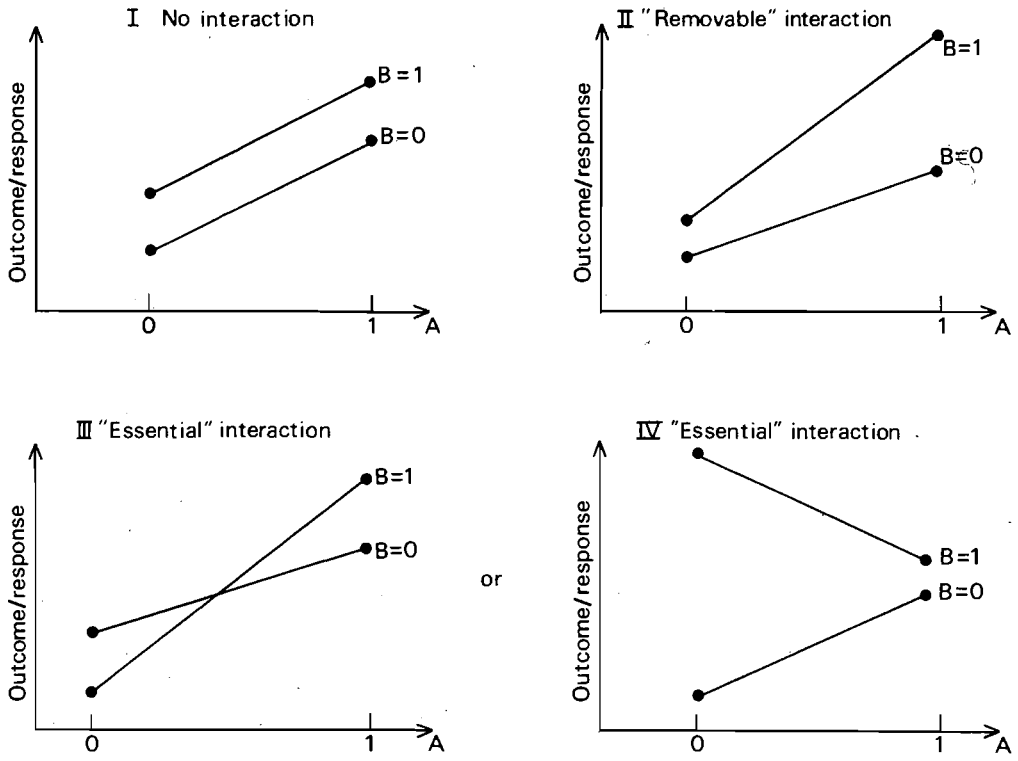
produced by the specific exposure were the same as that of the spontaneous cases, the differences in age-specific rates would be greater for the ages in which the spontaneous incidence was higher, even if the general and specific exposures had operated independently of each other early on. Nonetheless (2.7) may be postulated *ad hoc,* and if it appears to correspond reasonably well to the data, the estimate of b derived from the fitted model may be used as an overall measure of the effect of exposure.

In technical statistical terms, this model states that there are no *interactions* between the additive effects of exposure and strata on incidence rates; exposure to the risk factor has the same effect on disease incidence rates in each of the population strata. More generally, the absence of interactions between two factors, A and B, means that the effects of Factor B on outcome do not depend on the levels of Factor A. It is important to recognize, however, that what we mean by the effect of a factor depends very much on the scale of measurement. Since the rates are expressed on a simple arithmetic scale in (2.7), we speak of additive effects. As the following example shows, whether or not there are statistical interactions in the data may depend on the scale on which the outcome or response variable is measured.

Fig. 2.6 Schematic illustration of concept of statistical interaction.



Example: Figure 2.6 illustrates the concept of interaction schematically. Conditions for no interaction hold when the two response curves are parallel (Panel I). Note that the definition of interaction is completely symmetric; the diagram shows also that the effect of Factor A is independent of the level of Factor B.

The non-parallel response curves shown in Panel II of the figure indicate that Factor B has a greater effect on outcome at level 1 of Factor A than it does at level 0. It is apparent, however, that if the outcome variable were expressed on a different scale, for example a logarithmic or square root scale which tended to bring together the more extreme outcomes, the interaction could be made to disappear. In this sense we may speak of interactions which are "removable" by an appropriate choice of scale.

The situation in Panels III and IV, characterized by the response curves either crossing over or having slopes of different signs, allows for no such remedy. In Panel III the effect of Factor B is to increase the response at one level of Factor A, and to decrease it at another, while in Panel IV it is the sign of the A effect which changes with B. In the present context this would mean that exposure to the risk factor increased the rate of disease for one part of the population and decreased it for another. No change of the outcome scale could alter this essential difference.

While the excess risk is a useful measure in certain contexts, the bulk of this monograph deals with another measure of association, for reasons which will be clarified below. This is the *relative risk* of disease, defined as the *ratio* of the stratum-specific incidences:

$$r_i = \frac{\lambda_{1i}}{\lambda_{0i}}.$$

The assumed effect of exposure is to *multiply* the background rate λ_{0i} by the quantity r_i. Absence of interactions here leads to a *multiplicative model* for the rates such that, within the limits of statistical error, these may be expressed as the product of two terms, one representing the underlying natural disease incidence in the stratum and the other representing the relative risk r. More precisely, the model states

$$\lambda_{1i} = \exp(\beta)\lambda_{0i} \tag{2.8}$$

where $\beta = \log(r)$. Alternatively, if the incidence rates are expressed on a logarithmic scale, it takes the form

$$\log \lambda_{1i} = \log \lambda_{0i} + \beta.$$

Comparing this with equation (2.7) it is evident that they have precisely the same structure, except for the choice of scale for the outcome measure (incidence rate). In other words, the multiplicative model (2.8) is identical to an additive model in log rates. Such models are called *log-linear*.

