Chapter 11
Studies based on routine data

Routine-data-based studies are characterized by the fact that data on both the exposure(s) and the outcome(s) of interest are obtained from routine data-collection systems (e.g., cancer registries, hospital registries, death notification, etc.). Thus, studies of this type can be carried out relatively quickly and cheaply, without the need to contact the study subjects. Their main limitation, however, is that the number of variables available from routine surveillance systems is generally limited.

Studies based on routine data can be carried out at an *individual* or at an *aggregated* level.

11.1 Individual level

Most routine data-collection systems collect data on personal attributes such as age, sex, place of birth, place of residence, occupation, etc. Cancer occurrence can then be examined in relation to these variables, either to search for patterns that may suggest etiological hypotheses or to confirm a specific hypothesis.

11.1.1 Place of residence

Routine-data-based studies of the variability in cancer incidence across the world have given an indication of the extent to which environmental factors are implicated in the causation of each cancer type (Example 11.1).

Examination of geographical variations in cancer incidence may provide important etiological clues and stimulate further investigations. Cancer of the oesophagus is one of the cancer sites with the most striking patterns of geographical variation. Its incidence varies sharply from one area to another, with foci of very high risk completely surrounded by areas of very low risk. The remarkable geographical variation in the risk of this cancer in central Asia (Example 11.2) stimulated the conduct of a series of cross-sectional and case–control studies in northern Iran with the specific aim of investigating possible risk factors. These studies were interrupted in 1978 by the civil unrest in the area but initial results suggested a strong association with a diet poor in vegetables and with the use of opium (Joint Iran–International Agency for Research on Cancer Study Group, 1977).

Part of the international differences in cancer risks may be genetic rather than environmental. But for many cancer sites, there are still marked variations in incidence within countries, even when their populations are genetically fairly homogeneous.

The distribution shown in Example 11.3 was initially assumed to be due

Example 11.1. *Data from population-based cancer registries located throughout the world were used to examine international variations in cancer incidence. For most cancer sites, there was a more than ten-fold variation between the highest and the lowest recorded incidence rates (Table 11.1).*

Table 11.1.
International variations in recorded incidence for selected cancer sites: the highest and lowest rate among all the population-based cancer registries included in *Cancer Incidence in Five Continents*, Vol. V. Rates per 100 000 pyrs, age-standardized to the world standard population.[a]

Cancer site (ICD 9 code)	Males			Females		
	Highest rate (H)	Lowest rate (L)[b]	H:L ratio	Highest rate (H)	Lowest rate (L)[b]	H:L ratio
Lip (140)	15.1	0.1	151	1.6	0.1	16
Tongue (141)	9.4	0.4	24	3.4	0.2	17
Mouth (143–145)	13.5	0.5	27	15.7	0.2	79
Nasopharynx (147)	30.0	0.3	100	12.9	0.1	129
Pharynx (146, 148–149)	31.3	0.4	78	4.3	0.2	22
Oesophagus (150)	29.2	1.2	24	12.4	0.3	41
Stomach (151)	82.0	3.7	22	36.1	3.0	12
Colon (153)	34.1	1.8	19	29.0	1.8	16
Rectum (154)	21.5	3.0	7	13.4	1.3	10
Liver (155)	34.4	0.7	49	11.6	0.4	29
Larynx (161)	17.8	2.2	8	2.7	0.2	14
Lung (162)	110.0	5.8	19	68.1	1.2	57
Melanoma of skin (172)	30.9	0.2	155	28.5	0.2	143
Breast (174/175)	3.4	0.2	17	93.9	14.0	7
Cervix uteri (180)	–	–	–	83.2	3.0	28
Ovary (182)	–	–	–	25.8	3.3	8
Prostate (185)	91.2	1.3	70	–	–	–
Testis (186)	8.3	0.6	14	–	–	–
Bladder (188)	27.8	1.7	16	8.5	0.8	11
Nervous system (191–192)	9.7	1.1	9	10.0	0.8	13
Non-Hodgkin lymphoma (200, 202)	11.4	1.5	8	8.7	0.9	10
Hodgkin's disease (201)	4.8	0.5	10	3.9	0.3	13

[a] Data from Whelan *et al.* (1990).

