workers

Often one scale is clearly superior for the summary of a complex body of data. In the examples given above, a single rate ratio could pretty well describe overall effects of heavy smoking on CHD mortality. The ratio was nearly the same in the two age strata presented ($RR \approx 1.8$), whereas no single rate difference could even approximately describe both strata.

Effect relations that are too complex to summarize in a table can sometimes be described easily with a model of incidence. The annual incidence of mesothelioma in North American insulation workers employed before 1960 appears to be well described at any time t by the relation[19]

$$IR(t) = 0.00437\ t^{3.2} \text{ cases / 100,000 man years}$$

where t is the elapsed time in years since an individual's first employment as an insulation worker. The relation is charted in Figure 2.1 on a log-log scale (that is with both axes scaled to the logarithm of the reported values).[20] Very similar relations have been seen in other populations occupationally exposed to asbestos, and in the resident population of Karain, Turkey, where fibrous minerals are omnipresent in exposed veins, in the dust on roads and in homes, and are used as the basic construction materials for all housing.[21] In Karain, t_0 is your day of birth.

Notice that there are two estimates of effect in the incidence model for mesothelioma among insulation workers. The first constant term (0.00437) gives the height of the curve; it reflects the intensity of asbestos exposure endured by the insulation workers. The second constant term (3.2) is an exponent to elapsed time, and presents a curvilinear "effect" of time on mesothelioma mortality.

19. Peto J, Seidman H, Selikoff IJ. Mesothelioma mortality in asbestos workers: Implications for models of carcinogenesis and risk assessment. Br J Cancer 1982;45:124-35

20. That Figure 2.1 should present a straight line follows immediately from the linear form of the incidence equation when logarithms are taken:

$$\ln(IR) = \ln(0.00437) + 3.2\ln(t)$$

The leading constant is an intercept. It gives the incidence at $\ln(t) = 0$, that is at $t =$ 1 year.

21. Saracci R, Simonato L, Baris Y, Artvinli M, Skidmore J. The age-mortality curve of endemic pleural mesothelioma in Karain, Central Turkey. Br J Cancer 1982;45:147-9

Group Comparisons and Individual Cause

Consider ten men who smoke cigarettes and among whom, apart from the effect of their workplace exposures, there would have been three lung cancer deaths at the ages of 55, 60, and 65.[22] Imagine that the effects of an inhaled workplace carcinogen were twofold: first, to induce changes in the composition of bronchial mucus so that particulate matter from cigarette smoke would be more effectively removed from the lungs; second and independently, to induce lung cancer through some direct effect on lung tissue. Imagine further that the net result of the opposing effects is a balance, so that among the ten workers there were, again, three deaths from lung cancer at ages 55, 60, and 65, but in different men. What is the effect of workplace exposure?

Clearly the measurable effect would be nil, unless it were possible to distinguish between those cases saved from a smoking-attributable death and those cases caused by the carcinogen. All of the exposed lung cancer deaths are attributable in a mechanistic sense to the occupational exposure. Three deaths were also prevented, so that the magnitude of the disease burden in the population is unchanged. Any epidemiologic measure of the relation of disease to exposure describes a net change in the population, and does not capture the effect of exposure on an individual level. No one can say on the basis of population data what fraction of exposed cases would have had their disease in the absence of exposure, nor what fraction of potential cases were averted by exposure.

The lack of a basis for individual causal inference is not limited to situations where the overall effect is absent, nor is it simply a matter of substitution of cases. As an alternative to the example above, imagine that three exposed men died of lung cancer at ages 50, 55, and 60. If the effect of exposure had been to change the age at death from 65 to 50 years in a single man, we might be tempted to say that a third of the lung cancer mortality had been affected by exposure. If the effect had been to shift each death forward by five years, then the timing of all the deaths would be attributable to

22. The example is reworked from Robins JM and Greenland S, Estimability and estimation of excess and etiologic fractions. Statist Med 1989; in press. See also Greenland S and Robins JM, Conceptual problems in the definition and interpretation of attributable fractions. Am J Epidemiol 1988;128:1185-97

exposure. The two situations, though quite different from the perspective of the men involved, are epidemiologically indistinguishable.

What then of the tenets with which the chapter opened? If epidemiology is to make any contribution to science, then generalizability and the reciprocal relation between individual and group phenomena must hold. There is no guarantee that they do hold, and so they must be tested, through replication of studies. If differences in disease frequency are the result of general processes that have been adequately tagged by the comparisons undertaken, then similar comparisons elsewhere should yield similar results. The modulations of exposure effect outlined above result from the superimposed effects of characteristics or exposures that bear only a coincidental relation to the exposure under study. In the first example, the apparent effect of workplace exposure would be very different in smoking and nonsmoking men. In the second, the apparent effect would be sensitive to the baseline disease rates and to the time periods at which men are observed in both scenarios.

When the measure of comparison, be it a difference or a ratio or one drawn from a statistical model, does not correspond to some essential feature of the disease generating process, then it is unlikely that the results of the same comparison undertaken in different populations will be the same. Neither ratio nor difference measures capture the acceleration of disease posed in the second example given above, and the epidemiologist will find it difficult to reproduce the "effect" under circumstances that vary only slightly in the timing of observation. Replication of results in different populations is not proof that a studied relation is valid, since it is possible to repeat mistakes. Nonetheless, replicability is a strong test of the hypothesis of generalizability, and relations that withstand the test are better substantiated than those that have not been subjected to it.