CHAPTER II

FUNDAMENTAL MEASURES OF DISEASE OCCURRENCE AND ASSOCIATION

The occurrence of particular cancers varies remarkably according to a wide range of factors, including age, sex, calendar time, geography and ethnicity. Etiological studies attempt to explain such variation by relating disease occurrence to genetic markers, or to exposure to particular environmental agents, which may have a similar variation in time and space. The cancer epidemiologist studies how the disease depends on the constellation of risk factors acting on the population and uses this information to determine the best measures for prevention and control. This process requires a quantitative measure of exposure, as well as one of disease occurrence, and some method of associating the two.

In this chapter we introduce the fundamental concepts of disease incidence rates, cumulative incidence, and risk. These will allow us to make a precise comparison of disease occurrence in different populations. Relative risk is defined and shown to have both empirical and logical advantages as a measure of disease/risk factor association, especially in connection with case-control studies. The close connection between cohort and case-control studies is emphasised throughout.

2.1 Measures of disease occurrence

Two measures of disease frequency, incidence and prevalence, are commonly introduced in textbooks on epidemiology. *Point prevalence* is the proportion of a defined population affected by the disease in question at a specified point in time. The numerator of the proportion comprises all those who have the disease at that instant, regardless of whether it was contracted recently or long ago. Thus, diseases of long duration tend to have a higher prevalence than short-term illnesses, even if the total numbers of affected individuals are about equal.

Incidence refers to new cases of disease occurring among previously unaffected individuals. This is a more appropriate measure for etiological studies of cancer and other chronic illnesses, wherein one attempts to relate disease occurrence to genetic and environmental factors in a framework of causation. The duration of survival of patients with a given disease, and hence its prevalence, may be influenced by treatment and other factors which come into play after onset. Early reports of an association between the antigen HL-A2 and risk for acute leukaemia (Rogentine et al., 1972), for example, were later corrected when it was shown that the effect was on survival rather than on incidence (Rogentine et al., 1973). Since causal factors necessarily

operate prior to diagnosis, a more sensitive indication of their effects is obtained by using incidence as the fundamental measure of disease.

Rates, as opposed to frequencies, imply an element of time. The rate of occurrence of an event in a population is the *number of events which occur during a specified time interval, divided by the total amount of observation time accumulated during that interval.* For an incidence rate, the events are new cases of disease occurring among disease-free individuals. The denominator of the rate can be calculated by summing up the length of time during the specified interval that each member of the population was alive and under observation, without having developed the disease. It is usually expressed as the number of person-years of observation. Mortality rates, of course, refer to deaths occurring among those who remain alive.

The annual incidence rate for a particular calendar year is the number of new cases diagnosed during the year, divided by an approximation of the person-years of observation, such as the midyear population. If the disease is a common one, the denominator should refer more specifically to the subjects who are disease-free at midyear and hence at risk of disease development. This correction is rarely needed for cancer occurring at specific sites because the number of people alive with disease will be relatively small. One exception to this which illustrates the general principle is that of uterine cancer. In societies where a substantial fraction of older women have undergone a hysterectomy, the denominators used to calculate rates of cervical or endometrial cancer should include only women with an intact uterus, as the remainder are no longer at risk for the particular disease. This adjustment is particularly important when comparing cancer incidence among populations with different hysterectomy rates.

In calculating incidence rates *time* is usually taken to be *calendar time.* An annual rate is thus based on all cases which occur between January 1 and December 31 of a given year. However, there are other ways of choosing the origin of the time-scale besides reference to a particular date on the calendar.

Chronological age, for example, is simply elapsed time from birth. The fact that cancer incidence rates are routinely reported using age as the fundamental "time" variable reflects the marked variation of incidence with age which is found for most cancer sites. A typical practice is to use J = 18 age intervals, each having a constant length of five years (0–4, 5–9, ... 80–84, 85–89), ignoring cases occurring at age 90 or over. Sometimes the first interval is chosen to be of length $l_1 = 1$ (first year of life), the second of length $l_2 = 4$ (ages 1–4) and the remainder to have a constant length of 5 years. Cases of disease are allocated to each interval according to the age at diagnosis. Since individual ages will change during the period of observation, the same person may contribute to the person-years denominators for several age intervals.

Yet another possibility for the time variable is *time on study.* In prospective epidemiological investigations of industrial populations, for example, workers may enter the study after two or five years of continuous employment. Time is then measured as years elapsed since entry into the study. *Survival rates* for cancer and other diseases are presented in terms of elapsed months or years since diagnosis or definitive treatment. Here of course the endpoint is death for patients with disease. When using time on study as the fundamental time variable it is usually quite important to account also for the effects of age, whether one is calculating survival rates among cancer patients or cancer incidence rates among a cohort of exposed workers.

Fig. 2.1 Schematic illustration of age-specific incidence rates. (D = diagnosis of cancer; W = withdrawn, disease free.)

	Period 1	Period 2	Period 3
Number of events	2	4	3
Total observation time	59.7	93.5	43.6
rate	0.034	0.043	0.069