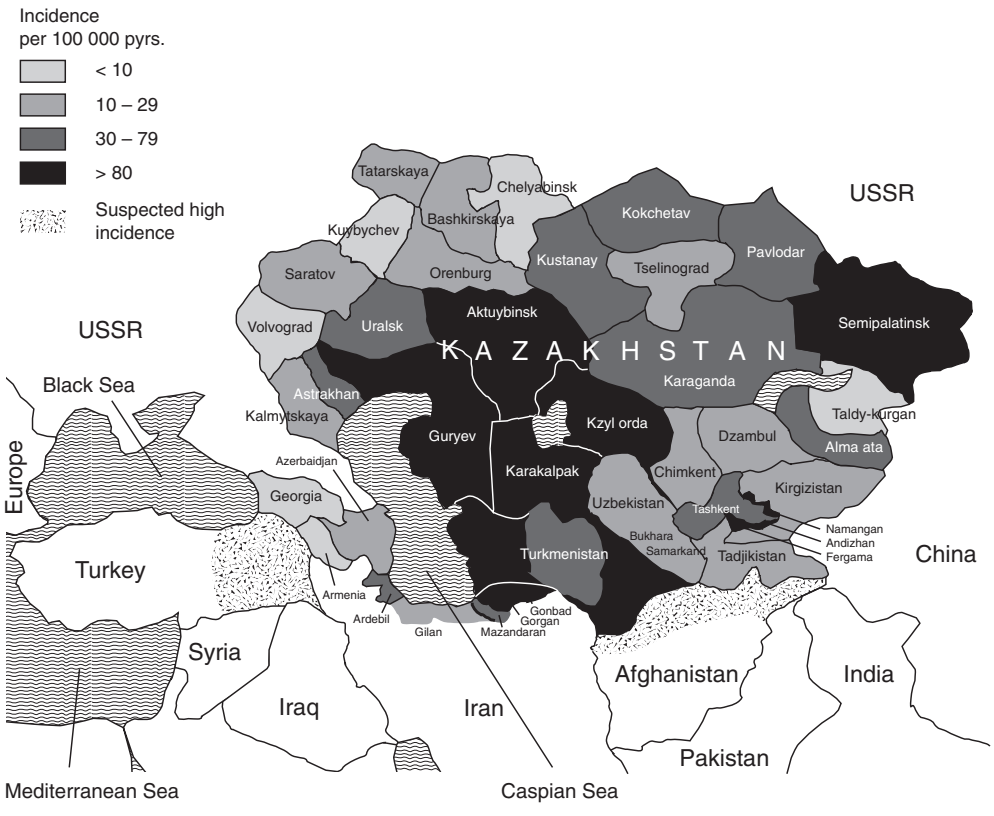
[b] Rates based on less than 10 cases were excluded.

to consumption of home-produced apple cider. Normandy and Brittany are the only French provinces where apple cider is produced in considerable quantities and this beverage is largely consumed in the provinces themselves (Tuyns *et al.*, 1983). Further research has shown that all types of alcoholic beverage appear to increase the risk of oesophageal cancer.

11.1.2 Place of birth and ethnicity

People who migrate from one country to another have lifestyle characteristics that are a combination of those prevailing in the host country and those from their homeland. Thus, evidence of a gradient of increasing, or decreasing, risks between population of origin, migrants and the

Example 11.2. *Data from local cancer registries were used to examine the geographical distribution of oesophageal cancer in central Asia (Muñoz & Day, 1996). The results from this analysis showed that the incidence of this cancer was extremely high in an area encompassing Kazakhstan, Uzbekistan and Turkmenistan, the north-east of Iran and northern Afghanistan (Figure 11.1).*



host population can suggest or confirm the importance of environmental factors over genetic factors in the etiology of a particular cancer.