While excess and relative risk are defined here in terms of differences and ratios of stratum-specific incidence rates, analogous measures for the comparison of cumulative rates and risks may be deduced directly from equations (2.2) and (2.4). Suppose, for example, that the two sets of incidence rates have a (constant) difference of 10 cases per 100 000 person-years observation for each year of a particular 15-year time period. Then the difference between the cumulative rates over this same period will be $10 \times 15 = 150$ cases per 100 000 population. On the other hand, if the two sets of rates have a (constant) ratio of 5 for each year, the ratio of the cumulative rates will also equal 5.

Because there is an exponential term in equation (2.4), the derived relationships between the probabilities, or risks, for this same time period are not so simple. Let $P_0(t)$ denote the net probability that a non-exposed person develops the disease during the time period from 0 to t years, and let $P_1(t)$ denote the analogous quantity for the exposed population. If the corresponding incidence rates satisfy the multiplicative equation $\lambda_1(u) = r\lambda_0(u)$ for all u between 0 and t, then

$$P_1(t) = 1-\{1-P_0(t)\}^r.$$

This relationship is well approximated by that for the cumulative rates

$$P_1(t) \approx rP_0(t),$$

providing the disease is sufficiently rare, or the time interval sufficiently short, so that both risks and rates remain small. In general, the ratio of disease risks is slightly less extreme, i.e., closer to unity, than is the ratio of the corresponding rates.

We have now introduced the two principal routes by which one may approach the statistical analysis of cancer incidence data: the additive model, where the fundamental measure of association is the excess risk, and the multiplicative model, where the effect of exposure is expressed in relative terms. In order to arrive at a choice between these two, or indeed to decide upon any particular statistical model, several considerations are relevant. From a purely empirical viewpoint, the most important properties of a model are simplicity and goodness of fit to the observed data. The aim is to be able to describe the main features of the data as succinctly as possible. Clarity is enhanced by avoiding models with a large number of parameters which must be estimated from the data. If, in one type of model many interaction terms (see § 6.1) are required to fit the data adequately, whereas with another only a few are required, the latter would generally be preferred.

The empirical properties of a model are not the only criteria. We also need to consider how the results of an analysis are to be interpreted and the meaning that will be attached to the estimated parameters. Excess and relative risks inform us about two quite different aspects of the association between risk factor and disease. Since relative risks for lung cancer among smokers *versus* non-smokers are generally at least five times those for coronary heart disease, one might be inclined to say that the lung cancer-smoking association is stronger, but this ignores the fact that the differences in rates are generally greater for heart disease. From a public health viewpoint the impact of smoking on mortality from heart disease may be more severe than its effect on lung cancer death rates. This fact has led some authors to advocate exclusive use of the additive measure (Berkson, 1958). Rothman (1976), as noted earlier, has argued that it is the most natural one for measuring interaction.

In spite of these considerations, the relative risk has become the most frequently used measure for associating exposure with disease occurrence in cancer epidemiology, both because of its empirical behaviour and because of several logical properties it possesses. Empirically it provides a summary measure which often requires little qualification in terms of the population to which it refers. Logically it facilitates the evaluation of the extent to which an observed association is causal. The next two sections

explore these important properties of the relative risk in some detail. We merely point out here that, once having obtained an estimate of the relative risk, it is certainly possible to interpret that estimate in terms of excess risk provided one knows the disease incidence rates for unexposed individuals in the population to which it refers. For example, if the baseline disease incidence is 20 cases per year per 100 000 population and the relative risk is 9, this implies that the difference in rates between the exposed and unexposed is $(9-1)\times 20 = 160$ cases per 100 000. In our opinion, the advantages of using the relative measure in the analysis far outweigh the disadvantage of having to perform this final step to acquire a measure of additive effect, if in fact that is what is wanted. No measure of association should be viewed blindly, but instead each should be interpreted using whatever information exists about the actual magnitude of the rates.

2.5 Empirical behaviour of the relative risk

Several examples from the literature of cancer epidemiology will illustrate that the relative risk provides a stable measure of association in a wide variety of human populations. When there are differences in the (multiplicative) effect of exposure for different populations, it is often true that the levels of exposure are not the same, or that there are definite biological reasons for the discrepancies in the response to the same exposure.

Temporal variation in age-specific incidence

Table 2.5 shows the age-specific incidence rates for breast cancer in Iceland for two of the birth cohorts represented in Figure 2.4. The ratios of these rates for the two cohorts are remarkably stable in the range 1.66–1.81, whereas the differences between them triple over the 50-year age span. Thus, while one can describe the relationship between birth cohort and incidence by saying that the age-specific rates for the later cohort are roughly 1.7 times those for the earlier one, no such simple summary is possible using the excess risk as a measure of association. Note that the ratio of the cumulative rates summarizes that for the age-specific ones, and that the cumulative risk ratio is only slightly less than the rate ratio despite the 50-year age span.