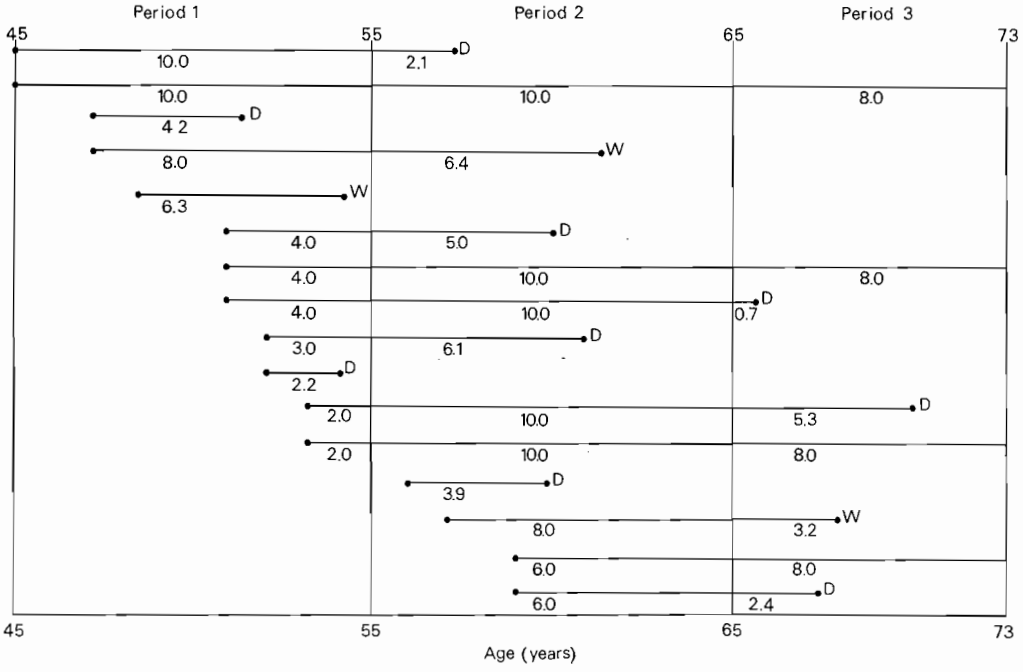


Figure 2.1 illustrates schematically the method of calculation of incidence rates for a study in which the time axis is divided into intervals: 45–54, 55–64 and 65–72 years inclusive. In this case time in fact means age. Subjects are arranged according to their age at entry to the study, which for simplicity has been taken to correspond to a birthday. The first subject, who entered the study on his 45th birthday and developed the disease (D) early in his 57th year, contributes 10 years of observation and no events to the 45–54 age period and 2.1 years and one event to the 55–64 age period. The third subject, who entered the study at age 47, was withdrawn (W) from observation during his 61st year (perhaps due to death from another disease) and hence contributes only to the denominator of the rate.

The least ambiguous definition of a rate results from making the time intervals short. This is because populations themselves change over time, through births, deaths or migrations, so that the shorter the time interval, the more stable the denominator used in the rate calculations. Also, the rate itself may be changing during the interval. If the

change is rapid it makes sense to consider short intervals so that information about the magnitude of the change is not lost; but if the intervals are too short only a few events will be observed in each one. The instability of the denominator must be balanced against statistical fluctuations in the numerator when deciding upon an appropriate time interval for calculation of a reasonably stable rate.

If an infinite population were available, so that statistical stability was not in question, one could consider making the time intervals used for the rate calculation infinitesimal. As the length of each interval approaches zero, one obtains in the limit an *instantaneous rate* $\lambda(t)$ defined for each instant t of time. This concept has proved very useful in actuarial science, where, with the event in question being death, $\lambda(t)$ represents the *force of mortality*. In the literature of reliability analysis, where the event is failure of some system component, $\lambda(t)$ is referred to as the *hazard* rate. When the endpoint is diagnosis of disease in a previously disease-free individual, we can refer to the instantaneous incidence rate as the *force of morbidity*.

The method of calculation of the estimated rate will depend upon the type of data available for analysis. It is perhaps simplest in the case of a longitudinal follow-up study of a fixed population of individuals, for example: mice treated with some carcinogen who are followed from birth for appearance of tumours; cancer patients followed from time of initial treatment until relapse or death; or employees of a given industry or plant who are followed from date of employment until diagnosis of disease. A common method of estimating incidence or mortality rates with such data is to divide the time axis into J intervals having lengths l_j and midpoints t_j. Denote by n_j the number of subjects out of the original population of n_0 who are still under observation and at risk at t_j. Let d_j be the number of events (diagnoses or deaths) observed during the j^{th} interval. Then the incidence at time t_j may be estimated by

$$\lambda(t_j) = \frac{d_j}{l_j \times n_j} \tag{2.1}$$

that is, by the number of events observed *per subject, per unit time* in the population at risk during the interval. Of course the denominator in equation (2.1) is only an approximation to the total observation time accumulated during the interval, which should be used if available.

Example: An example of the calculation of incidence rates from follow-up studies is given in Table 2.1 which lists the days until appearance of skin tumours for a group of 50 albino mice treated with benzo[*a*]-pyrene (Bogovski & Day, 1977). For the purpose of illustration, the duration of the study has been divided into four periods of unequal length: 0–179 days, 180–299 days, 300–419 days and 420–549 days. These are rather wider than is generally desirable because of limited data. Nineteen of the animals survived the entire 550 days without developing skin tumours, and are listed together at the bottom of the table. The contribution of each animal to the number of tumours and total observation time for each period are shown. Thus, the mouse developing tumour at 377 days contributes 0 tumours and 180 days observation to Period 1, 0 tumours and 120 days observation to Period 2, and 1 tumour and 78 days observation to Period 3.

Tumour incidence rates shown at the bottom of Table 2.1 were calculated in two ways. The first used the actual total observation time in each period, while the second used the approximation to this based on the number of animals alive at the midpoint (equation 2.1). Thus the incidence rate for Period 1 is 0 as no tumours were observed. For Period 2, 7 tumours were seen during 5 415 mouse-days of observa-

tion for a rate of (7/5 415) × 1 000 = 1.293 per 1 000 mouse-days. The approximate rate is [7/(47 × 120)] × 1 000 = 1.241 tumours per 1 000 mouse-days. The rate increases during the third period and then falls off.