This approach can be refined by adding a time component to it. The degree of cultural integration into the host country can be measured indirectly by information on age at migration and/or time since migration.

In Example 11.4, European immigrants were at lower risk of dying from melanoma than the Australian-born, particularly southern Europeans. This finding may reflect the protective effect of darker complexions. The data also suggested that migration in childhood was associated with a higher risk than migration later in life, but it is difficult to separate this effect from that of duration of stay (i.e., migration at an early age was inevitably associated with a longer stay than migration at an older age).

Figure 11.1.
Oesophageal cancer among men in central Asia. Incidence rates age-standardized to the world standard population (reproduced, by permission of Oxford University Press, from Muñoz & Day, 1996).

Example 11.3. *Routinely collected mortality data were used to examine the geographical distribution of oesophageal cancer in Brittany and the Normandy departments of Calvados and Orne (Figure 11.2). Although rates in most cantons were similar to the average rate in France, in some cantons in eastern Brittany and north-western Orne, mortality was five to ten times higher (Tuyns & Vernhes, 1981).*

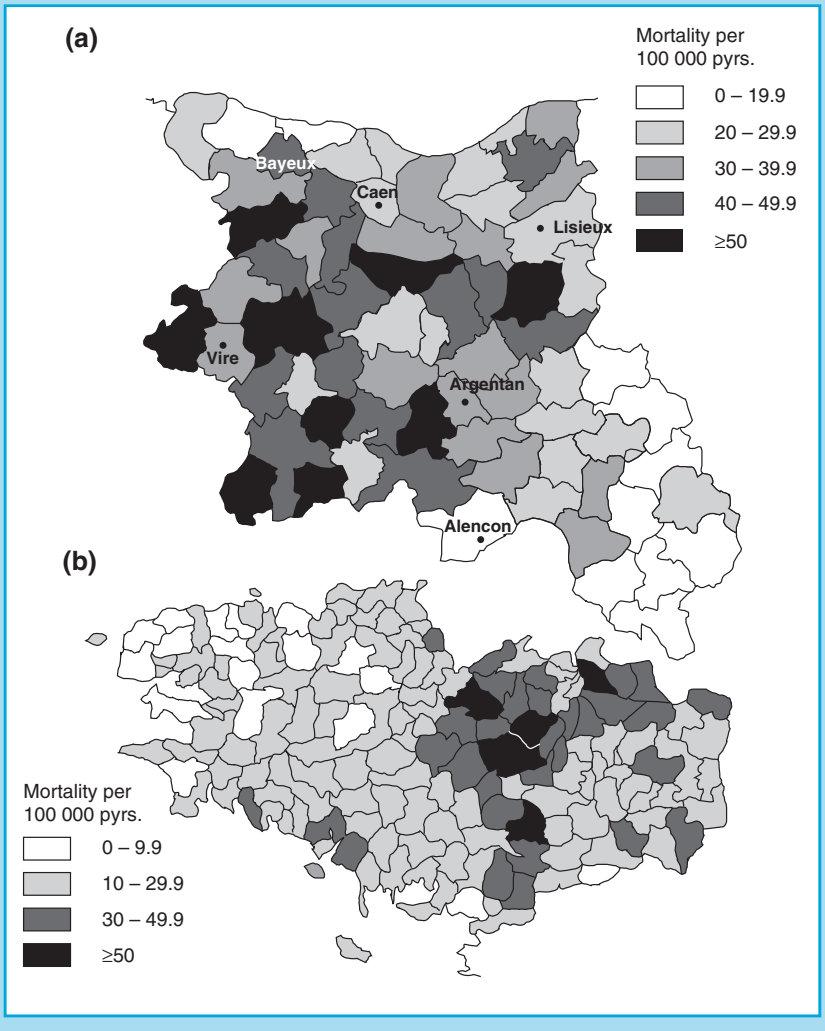


Figure 11.2.
Age-standardized mortality rates for cancer of the oesophagus among males, by canton in (*a*) the Normandy departments of Calvados and Orne (reproduced with permission from Tuyns & Vernhes, 1981 © Masson Editeur, 1981), and (*b*) Brittany, 1958–66 (modified from Tuyns & Massé, 1973).