Table 2.5 Average annual incidence rates for breast cancer in Iceland, 1910–72, per 100 000 population[a]

Year of birth	Age (years) 40–49	50–59	60–69	70–79	80–89	Cumulative (ages 40–89) Rate (%)	Risk (%)
1880–1909	65.90	95.10	129.50	140.10	227.90	6.59	6.38
1840–1879	38.70	53.80	71.70	81.10	136.90	3.82	3.75
Difference	27.20	41.30	57.80	59.00	91.00	2.77	2.63
Ratio	1.70	1.78	1.81	1.73	1.66	1.73	1.70

[a] From Bjarnasson et al. (1974)

Geographical variation in age-specific incidence

Figure 2.7 gives a plot of incidence rates against age for stomach cancer occurring in males in three countries (Waterhouse et al., 1976). In calculating these rates, six 5-year age intervals were used: 35–39, 40–44, 45–49, 50–54, 55–59, 60–64. Since a logarithmic scale is used for both axes, the plotted points appear to lie roughly on three parallel straight lines, each with a slope of about 5 or 6. This quantitative relationship, which is common for many epithelial tumours, may be expressed symbolically as follows. Denote by $\lambda_i(t)$ the average annual incidence rate for the i^{th} area at age t, where t is taken to be the midpoint of the respective age interval: t = 37.5, 42.5, etc. The fact that the log-log plots are parallel and linear means that approximately

$$\log \lambda_i(t) = \alpha + \beta_i + \gamma \log(t), \tag{2.9}$$

where we arbitrarily set $\beta_1 = 0$, thus using country 1 as a baseline for comparison. Raising each side of this equation to the power e, the relationship may also be expressed as

$$\lambda_i(t) = e^{\alpha} r_i t^{\gamma}, \tag{2.10}$$

where $r_i = \exp(\beta_i)$.

The values of the parameters in (2.9) which give the best "fit" to the observed data points, using a statistical technique known as 'weighted least squares regression' (Mosteller & Tukey, 1977, p. 346), are $\alpha = -18.79$, $\beta_1 = 0$, $\beta_2 = 0.67$, $\beta_3 = 1.99$ and $\gamma = 5.49$. Although the deviations of the plotted points about the fitted regression lines are slightly larger than would be expected from purely random fluctuations, the equations well describe the important features of the data.

The parameters r (= exp β) describe the relative positions of the age-incidence curves for the three countries. By considering ratios of incidence rates, the relative risk of stomach cancer in males in Japan *versus* those in Connecticut is

$$\frac{\lambda_3(t)}{\lambda_1(t)} = \frac{r_3 t^{\gamma}}{r_1 t^{\gamma}} = \exp(\beta_3 - \beta_1) = 7.3$$