Except in rare instances, cancer incidence rates are not obtained by continuous observation of all members of a specified population. Since the production of stable rates for cancers at most individual sites requires a population of at least one million subjects, the logistic and financial problems of attempting to maintain a constant sur-

Table 2.1 Calculation of incidence rate of skin tumours in mice treated with benzo[*a*]pyrene[a]

No. of animals if greater than one	Day of tumour appearance or day of death without tumour (*)	No. of animals at risk at start of each day	Contribution to rate calculation by period							
			Period 1 (0–179 days)		Period 2 (180–299 days)		Period 3 (300–419 days)		Period 4 (420–549 days)	
			No.[b]	Days[c]	No.	Days	No.	Days	No.	Days
	178*	50		179						
	187	49		180	1	8				
	194	48		180	1	15				
(3)	243	47		540	3	192				
	257	44		180	1	78				
	265*	43		180		86				
	297	42		180	1	118				
	297*	41		180		118				
(2)	327	40		360		240	2	56		
(2)	336	38		360		240	2	74		
	377	36		180		120	1	78		
	379	35		180		120	1	80		
	390*	34		180		120		91		
(2)	399	33		360		240	2	200		
	413	31		180		120	1	114		
	431*	30		180		120		120		12
	432*	29		180		120		120		13
(2)	444*	28		360		240		240		50
	482*	26		180		120		120		63
	495*	25		180		120		120		76
	515*	24		180		120		120		96
	522*	23		180		120		120		103
(2)	544*	22		360		240		240		250
	549	20		180		120		120	1	130
(19)	550*	19		3 420		2 280		2 280		2 470
Totals			0	8 999	7	5 415	9	4 293	1	3 263
No. animals at risk at midpoint				50		47		36		25
Length of interval (days)				180		120		120		130
Rate[d] (per 1 000 mouse-days)				0		1.293		2.096		0.306
Rate[e] (per 1 000 mouse-days)				0		1.241		2.083		0.308

[a] From Bogovski and Day (1977)
[b] No. of tumours observed during period
[c] Contribution to observation time during period
[d] Rate calculated using total observation time in denominator
[e] Rate calculated from equation (2.1)

veillance system are usually prohibitive. The information typically available to a cancer registry for calculation of rates includes the cancer cases, classified by sex, age and year of diagnosis, together with *estimates* of the population denominators obtained from the census department. How good the estimated denominators are depends on the frequency and accuracy of the census in each locality.

Example: Table 2.2 illustrates the calculation of the incidence of acute lymphatic leukaemia occurring among males aged 0–14 years in Birmingham, UK, during 1968–72 (Waterhouse et al., 1976). The numbers of cases (d_j), classified by age, and the number of persons (n_j) in each age group in 1971, the mid-year of the observation period, are shown. In order to approximate the total person-years of observation, n_j is multiplied by the length of the observation period, namely five years. While this is adequate if the population size and age distribution remain fairly stable, this procedure would not suffice for times of rapid change in population structure. A better approximation to the denominator for the 1–4 year age group, for example, would be to sum up the numbers of 1–4 year-olds in the population at mid–1968 plus those at mid–1969 and so on to 1972. As is standard for cancer incidence reporting, the rates are expressed as numbers of cases per 100 000 person-years of observation. Table 2.3 presents the calculated rates for three additional sites and a larger number of age groups.

Table 2.2 Average annual incidence rates of acute lymphatic leukaemia for males aged 0–14 Birmingham region (1968–72)[a]

Age (years)	Interval length (l)	No. of cases (d)	Population (1971) (n)	No. of years of observation (1968–72)	Rate[b] (per 100 000 person-years)
0	1	2	45 300	5	0.88
1–4	4	47	182 400	5	5.15
5–9	5	30	228 300	5	2.63
10–14	5	13	202 500	5	1.28

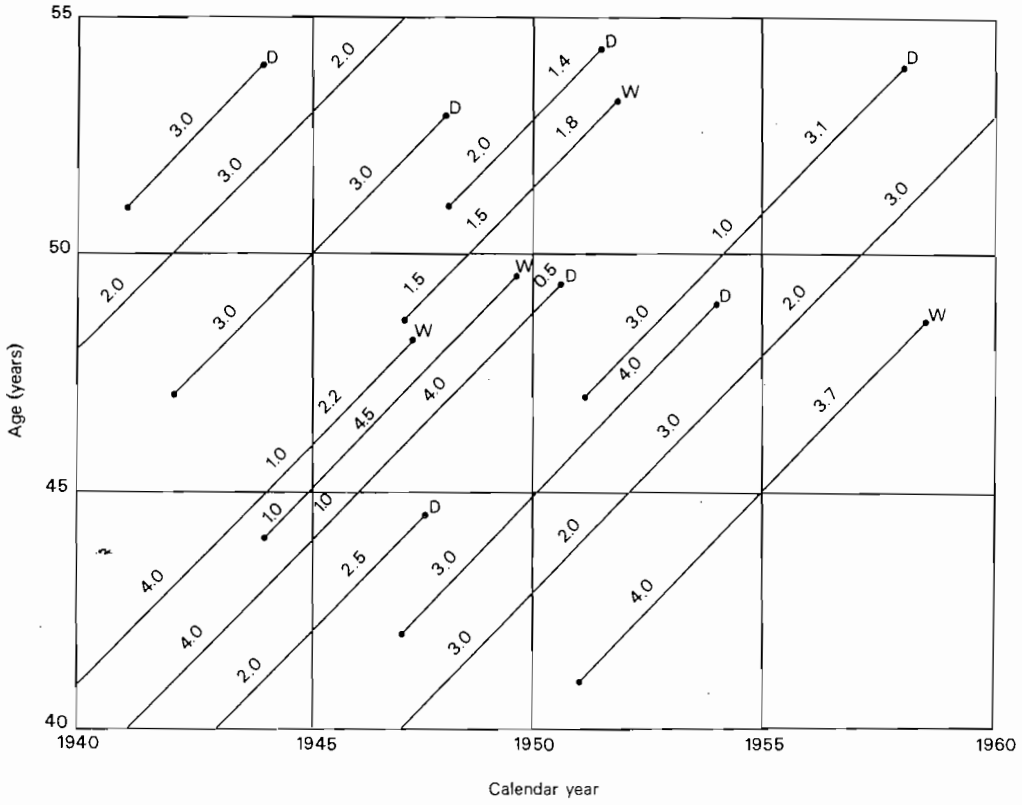
[a] From Waterhouse et al. (1976)