If information on both place of birth and ethnicity is available, it is possible to distinguish first-generation immigrants (born in the country of origin) from their children, often born in the host country, who are considered as second-generation immigrants. This distinction provides another useful indicator of the likely degree of assimilation by the immigrants of the lifestyle characteristics of the host country, which tends to be more marked in the second than in the first generation.

Example 11.4. *Data on deaths registered in Australia during the period 1964–85 were obtained from the Australian Bureau of Statistics to examine mortality from malignant melanoma of the skin in immigrants compared with Australian-born individuals, and to investigate changes in risk with age at arrival and duration of stay. Each death record contained information on the following items: sex, country of birth, duration of stay in Australia, year of death, age at death, and cause of death (Khlat* et al.*, 1992). Some of the results from this study are shown in Figure 11.3.*

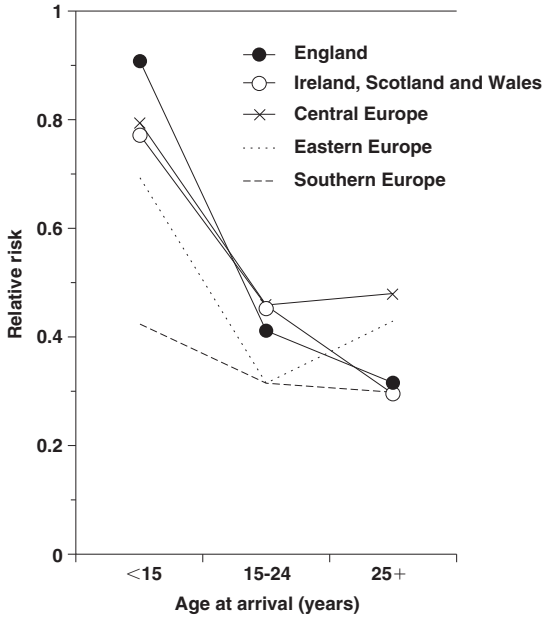


Figure 11.3.
Relative risks (RRs) of mortality from melanoma in male European immigrants to Australia, by region of birth and age at arrival compared to Australian-born (taken as the baseline group: RR=1); Australia, 1964–85 (reproduced with permission from Khlat *et al.*, 1992).

In Example 11.5, there was a consistent pattern: the risk for all four cancer sites converged towards the risk in the host country. For stomach and liver cancer, both of which were more common in Japan than among whites in the United States, men who were born in Japan but migrated to California had considerably lower risks of death than men of the same age in Japan. But risks in the second generation were still lower. In contrast, the risk of colon and prostatic cancers, which were common in California but rare in Japan, rose to approach that of white men in California with migration. These findings clearly show the importance of environmental factors over genetic factors linked to ethnicity. There was still a residual difference between rates in second-generation immigrants and rates in California. These residual differences might reflect differences in genetic susceptibility to these malignancies or, alternatively, they might indicate that second-generation immigrants maintained some of the 'traditional' Japanese lifestyle.

Example 11.5. *In a study of Japanese migrants living in California, the mortality from common forms of cancer in first- and second-generation migrants was compared with the corresponding rates for California and Japan (Buell & Dunn, 1965). Figure 11.4 shows results for four of these common cancer sites: stomach, liver, colon and prostate.*