while the relative risk in Birmingham *versus* that in Connecticut is

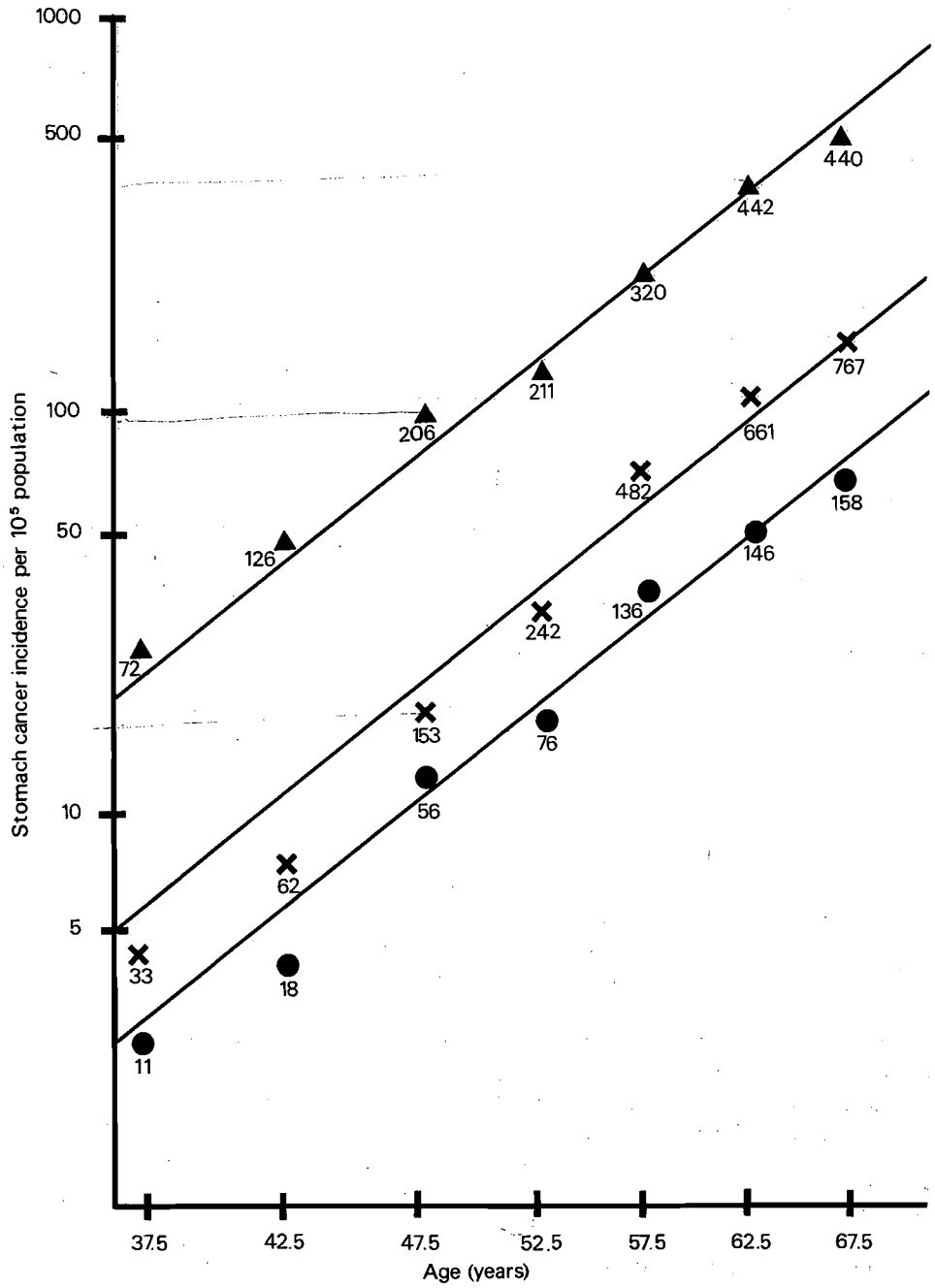
$$\exp(\beta_2 - \beta_1) = 1.9.$$

The most important feature of the above relationships is that, to the extent that equations (2.9) or (2.10) hold, the relative risks between different areas *do not vary with age.* The chance that a Birmingham male of a given age contract stomach cancer during the next year is roughly twice that of his New England counterpart, and the same applies whether he is 45, 55 or 65 years old. On the other hand, the absolute differences in the age-specific rates, i.e., $\lambda_2(t) - \lambda_1(t)$, vary markedly with age. The percentage increase in incidence associated with each 10% increase in age is related to the parameter γ through the equation

$$\left(\frac{\lambda_i(1.1t)}{\lambda_i(t)} - 1\right) \times 100\% = \left((1.1)^{\gamma} - 1\right) \times 100\% = \left((1.1)^{5.49} - 1\right) \times 100\% = 69\%,$$

and varies neither with age nor with area.

Fig. 2.7 Age-specific incidence of stomach cancer in three populations. From Waterhouse et al. (1976). Number of cases shown by each point. (▲ = Japan (Miyagi); × = UK (Birmingham); ● = US (Connecticut).)



As shown by Cook, Doll & Fellingham (1969), most epithelial tumours have age-incidence curves of a similar shape to that of gastric cancer, differing between populations only by a proportionality constant, i.e., relative risk. This is a good technical reason for choosing the ratio as a measure of association, since it permits the relationship between each pair of age-incidence curves to be quite accurately summarized in a single number.

The two epithelial tumours which deviate most markedly from this pattern are those of the lung and the breast. For breast cancer we have already shown how irregularities in the cross-sectional age curves reflect a changing incidence by year of birth, and that a basic regular behaviour is seen when the data are considered on a cohort basis (Figures 2.3 and 2.4; Bjarnasson et al., 1974). A similar phenomenon has been noted for lung cancer, where a large part of the inter-cohort differences are presumably due to increasing exposure to tobacco and other exogenous agents (Doll, 1971).

Risk of cancer following irradiation

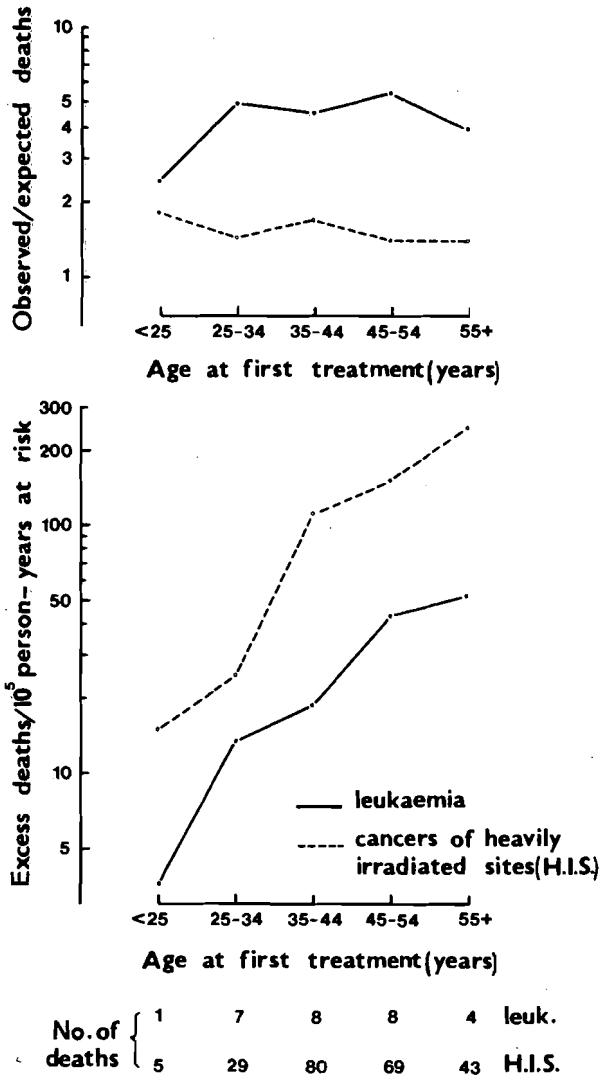
Radiation induces tumours at a wide range of sites, and its carcinogenic effects have been studied in a variety of population groups, including the atomic bomb survivors in Japan and people treated by irradiation for various conditions. As discussed in the previous example, the "natural" incidence of most cancers varies widely with age at diagnosis. Here we examine how the carcinogenic effect of radiation varies according to *age at exposure,* i.e., the age of the individual when irradiated.

In the mid 1950s, Court Brown and Doll (1965) identified over 14 000 individuals who had been treated by irradiation for ankylosing spondylitis between 1935 and 1954 in the United Kingdom. The latest report analyses the mortality of this group up to 1 June 1970 (Smith, 1979). In Figure 2.8 we show the change with age at exposure of the relative risk and of the absolute risks for leukaemia and for other heavily irradiated sites. For both types of malignancy, the relative risk varies little with age at exposure, whereas the absolute risk increases rapidly as age at treatment increases. The effect of the radiation is thus to multiply the incidence which would be expected among people in the general population of the same age by a factor of roughly 4.8 for leukaemia and 1.5 for other heavily irradiated sites. As a function of *time since exposure,* the relative risk for leukaemia appears to reach a peak after 3–5 years and then decline to zero, whereas the effect on heavily irradiated sites may persist for 20 or more years after exposure.