[b] Rate $= \frac{d}{n \times 5} \times 100\,000$

2.2 Age- and time-specific incidence rates

If the population has been under observation for several decades, cases of disease and person-years at risk may be classified usefully by both calendar year and age at diagnosis. The situation is illustrated in Figure 2.2. As each study subject is followed forward in time, he traces out a 45° trajectory in the age × time plane. Person-years of observation are allocated to the various age × time cells traversed by this path, and diagnoses of cancer or other events are assigned to the cell in which they occur. Thus, the upper left-hand cell in Figure 2.2, corresponding to ages 50–54 years and the 1940–44 time period, contains 1 death and 6 person-years of observation for a rate of $1/6 \times 100 = 16.7$ events per 100 person-years. An analysis of age-specific rates averaged over a certain calendar period would ignore the time axis in this diagram (as in Figure 2.1), while an analysis of time-specific rates would ignore the age classification. Typical practice is to consider five-year intervals of age and time, so as to be

Fig. 2.2 Schematic diagram of a follow-up study with joint classification by age and year. (D = diagnosis of cancer; W = withdrawn, disease free.)



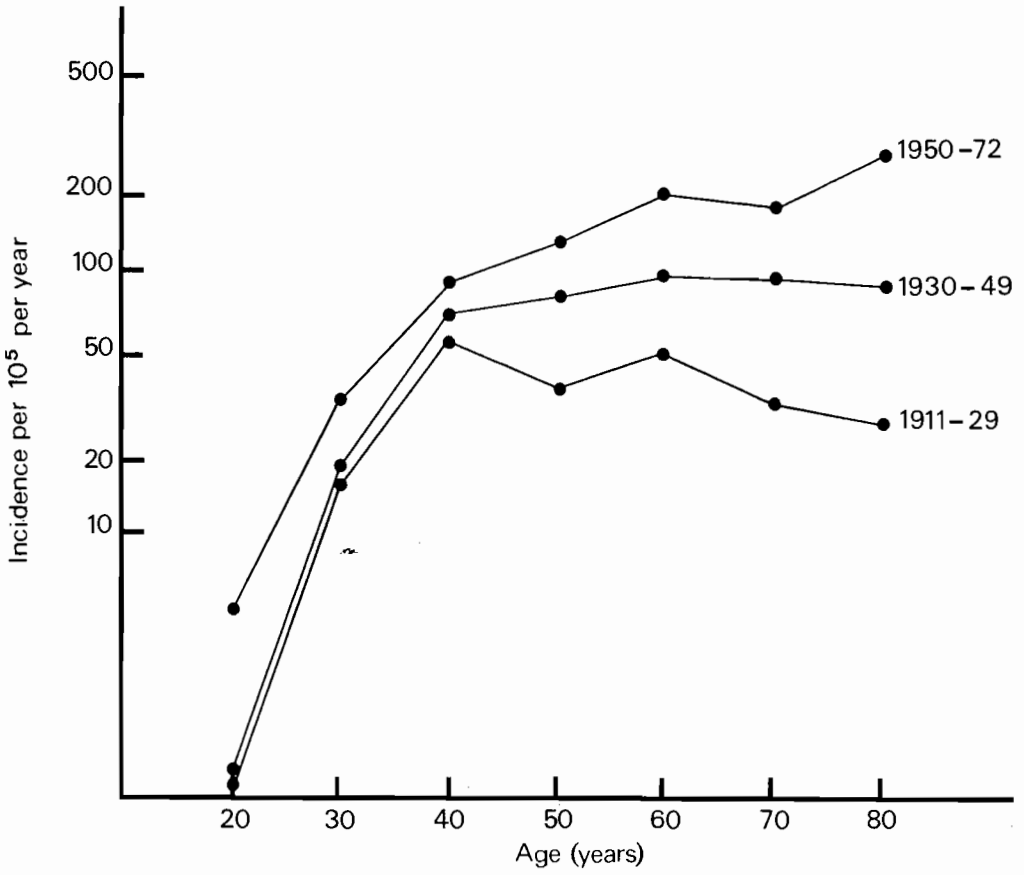
able to study the reasonably fine details of the variation in rates; but this will depend on the amount of data available.

A *cross-sectional* analysis results from fixing the calendar periods and examining the age-specific incidences. Alternatively, in a *birth-cohort* analysis, the same cancer cases and person-years are classified according to year of birth and age. This is possible since any two of the three variables (1) year of birth, (2) age and (3) calendar year determine the third. In Figure 2.2, for example, the 1890–99 birth cohort would be represented by the diagonal column of 45° lines intersecting the vertical axis between 40 and 50 years of age in 1940.

Example: Figure 2.3 shows the age-specific incidence of breast cancer in Iceland during the three calendar periods 1910–29, 1930–49 and 1950–72 (Bjarnasson et al., 1974). While the three curves show a general increase in incidence with calendar time, they also have rather different shapes. There was a decline in incidence with age after 40 years during the 1911–29 period, a fairly constant incidence during 1930–49 and an increase in incidence with age during the latest calendar period.

If the data are rearranged into birth cohorts, a more coherent picture emerges. Figure 2.4 shows the age incidence curves for three cohorts of Icelandic women born in 1840–79, 1880–1909 and 1910–49,

Fig. 2.3 Age-specific incidence of breast cancer in Iceland for the three time periods 1911–29, 1930–49, 1950–72. From Bjarnasson et al. (1974).



respectively. Because the period of case ascertainment was limited to the years 1910–72, the age ranges covered by these three curves are different. However, their shapes are much more similar than for the cross-sectional analysis of Figure 2.3; there is a fairly constant distance between the three curves on the semi-logarithmic plot. Since the ratios of the age-specific rates for different cohorts are therefore nearly constant across the age span, one may conveniently summarize the inter-cohort differences in terms of ratios of rates.