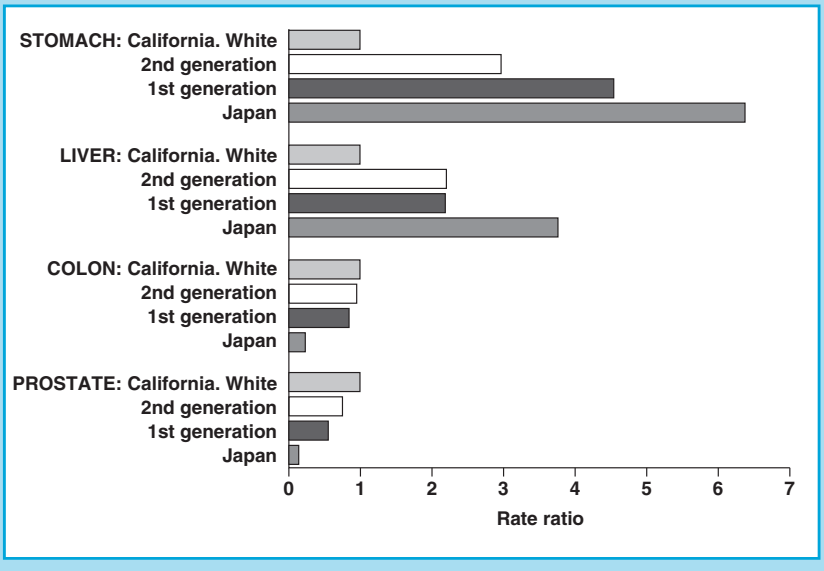


Figure 11.4.
Age-adjusted mortality rate ratios (RRs) for cancer of the stomach, liver, colon and prostate among Japanese men in Japan, and first- and second-generation Japanese immigrants to California, compared with white men in California (taken as the baseline group: RR=1); Japan, 1958–59 and California, 1956–62 (data from Buell & Dunn, 1965).

Any differences in cancer risk between migrants and those who remained in their country of origin must be interpreted cautiously, however, since migrants are usually a self-selected group not representative of the population of their country of origin. Migrants are also likely to differ from the host population in a number of demographic and socioeconomic characteristics that should be taken into account when comparing risks. In Example 11.6, most of the migrants originated from the southern part of Italy, which had lower mortality from pancreatic cancer than the country as a whole.

In routine-data-based studies, migrants are identified on the basis of place or country of birth and ethnicity and, when information on these two variables is not available, on the basis of name analysis. More rarely, information on language and religion has also been used. It should, however, be kept in mind that these approaches do not yield similar results. For instance, some of the people born in the Indian subcontinent who migrated to England and Wales are, in fact, of Caucasian ethnicity. Thus, analyses exclusively based on country of birth will include first-generation migrants regardless of their ethnicity. By contrast, analyses exclusively based on ethnicity will include migrants of a particular ethnic group regardless of whether they are of first or of subsequent generations.

Example 11.6. *Mortality from cancer of the pancreas in Italian migrant men was compared with mortality in their country of birth (Italy) and with mortality in eight host countries. Most of the Italian migrants originated from southern Italy. In order to attempt to allow for selection bias, relative risks (RR) were examined for both southern Italy and the whole country. Figure 11.5 shows that, for instance, mortality from pancreatic cancer was much lower in southern Italy (RR = 0.38) and in the whole of Italy (RR = 0.65) than in Canada (taken as the baseline: RR = 1). However, the mortality in Italian migrants to Canada was close to that of the host population (RR = 0.93) (Balzi* et al.*, 1993).*