An analysis of the mortality among atomic bomb survivors for the period 1950–74 (Beebe, Kato & Land, 1977) demonstrates a similar uniformity of relative risk with age at exposure, and the corresponding sharp increase in absolute risk. There is, however, one major exception to the uniformity of the relative risk. For those aged less than ten years at exposure the relative risks are considerably higher than in subsequent age groups, which presumably indicates greater susceptibility among young children.

Studies of breast cancer induced by radiation include those of atomic bomb survivors (MacGregor et al., 1977) and of women treated by irradiation for tuberculosis (Boice & Monson, 1977) or a range of benign breast conditions (Shore et al., 1977). The relative risk appears higher among women exposed at younger ages and is particularly high among those exposed in the two years preceding menarche or during their first

Fig. 2.8 Ratio of observed to expected numbers of deaths and excess death rates from leukaemia and cancers of heavily irradiated sites according to age at first treatment with X-rays for ankylosing spondylitis. From Smith (1979).



pregnancy (Boice & Stone, 1979). The proliferation of breast tissue during menarche or first pregnancy would suggest an increased susceptibility to carcinogenic hazards.

The relative risk thus seems to provide a fairly uniform measure of the carcinogenic effect of radiation as a function of age at exposure, except where a difference in the relative risk probably reflects differences in tissue susceptibility.

Lung cancer and cigarette smoking

Smoking and irradiation are perhaps the most extensively studied of all carcinogenic exposures. Cigarette smoking is related to tumours at a number of sites including the respiratory tract, the oral cavity and oesophagus, and the bladder and pancreas. The relationship with cancer of the lung has been the most extensively studied, and the results of several large prospective studies have quantified the association in some detail.

Table 2.6 presents the change in incidence with age among continuing smokers and among non-smokers, as given by Doll (1971), the data for consecutive five-year age groups being averaged. The excess risk increases sharply with age, whereas the relative risk, although increasing, changes only slowly.

Table 2.6 Incidence of bronchial carcinoma among non-smokers and continuing smokers, per 100 000 person-years[a]

Age at risk (years)	Non-smokers	Smokers	Relative risk	Excess risk
35–44	2.8	5.2	1.9	2.4
45–54	5.8	67.0	11.6	61.2
55–64	13.9	221.8	16.0	207.9
65–74	25.6	482.2	18.8	456.6
75–84	49.4	860.5[b]	17.4	811.1

[a] From Doll (1971)
[b] Likely to be unreliable due to under-reporting

A more appropriate way of looking at the risk of lung cancer associated with cigarette smoking, however, is in terms of duration of smoking rather than simply age. Figure 2.9 presents the incidence of lung cancer for non-smokers as a function of age, and for smokers as a function of both age and duration of smoking. The increase in relative risk with age is clear, but more striking is the parallellism of the lines for non-smokers and for smokers when incidence is related to duration of smoking. Since for non-smokers we might regard exposure as lifelong, one could consider that the two time scales both refer to duration of exposure. The figure thus displays a constant relative difference in incidence when the more relevant time-scales are used.

Breast cancer and age at first birth

The large international study by MacMahon and associates (MacMahon et al., 1970) showed that age at first birth is the major feature of a woman's reproductive life which influences risk for breast cancer. Table 2.7, taken from their work, shows the uniformity of the relationship between risk and age at first birth over all centres in a collaborative study. Furthermore (not shown in the table), these relative risks change little with age at diagnosis. The populations included in the study showed a wide range of incidence levels, and had age-incidence curves of quite different shapes. The ability of the

Fig. 2.9 Age-specific mortality rates from lung cancer for smokers and non-smokers. From Doll (1971). (● —— ● = cigarette smokers by duration of smoking; ○——○ = cigarette smokers by age; ×——× = non-smokers by age.)

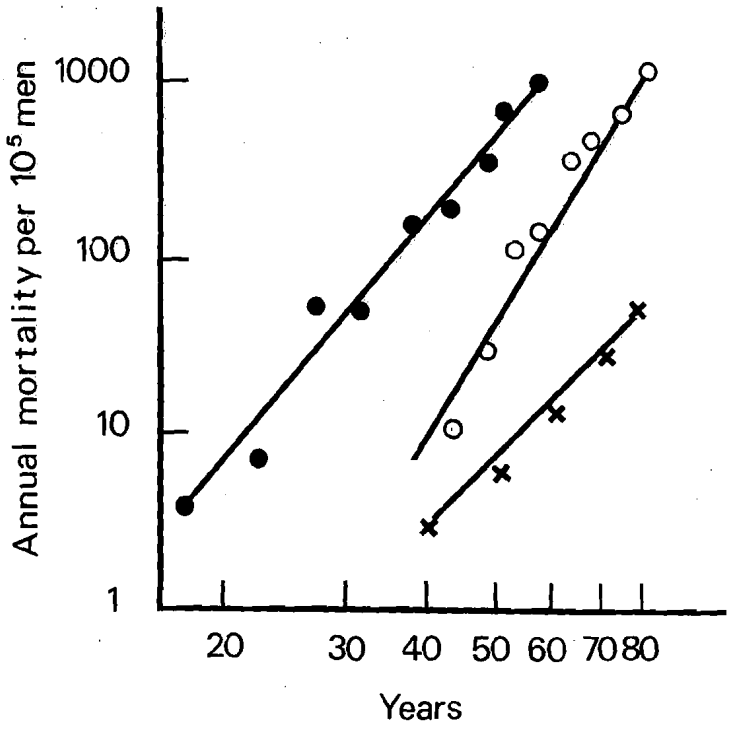


Table 2.7 Estimates of relative risk of breast cancer, by age at first birth[a, b]