2.3 Cumulative incidence rates

While the importance of calculating age- or time-specific rates using reasonably short intervals cannot be overemphasized, it is nevertheless often convenient to have a single synoptic figure to summarize the experience of a population over a longer time span or age interval. For example, in comparing cancer incidence rates between different countries, it is advisable to make one comparison for children aged 0–14, another for

$$\Lambda(t) = \sum_{n=0}^{t} \lambda(n)$$

where the $\lambda(n)$ give the annual age-specific rates. In precise mathematical terms, the cumulative incidence rate between time 0 and t is expressed by an integral

$$\Lambda(t) = \int_0^t \lambda(u)du \tag{2.2}$$

where $\lambda(u)$ represents the instantaneous rate. The cumulative incidence between 15 and 34 years, inclusive, would be obtained from yearly rates as

$$\Lambda(34) - \Lambda(14) = \sum_{n=15}^{34} \lambda(n).$$

In practice, age-specific rates may not be available for each individual year of life but rather, as in the previous example, for periods of varying length such as 5 or 10 years. Then the age-specific rate $\lambda(t_i)$ for the i^{th} period is multiplied by its length l_i before summing:

$$\hat{\Lambda}(t_j) = \sum_{i=1}^{j} l_i \lambda(t_i).$$

When calculating the cumulative rate from longitudinal data, we have, using (2.1),

$$\hat{\Lambda}(t_j) = \frac{d_1}{n_1} + \ldots + \frac{d_j}{n_j}, \tag{2.3}$$

where the d_i are the deaths and the n_i are the numbers at risk at the midpoint of each time interval.

One reason for interest in the cumulative incidence rate is that it has a useful probabilistic interpretation. Let P(t) denote the net *risk,* or *probability,* that an individual will develop the disease of interest between time 0 and t. We assume for this definition that he remains at risk for the entire period, and is not subject to the *competing risks* of loss or death from other causes. The instantaneous incidence rate at time t then has a precise mathematical definition as the rate of increase in P(t), expressed relative to the proportion of the population still at risk (Elandt-Johnson, 1975). In symbols

$$\lambda(t) = \frac{1}{1 - P(t)} \times \frac{dP(t)}{dt}.$$

From this it follows that

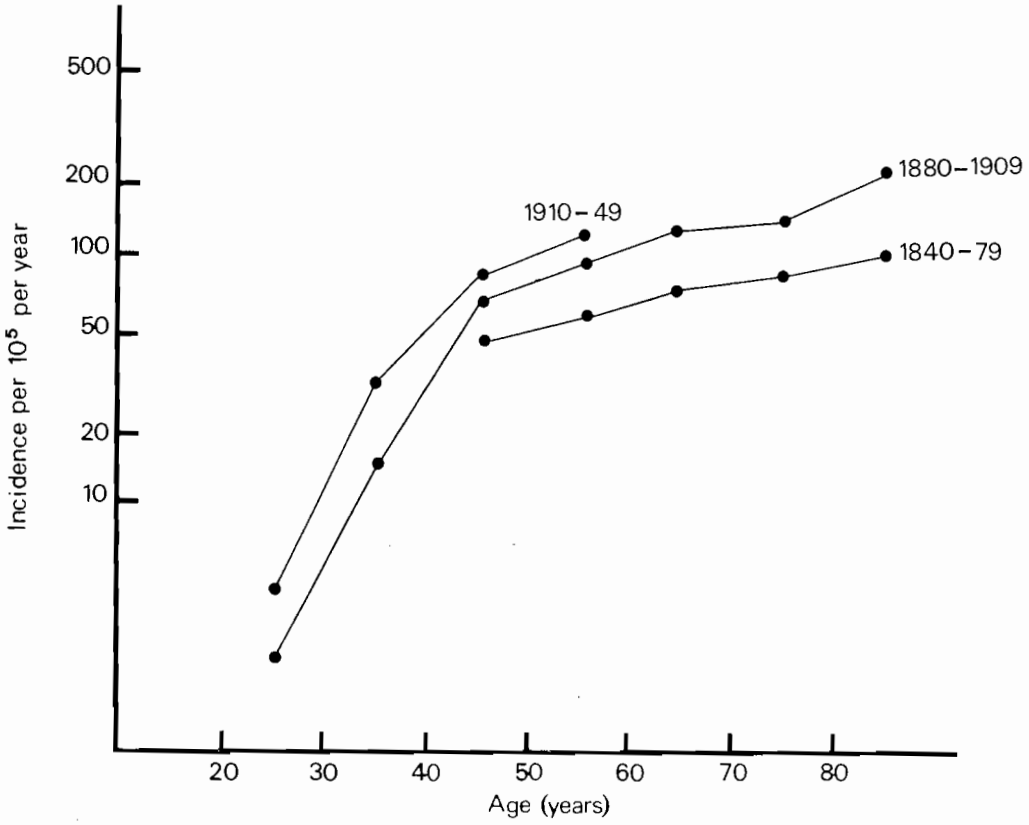
$$1\text{-}P(t) = \exp\{-\Lambda(t)\}, \tag{2.4}$$

or, using logarithms[1] rather than exponentials,

$$\Lambda(t) = -\log\{1\text{-}P(t)\}.$$

[1] log denotes the natural logarithm, i.e., to the base e, which is used exclusively throughout the text.

