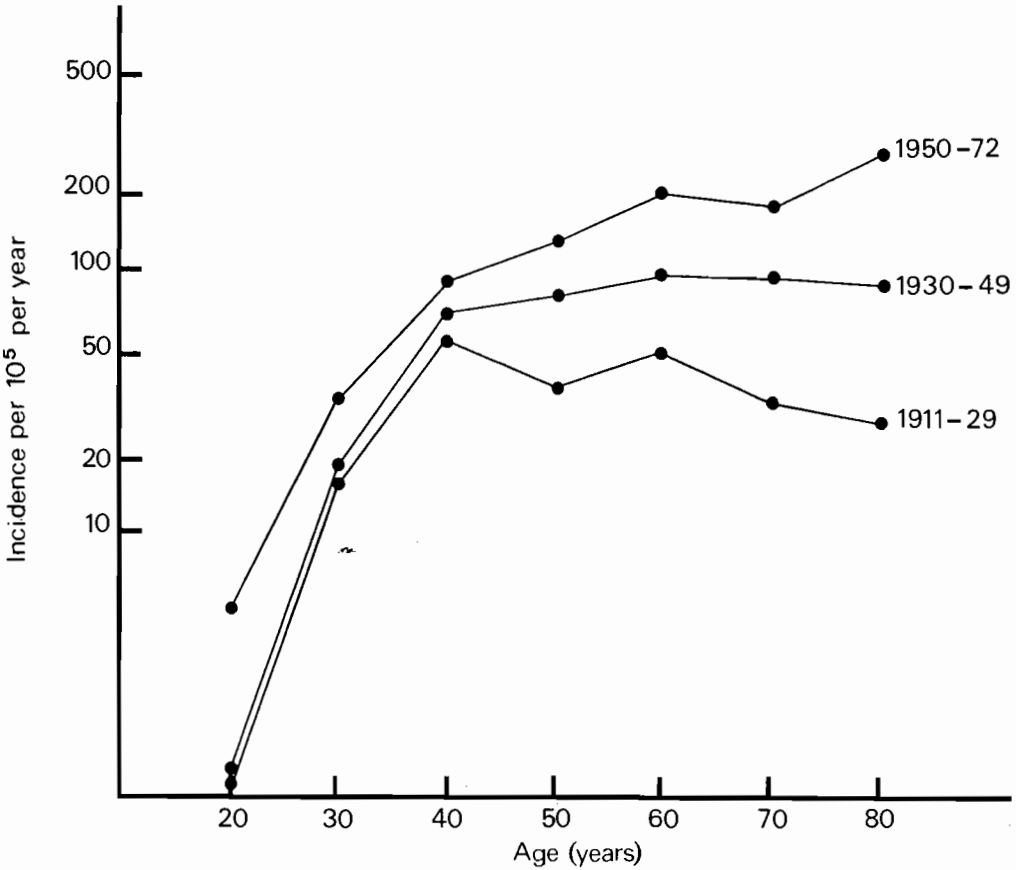
Fig. 2.3 Age-specific incidence of breast cancer in Iceland for the three time periods 1911–29, 1930–49, 1950–72. From Bjarnasson et al. (1974).



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

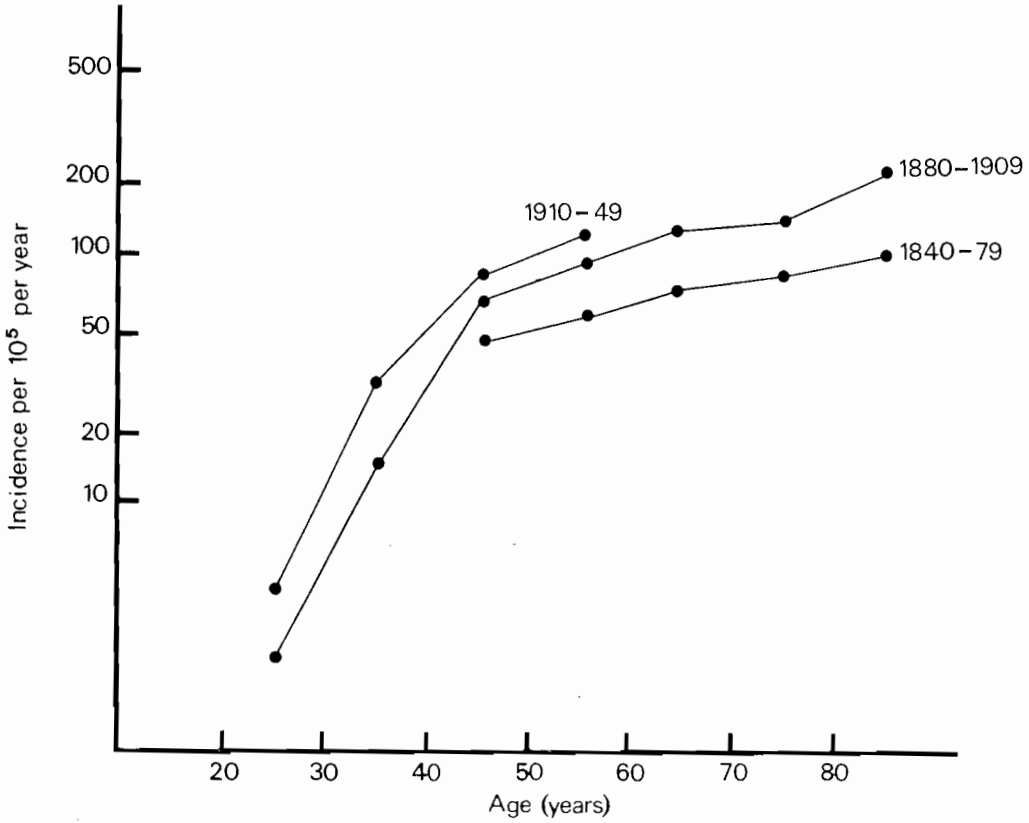
$$1\text{-}P(t) = \exp\{-\Lambda(t)\}, \tag{2.4}$$

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