Centre	Nulliparous	Parous, age at first birth (years): <20	20–24	25–29	30–34	35+
Boston	100	32	55	76	90	117
Glamorgan	100	38	49	67	73	124
Athens	100	51	71	79	106	127
Slovenia	100	81	74	94	112	118
Sao Paulo	100	49	65	94	84	175
Taipei	100	54	45	37	89	106
Tokyo	100	26	49	78	100	138
All centres	100	50	60	78	94	122

[a] From MacMahon et al. (1970)
[b] Estimated risk relative to a risk of 100 for the nulliparous; adjusted for age at diagnosis

relative risk to summarize the relationships among so wide an array of incidence patterns indicates that, at least in this situation, it reflects a fundamental feature of the disease. The absolute differences in age-specific incidence rates by age at first birth vary widely between the populations.

The failure of previous work on the influence of reproductive factors on risk of breast cancer to identify the basic importance of age at first birth was probably due to inappropriate measures of disease association. As MacMahon et al. concluded, "Previous workers seem not to have considered the differences of sufficient importance to warrant detailed exploration. An apparent lack of interest in the relationship may have resulted from failure to realize the magnitude of the differences in relative risk that underlie it. This lack of recognition of the strength of the relationship can be attributed primarily to analyses using summary statistics such as means ...".

2.6 Effects of combined exposures

The previous examples have illustrated the extent to which the relative risk remains constant over different age strata, or among different population groups. We shall now examine the extent to which the relative risk associated with one risk factor varies with changing exposure to a second risk factor, and we shall see that in this situation one also frequently observes relative uniformity. Consider the simplest situation, with two dichotomous variables A and B. There are four incidence rates, denoted λ_{AB}, λ_A, λ_B and λ_0 according to whether an individual is exposed to both, one or neither of the factors. The three relative risks, expressed using λ_0 as the baseline incidence, are $r_{AB} = \lambda_{AB}/\lambda_0$, $r_A = \lambda_A/\lambda_0$ and $r_B = \lambda_B/\lambda_0$, respectively.

Among those exposed to B, the relative increase in risk incurred by also being exposed to A is given by $\lambda_{AB}/\lambda_B = r_{AB}/r_B$. If the relative risk associated with exposure to A is the same, whether or not there is exposure to B, we say that the effects of the two factors are independent or do not interact (Figure 2.6). In this case $r_{AB}/r_B = r_A$, from which $r_{AB} = r_A r_B$. Thus, the independence of relative risks for two or more exposures implies a multiplicative combination for the joint effect. But, if the two risk factors each have additive rather than multiplicative effects on incidence, then similar calculations show that the relative risk for the joint exposure under the no interaction assumption is $r_{AB} = r_A + r_B - 1$.

The uniformity of relative risk for the exposures considered in the earlier examples can also be interpreted as a multiplicative combination of effects. Since the spontaneous incidence of leukaemia increases with age and radiation affects the spontaneous incidence proportionately, the joint effect is simply the product of the spontaneous rate and the radiation risk. Women in the United States have an incidence of breast cancer about six times higher than that of Japanese women. The joint action of the factor responsible for the elevated risk among United States women, whatever it may be, and age at first birth is clearly multiplicative.

Example: As an example of the joint effects of two risk factors, Table 2.8 summarizes results from a case-control study of oral cancer as related to alcohol and tobacco consumption (Rothman & Keller, 1972). The 483 cases and 492 controls were cross-classified according to four levels of consumption of each risk factor and also two age categories, under and over 60 years of age. Using methods which will be introduced in Chapter 4, age-adjusted relative risks of oral cancer were calculated for each of the 16

Table 2.8 Joint effect of alcohol and tobacco consumption on risk for oral cancer[a, b]

Alcohol (oz/day)	Tobacco (cigarette equiv./day) 0	1–19	20–39	40+	Alcohol risk (adjusted for tobacco)
0	1.0	1.6	1.6	3.4	1.0
0.1–0.3	1.7	1.9	3.3	3.4	1.8
0.4–1.5	1.9	4.9	4.9	8.2	2.9
1.6+	2.3	4.8	10.0	15.6	4.2
Tobacco risk (adjusted for alcohol)	1.0	1.4	2.4	4.2	

[a] From Rothman and Keller (1972)
[b] Relative risks adjusted for age at diagnosis

alcohol/tobacco categories shown. These may be denoted r_{ij}, where i refers to tobacco level and j to alcohol level. Since the category of lowest exposure to both factors is used as a baseline for comparison with other groups, $r_{11} = 1.0$.

The multiplicative hypothesis in this framework takes the form

$$r_{ij} = r_{i1}r_{1j}, \tag{2.11}$$

whereby the relative risk for a given category of tobacco/alcohol consumption is obtained as the product of a relative risk for the tobacco level times that for the alcohol level. Again, this expresses the idea that relative risks for different tobacco levels do not vary according to alcohol consumption, and vice versa. Of course the r_{ij} presented in Table 2.8 do not satisfy this requirement exactly. Procedures are presented in Chapter 6 for finding estimates of r_{i1} and r_{1j} which yield the *best fit* to the observed data under the model. These estimates, shown in the margins of Table 2.8, were used to calculate the expected number of cases in Table 2.9. Comparison of the observed numbers of cases with those expected under the model shows that agreement between the model and the data is about as good as can be expected, given the errors inherent in random sampling.