Fig. 2.4 Age-specific incidence of breast cancer in Iceland for three birth cohorts, 1840–1879, 1880–1909, 1910–1949. Adapted from Bjarnasson et al. (1974).



young adults aged 15–34, and a third for mature adults aged 35–69. Comparison of rates among the elderly may be inadvisable due to problems of differential diagnosis among many concurrent diseases.

The usual method of combining such age-specific rates for comparison across different populations is that of direct standardization (Fleiss, 1973). The *directly standardized* (adjusted) rate consists of a weighted average of the age-specific rates for each study group, where the weights are chosen to be proportional to the age distribution of some external standard population. Hypothetical standard populations have been constructed for this purpose, which reflect approximately the age structure of World, European or African populations (Waterhouse et al., 1976); however, the choice between them often seems rather arbitrary.

An alternative and even simpler summary measure is the *cumulative incidence rate,* obtained by summing up the annual age-specific incidences for each year in the defined age interval (Day, 1976). Thus the cumulative incidence rate between 0 and t years of age, inclusive, is

These equations tell us that when the disease is rare or the time period short, so that the cumulative incidence or mortality is small, then the probability of disease occurrence is well approximated by the cumulative incidence

$$P(t) \approx \Lambda(t). \tag{2.5}$$

Example: To illustrate the calculation of a cumulative rate, consider the age-specific rates of urinary tract tumours (excluding bladder) for Birmingham boys between 0 and 14 years of age (Table 2.3). These are almost entirely childhood tumours of the kidney, i.e., Wilms' tumours or nephroblastomas. The period cumulative rate is calculated as $(1 \times 2.2) + (4 \times 1.0) + (5 \times 0.4) + (5 \times 0.0) = 8.20$ per 100 000 population. Note that the first two age intervals have lengths of 1 and 4 years, respectively, while subsequent intervals are five years each. Table 2.4 shows the cumulative rates for all four tumours in Table 2.3 using three age periods: 0–14, 15 –34 and 35–69. Also shown are the cumulative risks, i.e., probabilities, calculated from the rates according to equation (2.4). With the exception of lung cancer, which has a cumulative rate approaching 0.1 for the 35–69 age group, the rates and risks agree extremely well.

Table 2.3 Average annual incidence per 100 000 population by age group for Birmingham region, 1968–72 (males)[a]

Age (years)	Tumour site			
	Urinary tract (excl. bladder)	Stomach	Lung	Lymphatic leukaemia
0	2.2	0.0	0.0	0.9
1–4	1.0	0.0	0.0	5.2
5–9	0.4	0.0	0.0	2.6
10–14	0.0	0.0	0.0	1.3
15–19	0.1	0.0	0.1	1.0
20–24	0.2	0.1	0.7	0.4
25–29	0.1	0.7	0.8	0.3
30–34	0.5	0.7	3.3	0.6
35–39	1.2	4.3	9.1	0.6
40–44	4.0	7.6	25.6	0.9
45–49	4.6	18.1	71.4	1.5
50–54	7.1	31.3	137.4	1.6
55–59	11.8	64.1	257.5	4.3
60–64	16.7	100.6	404.9	7.0
65–69	21.7	150.2	520.3	11.2

[a] From Waterhouse et al. (1976)

Estimates of the cumulative rate are much more stable numerically than are estimates of the component age- or time-specific rates, since they are based on all the events which occur in the relevant time interval. This stability makes the cumulative rate the method of choice for reporting results of small studies. An estimate of $\Lambda(t)$ for such studies may be obtained by applying equation (2.3), with the chosen intervals so fine that each event occupies its own separate interval. In other words, we simply sum up, for each event occurring before or at time t, the reciprocal of the number of subjects remaining at risk just prior to its occurrence.

Table 2.4 Cumulative rates and risks, in percent, of developing cancer between the indicated ages: calculated from Table 2.3

Age period (years)		Tumour site			
		Urinary tract (excl. bladder)	Stomach	Lung	Acute lymphatic leukaemia
0–14	Rate	0.0082	0.0	0.0	0.0412
	Risk	0.0082	0.0	0.0	0.0412
15–34	Rate	0.0045	0.0075	0.0245	0.0115
	Risk	0.0045	0.0075	0.0245	0.0115
35–69	Rate	0.3355	1.8810	7.1310	0.1355
	Risk	0.3349	1.8634	6.8827	0.1355

Example: Consider the data on murine skin tumours shown in Table 2.1. Since 49 animals remain at risk at the time of appearance of the first tumour, t = 187 days, the cumulative rate is estimated as $\hat{\Lambda}(187) = 1/49 = 0.020$. The estimate at t = 243 days is given by

$$\hat{\Lambda}(243) = \frac{1}{49} + \frac{1}{48} + \frac{1}{47} + \frac{1}{46} + \frac{1}{45} = 0.106.$$

Note that the contribution from the three tumours occurring at 243 days, when 47 animals remain at risk, is given by (1/47) + (1/46) + (1/45) rather than (3/47). This is consistent with the idea that the three tumours in fact occur at slightly different times, which are nevertheless too close together to be distinguished by the recording system.

Only 20 animals remain at risk at the time of the last observed tumour, 549 days, the others having already died or developed tumours. Hence this event contributes 1/20 = 0.05 to the cumulative rate, bringing the total to

$$\frac{1}{49} + \frac{1}{48} + \frac{1}{47} + \frac{1}{46} + \frac{1}{45} + \ldots + \frac{1}{20} = 0.457.$$

The risk of developing a skin tumour in the first 550 days is thus estimated to be $1 - \exp(-0.457) = 0.367$ for mice in this experiment who survive the entire study period. Figure 2.5 shows the cumulative incidence rate plotted as a function of days to tumour appearance.

In summary, three closely related measures are available for expressing the occurrence of disease in a population: the instantaneous incidence rate defined at each point in time; the cumulative incidence rate defined over an interval of time; and the probability or risk of disease, also defined over an interval of time. Our next task is to consider how exposure of the population to various risk factors may affect these same rates and risks of disease occurrence.