Table 2.9 Observed number of cases and controls by smoking and drinking category, and the number expected under the multiplicative model[a]

Alcohol (oz/day)	Tobacco. (cigarette equiv./day) 0 Cases	Controls	Expected cases	1–19 Cases	Controls	Expected cases	20–39 Cases	Controls	Expected cases	40+ Cases	Controls	Expected cases
0	10	38	7.67	11	26	9.91	13	36	17.54	9	8	7.87
0.1–0.3	7	27	7.36	16	35	16.34	50	60	47.08	16	19	18.21
0.4–1.5	4	12	5.14	18	16	15.64	60	49	61.37	27	14	26.86
1.6+	5	8	5.82	21	20	24.11	125	52	122.00	91	27	90.06

[a] From Rothman and Keller (1972)

The multiplicative effects of alcohol and tobacco have been demonstrated by Wynder and Bross (1961) for cancer of the oesophagus, and for cancer of the mouth in an earlier publication (Wynder, Bross & Feldman, 1957).

Example: A second example concerns the joint effect of asbestos exposure and cigarette smoking on risk for bronchogenic carcinoma. Selikoff and Hammond (1978) followed 17 800 asbestos insulation workers prospectively from 1 January 1967 to 1 January 1977. Smoking histories were obtained for the

majority of the cohort. Risk estimates for smoking obtained from the American Cancer Society prospective study (Hammond, 1966) were applied to generate expected numbers of deaths from lung cancer among the insulation workers. Table 2.10 gives the observed and expected numbers of lung cancer deaths among continuing smokers and among non-smokers.

Since the asbestos-related risks in the two groups are about equal, it follows that the risk for cigarette smoking asbestos insulation workers, compared with non-smokers not exposed to asbestos, is the product of their smoking risk, from which the expected numbers were derived, and their asbestos risk. Similar results have been reported by Berry, Newhouse and Turok (1972) and reviewed by Saracci (1977).

Table 2.10 The joint effect of cigarette smoking and asbestos exposure on risk for lung cancer. Lung cancer mortality among 17 800 asbestos insulation workers, 1967–77[a]

	Lung cancer deaths		
	Observed	Expected[b]	Relative risk
Non-smokers	8	1.82	4.40
Smokers	228	39.7	5.74

[a] From Hammond, Selikoff and Seidman (1979)
[b] Based on age-specific general population rates for men smoking equivalent numbers of cigarettes

The epidemiology of cancer thus provides empirical reasons for choosing relative risk as the natural measure of association of cancer and exposure. On many occasions similar exposures lead to similar relative risks, almost independent of the population group exposed. When appreciable differences in relative risk are observed, these often can be expected to reflect real differences in susceptibility or exposure which may not be immediately apparent. As an interesting contrast, Table 2.11 gives data for ischaemic heart disease (Doll & Peto, 1976), where the biological processes are presumably different. The relative risks change markedly with age, and a different measure of association might be more appropriate.

Table 2.11 Smoking and risk for ischaemic heart disease, by age[a]

Age (years)	Annual death rate per 100 000 men[b] (no. of deaths in parentheses)											
	Non-smokers		RR	Current smokers, smoking cigarettes only (no./day)								
				1–14		RR	15–24		RR	25+		RR
< 45	7	(3)	1.0	46	(12)	6.6	16	(22)	2.3	104	(18)	14.9
45–54	118	(32)	1.0	220	(38)	1.9	368	(90)	3.1	383	(69)	3.3
55–64	531	(79)	1.0	742	(91)	1.4	819	(123)	1.5	1 025	(125)	1.9
< 65	166	(114)	1.0	278	(141)	1.7	358	(235)	2.2	427	(212)	2.6
65–74	1 190	(83)	1.0	1 866	(134)	1.6	1 511	(101)	1.3	1 731	(81)	1.5
75+	2 432	(92)	1.0	2 719	(113)	1.1	2 466	(50)	1.0	3 247	(27)	1.3

[a] From Doll and Peto (1976)
[b] Indirectly standardized for age to make the four entries in any one line comparable

2.7 Logical properties of the relative risk

In addition to an empirical justification for its use, the relative risk has some properties of a logical nature which are useful for appraising the extent to which the observed association may be explained by the presence of another agent, or may be specific to a particular disease entity. Cornfield et al. (1959) gave a precise statement and formal proof of these properties (see also § 2.9).

> "If an agent, A, with no causal effect upon the risk of disease, nevertheless, because of a positive correlation with some other causal agent, B, shows an apparent risk, r, for those exposed to A, relative to those not so exposed, then the prevalence of B, among those exposed to A, relative to the prevalence among those not so exposed, must be greater than r."

Thus, in order that the smoking-lung cancer association be explained by a tendency for people with a cancer-causing genotype to smoke, the putative genetic trait must carry a risk of at least ninefold in addition to being at least nine times more prevalent among smokers. Spurious associations due to confounding are always weaker than the underlying genuine associations when strength of association is measured by relative risk.

Cornfield et al. also note that the relative measure is a sensitive indicator of the *specificity* of the association with a particular disease entity:

> "If a causal agent A increases the risk for disease I and has no effect on the risk for disease II, then the relative risk of developing disease I, alone, is greater than the relative risk of developing disease I and II combined, while the absolute measure is unaffected."

Thus, if the agent in question increases the risk of a certain histological type of cancer at a given site (e.g., "epidermoid" as opposed to other types of lung cancer) but has little or no effect on other types, a greater relative risk is obtained when the calculation is restricted to the particular histological type than when all cancers at that site are considered. But, it makes no difference to the excess risk if the other histological types are included or not.

Finally, from the point of view of case-control studies, there is one compelling reason for adopting the relative risk as the primary measure of association even in the absence of other considerations. This is simply that, as shown in the next section, the *relative risk is in principle directly estimable from data collected in a case-control study.* Additional information, namely knowledge of actual incidence rates for at least one of the exposed or non-exposed populations, is required to estimate the excess risk.