2.4 Models of disease association

The simplest types of risk factors are the *binary* or "all or none" variety, as exemplified by the presence or absence of a particular genetic marker. Environmental variables are usually more difficult to quantify since individual histories vary widely with respect to the onset, duration and intensity of exposure, and whether it was continuous or intermittent. Nevertheless it is often possible to make crude classifications into an

both even greater than R. Consider the above formulation in the case $R = 9$, $R_0 = 1$ and $K = 1$. Let p_2 denote the proportion of exposed individuals in population 2 and let p_1 be the same for population 1. In order for the difference between these two proportions to explain completely the ninefold excess we must have $w>9$, i.e., $(1-p_2) + rp_2 > 9\{(1-p_1)+rp_1\}$, which implies both $p_2 > 9p_1 + 8/(r-1) > 9p_1$ and $r>9$.

We end this chapter with a brief word of caution regarding the interpretation of attributable risks, whether relative or absolute. For pedagogic reasons, language was occasionally used which seemed to imply that the elimination of a particular risk factor would result in a measured reduction in incidence. This of course supposes that the association between risk factor and disease as estimated from the observational study is in fact a causal one. Unfortunately, the only way to be absolutely certain that a causal relationship exists is to intervene actively in the system by removing the disputed factor. In the absence of such evidence, a more cautious interpretation of the attributable risk measures would be in terms of the proportion of risk *explained* by the given factor, where "explain" is used in the limited sense of statistical association. The next chapter considers in some detail the problem of drawing causal inferences from observational data such as those collected in case-control studies.

REFERENCES

Beebe, G.W., Kato, H. & Land, C.E. (1977) Mortality experience of atomic bomb survivors 1950–74. *Life Span Study Report 8,* Hiroshima, Radiation Effects Research Foundation

Berkson, J. (1958) Smoking and lung cancer: some observations on two recent reports. *J. Am. stat. Assoc., 53,* 28–38

Berry, G., Newhouse, M.L. & Turok, M. (1972) Combined effect of asbestos exposure and smoking on mortality from lung cancer in factory workers. *Lancet, ii,* 476–479

Bjarnasson, O., Day, N.E., Snaedal, G. & Tulinius, H. (1974) The effect of year of birth on the breast cancer incidence curve in Iceland. *Int. J. Cancer, 13,* 689–696

Bogovski, P. & Day, N.E. (1977) Accelerating action of tea on mouse skin carcinogenesis. *Cancer Lett., 3,* 9–13

Boice, J.D. & Monson, R.R. (1977) Breast cancer in women after repeated fluoroscopic examinations of the chest. *J. natl Cancer Inst., 59,* 823–832

Boice, J.D. & Stone, B.J. (1979) *Interaction between radiation and other breast cancer risk factors.* In: *Late Biological Effects of Ionizing Radiation, Vol. I,* Vienna, International Atomic Energy Agency, pp. 231–249

Cole, P. & MacMahon, B. (1971) Attributable risk percent in case-control studies. *Br. J. prev. soc. Med., 25,* 242–244

Cook, P., Doll, R. & Fellingham, S.A. (1969) A mathematical model for the age distribution of cancer in man. *Int. J. Cancer, 4,* 93–112

Cornfield, J. (1951) A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *J. natl Cancer Inst., 11,* 1269–1275

Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B. & Wynder, E.L. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *J. natl Cancer Inst., 22,* 173–203

Court Brown, W.M. & Doll, R. (1965) Mortality from cancer and other causes after radiotherapy for ankylosing spondylitis. *Br. med. J., ii,* 1327–1332

Day, N. (1976) *A new measure of age standardized incidence, the cumulative rate.* In: Waterhouse, J.A.H., Muir, C.S., Correa, P. & Powell, J., eds, *Cancer Incidence in Five Continents,* Vol. III, Lyon, International Agency for Research on Cancer *(IARC Scientific Publications No. 15),* pp. 443–452

Doll, R. (1971) The age distribution of cancer: implications for models of carcinogenesis. *J. R. stat. Soc. A, 132,* 133–166

Doll, R. & Peto, R. (1976) Mortality in relation to smoking: 20 years' observations on male British doctors. *Br. med. J., ii,* 1525–1536

Elandt-Johnson, R. (1975) Definition of rates: some remarks on their use and misuse. *Am. J. Epidemiol., 102,* 267–271

Eyigou, A. & McHugh, R. (1977) On the factorization of the crude relative risk. *Am. J. Epidemiol., 106,* 188–193

Fleiss, J.L. (1973) *Statistical Methods for Rates and Proportions.* New York, Wiley

Hammond, E.C. (1966) Smoking in relation to the death rates of one million men and women. *Natl Cancer Inst. Monogr., 19,* 127–204

Hammond, E.C., Selikoff, I.J. & Seidman, H. (1979) Asbestos exposure, cigarette smoking and death rates. *Ann. N.Y. Acad. Sci., 330,* 473–490

Koopman, J.S. (1977) Causal models and sources of interaction. *Am. J. Epidemiol., 106,* 439–444

Levin, M.L. (1953) The occurrence of lung cancer in man. *Acta Unio Int. Cancer, 9,* 531–541

Liddell, F.D.K., McDonald, J.C. & Thomas, D.C. (1977) Methods of cohort analysis: appraisal by application to asbestos mining. *J. R. stat. Soc. Ser. A, 140,* 469–491

MacGregor, P.H., Land, C.E., Choi, K., Tokuota, S., Liu, P.I., Wakabayoshi, T. & Beebe, G.W. (1977) Breast cancer incidence among atomic bomb survivors, Hiroshima and Nagasaki, 1950–69. *J. natl Cancer Inst., 59,* 799–811

MacMahon, B., Cole, P., Lin, T.M., Lowe, C.R., Mirra, A.P., Ravnihar, B., Salber, E.J., Valaoras, V.G. & Yuasa, S. (1970) Age at first birth and breast cancer risk. *Bull. World Health Org., 43,* 209–221

Miettinen, O.S. (1972) Components of the crude risk ratio. *Am. J. Epidemiol., 96,* 168–172

Mosteller, F. & Tukey, J. (1977) *Data Analysis and Regression: A Second Course in Statistics,* Reading, MA, Addison & Wesley

Prentice, R.L. & Breslow, N.E. (1978) Retrospective studies and failure time models. *Biometrika, 65,* 153–158

Rogentine, G.N., Yankee, R.A., Gart, J.J., Nam, J. & Traponi, R.J. (1972) HL-A antigens and disease. Acute lymphocytic leukemia. *J. clin. Invest., 51,* 2420–2428

Rogentine, G.N., Traponi, R.J., Yankee, R.A. & Henderson, E.S. (1973) HL-A antigens and acute lymphocytic leukemia: the nature of the HL-A2 association. *Tissue Antigens, 3,* 470–475

Rothman, K. (1976) The estimation of synergy or antagonism. *Am. J. Epidemiol., 103,* 506–511

Rothman, K.J. & Keller, A.Z. (1972) The effect of joint exposure to alcohol and tobacco on risk of cancer of the mouth and pharynx. *J. chron. Dis., 23,* 711–716

Saracci, R. (1977) Asbestos and lung cancer: an analysis of the epidemiological evidence on the asbestos-smoking interaction. *Int. J. Cancer, 20,* 323–331

Schlesselman, J.J. (1978) Assessing the effects of confounding variables. *Am. J. Epidemiol., 108,* 3–8

Selikoff, I.O. & Hammond, E.C. (1978) Asbestos associated disease in United States shipyards. *CA: A Cancer Journal for Clinicians, 28,* 87–99

Shore, R.E., Hempelmann, L.A., Kowaluk, E., Mansur, P.S., Pasternack, B.S., Albert, R.E. & Haughie, G.E. (1977) Breast neoplasms in women treated with X-rays for acute postpartum mastitis. *J. natl Cancer Inst., 59,* 813–822

Smith, P.G. (1979) *Some problems in assessing the carcinogenic risk to man of exposure to ionizing radiations.* In: Breslow, N. & Whittemore, A., eds, *Energy and Health,* Philadelphia, Society for Industrial and Applied Mathematics, pp. 61–80

Walter, S.D. (1975) The distribution of Levin's measure of attributable risk. *Biometrika, 62,* 371–374

Waterhouse, J., Muir, C., Correa, P. & Powell, J., eds (1976) *Cancer Incidence in Five Continents,* Vol. III, Lyon, International Agency for Research on Cancer *(IARC Scientific Publications No. 15)*

Wynder, E.L. & Bross, I.J. (1961) A study of etiological factors in cancer of the esophagus. *Cancer, 14,* 389–413

Wynder, E.L., Bross, I.J. & Feldman, R.M. (1957) A study of etiological factors in cancer of the mouth. *Cancer, 10,* 1300–1323

LIST OF SYMBOLS – CHAPTER 2 (in order of appearance)

l_j	length of j^{th} time interval for rate calculation
$\lambda(t)$	instantaneous event (e.g., incidence) rate at time t
t_j	midpoint of j^{th} time interval for rate calculation
d_j	number of events (e.g., cancer diagnoses) in j^{th} time interval
n_j	number of subjects under observation at midpoint of j^{th} time interval
$\Lambda(t)$	cumulative event (e.g., incidence) rate at time t
P(t)	cumulative risk or probability of occurrence of an event (e.g., diagnosis of disease) by time t
$\approx$	approximate equality
$\hat{\Lambda}(t)$	estimated cumulative rate
λ_{1i}	disease incidence rate in i^{th} stratum among persons exposed to risk factor
λ_{0i}	disease incidence rate in i^{th} stratum among persons not exposed to risk factor
b_i	difference in incidence rates for exposed *versus* non-exposed in i^{th} stratum
b	difference in incidence rates for exposed *versus* non-exposed in additive model
r_i	ratio of incidence rates for exposed *versus* non-exposed in i^{th} stratum
r	ratio of incidence rates for exposed *versus* non-exposed in multiplicative model; rate ratio; relative risk
β	logarithm of relative risk for exposed *versus* non-exposed

P_0	cumulative risk or probability of disease diagnosis among those not exposed to the risk factor
P_1	cumulative risk or probability of disease diagnosis among those exposed to the risk factor
$\lambda_i(t)$	average annual incidence rate for i^{th} area at age t
β_i	logarithm of relative risk of stomach cancer for country i *versus* country 1
γ	slope in fit of straight line to log-log plot of age-incidence data
r_i	relative risk of stomach cancer for country i *versus* country 1
r_{ij}	relative risk of exposure to level i of one risk factor and level j of another, with reference to the non-exposed
p	proportion of population exposed to risk factor
$Q_0 = 1-P_0$	proportion of non-exposed population which remains disease-free
$Q_1 = 1-P_1$	proportion of exposed population which remains disease-free
ψ	$P_1Q_0/(Q_1P_0)$; odds ratio of disease probabilities for exposed *versus* non-exposed groups
p_1	probability of exposure among diseased
p_0	probability of exposure among disease-free
AR	population attributable risk
p_{1k}	proportion of first population exposed to level k of a risk factor
p_{2k}	proportion of second population exposed to level k of a risk factor
R	crude ratio of incidence rates between two populations
λ_{10}	incidence rate for non-exposed in population 1
λ_{20}	incidence rate for non-exposed in population 2
R_0	ratio of incidence rates for non-exposed, population 2 to population 1
w	(multiplicative) confounding factor
RAR	relative attributable risk
AR_1	attributable risk for population 1
AR_2	attributable risk for population 2