2.8 Estimation of the relative risk from case-control studies – basic concepts

A full understanding of how the data from a case-control study permit estimation of the relative risk requires careful description of how cases and controls are sampled from the population. The studies whose analysis is considered in this monograph involve the ascertainment of new (incident) cases which occur in a defined study period. Ideally these cases are identified through a cancer registry or some other system which

covers a well-defined population; with hospital-based studies the referent population, consisting of all those "served" by the given hospital, may be more imaginary than real. Most commonly the sample will contain all new cases arising during the study period, or at least all those successfully interviewed. Otherwise they are assumed to be a random sample of the actual cases.

The controls in a case-control study are assumed to represent a random sample of the subjects who are disease-free, though otherwise at risk. The control sample may be stratified, for example on the basis of age and sex, so that it has roughly the same age and sex distribution as the cases. Or, the controls may be individually matched to cases on the basis of family membership, residence or other characteristics. Under such circumstances the controls are assumed to constitute a random sample from within each of the subpopulations formed by the stratification or matching factors.

If infinite resources were available, one would ideally conduct a prospective investigation of the entire population. Subjects would be classified at the beginning of the study period on the basis of exposure to the risk factor, and at the end of the period according to whether or not they had developed the disease. Suppose that a proportion p of the individuals at risk in a particular stratum were exposed at the beginning of the study. Denote by $P_1 = P_1(t)$ the probability that an exposed person in this stratum develops the disease during a study period of length t, and by $P_0 = P_0(t)$ the analogous quantity for the unexposed. Let $Q = 1-P$ and $q = 1-p$. Then the *expected* proportions of individuals who fall into each of the resulting four categories or cells may be represented thus:

	Exposed	Unexposed	Total
Diseased	pP_1	qP_0	$pP_1 + qP_0$
Disease-free	pQ_1	qQ_0	$pQ_1 + qQ_0$
Total	p	q	1

(2.12)

If the study period is reasonably short, which means of the order of a year or two for most cancers and other chronic disease, the probabilities P_1 and P_0 will be quite small. According to § 2.4, their ratio will thus be a good approximation to the ratio r of stratum-specific incidence rates averaged over the study period. In other words, we have as an approximation $r = \lambda_1/\lambda_0 \approx P_1/P_0$. Since $Q_1 \approx Q_0 \approx 1$ under these same circumstances, it follows that $P_1/Q_1 \approx P_1$ and $P_0/Q_0 \approx P_0$, and thus that the relative risk is also well approximated by the *odds ratio* ψ of the disease probabilities:

$$\psi = \frac{P_1 Q_0}{P_0 Q_1} \approx \frac{P_1}{P_0} \approx r. \tag{2.13}$$

The term "odds ratio" derives from the fact that ψ may also be written in the form $(P_1/Q_1) \div (P_0/Q_0)$, i.e., as the ratio of the "odds" of disease occurrence in the exposed and non-exposed sub-groups.

Example: Suppose the average annual incidence rates for the exposed and non-exposed substrata are $\lambda_1 = 0.02$ and $\lambda_0 = 0.01$ and that the study lasts three years. Then the cumulative rates are $\Lambda_1 = 0.06$ and $\Lambda_0 = 0.03$, while the corresponding risks (2.4) are $P_1 = 1 - \exp(-0.06) = 0.05824$ and $P_0 = 1 - \exp(-0.03) = 0.02956$. It follows that the odds ratio is

$$\psi = \frac{0.05824 \times 0.97044}{0.02956 \times 0.94176} = 2.03,$$

as compared with a relative risk $r = \lambda_1/\lambda_0$ of exactly 2.

As Cornfield (1951) observed, the approximation (2.13) provides the critical link between prospective and retrospective (case-control) studies *vis-à-vis* estimation of the relative risk. If the entire population were kept under observation for the duration of the study, separate estimates would be available for each of the quantities p, P_1 and P_0, so that one could determine all the probabilities shown in (2.12). If we were to take samples of exposed and unexposed individuals at the beginning of the study and follow them up, this would permit estimation of P_1 and P_0 and thus of both excess and relative risks, but not of p; of course such samples would have to be rather large in order to permit sufficient cases to be observed to obtain good estimates. With the case-control approach, on the other hand, sampling is done according to disease rather than exposure status. This ensures that a reasonably large number of diseased persons will be included in the study. From such samples of cases and controls one may estimate the exposure probabilities given disease status, namely:

$$p_1 = \text{pr(exposed}\,|\,\text{case)} = \frac{pP_1}{pP_1 + qP_0} \quad \text{and}$$

$$p_0 = \text{pr(exposed}\,|\,\text{control)} = \frac{pQ_1}{pQ_1 + qQ_0}.$$

It immediately follows that the odds ratio calculated from the exposure probabilities is identical to the odds ratio of the disease probabilities, or in symbols:

$$\psi = \frac{p_1 q_0}{p_0 q_1} = \frac{P_1 Q_0}{P_0 Q_1}. \tag{2.14}$$

Consequently the ratio of disease incidences, as approximated by the odds ratio of the corresponding risks, can be directly estimated from a case-control study even though the latter provides no information about the absolute magnitude of the incidence rates in the exposed and non-exposed subgroups.

Example: As an illustration of this phenomenon, suppose the incidence rates from the previous example applied to a population of 10 000 persons, of whom 30% were exposed to the risk factor. If the entire population were kept under observation for the study period one would expect to find $P_1 \times 3\,000 = 175$ exposed cases and $P_0 \times 7\,000 = 207$ non-exposed cases. The data could thus be summarized:

	Exposed	Unexposed	Total
Diseased	175	207	382
Disease-free	2 825	6 793	9 618
Total	3 000	7 000	10 000