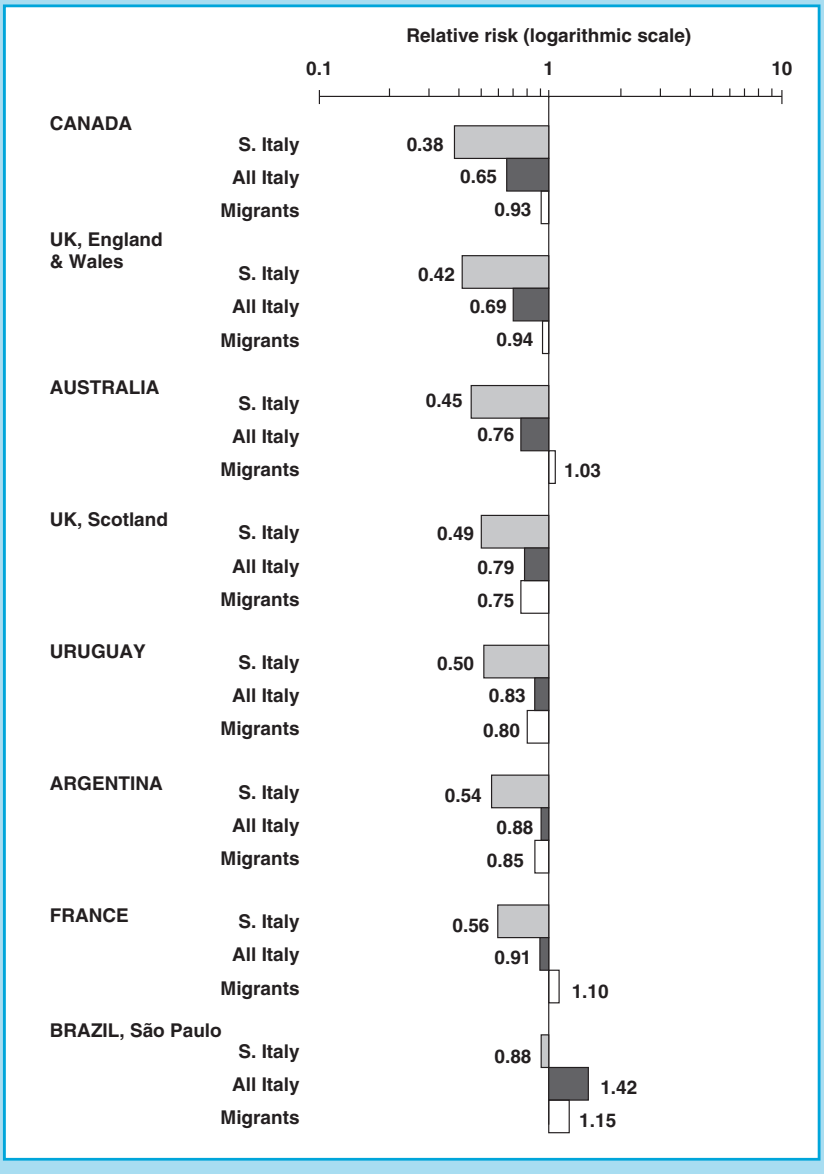


Figure 11.5.
Relative risk (RR) of mortality from pancreatic cancer in Italy (South, all Italy), and in eight male Italian immigrant populations compared with the host country populations (each taken as the baseline group: RR=1) (reproduced with permission from Balzi *et al.*, 1993).

Example 11.7. *Death records from residents in São Paulo County (Brazil) during 1978–82 were examined to investigate socioeconomic differentials in cancer mortality. The death certificates contained information on sex, age, place of birth and residence, cause of death, marital status, education and last occupation of the deceased. Information on education was provided by the family of the deceased, as years of schooling, recorded in five categories: less than 1, 1 to 8, 9 to 11, 12 or more, unknown (Bouchardy* et al.*, 1993). Table 11.2 shows educational differentials in mortality in females from cancer at selected sites.*

Table 11.2.
Mortality from selected cancer sites by years of schooling, for females aged 35–64 years in São Paulo (Brazil), 1978–82.[a]

Cancer site (ICD-9 code)	No. of deaths	Educational level unknown (%)	Odds ratio by years of education[b]				χ^2 test for trend[c]
			<1[d]	1–8	9–11	>11	
Stomach (151)	691	6.2	1.0	0.7	0.3	0.3	43.5; P<0.001
Colon (153)	338	7.1	1.0	1.4	2.1	2.2	13.3; P<0.001
Breast (174)	1744	5.9	1.0	1.6	2.4	2.6	77.5; P<0.001
Cervix uteri (180)	645	7.2	1.0	0.7	0.4	0.2	50.2; P<0.001

[a] Data from Bouchardy *et al.* (1993)

[b] Controls are all cancers except the one under investigation (see Section 11.1.6 for discussion of the method of analysis). Odds ratios are adjusted for age, civil status, and country of birth.

[c] After exclusion of subjects with unknown level of education.

[d] Taken as the baseline category.

11.1.3 Socioeconomic status and occupation

Analyses of cancer risks by socioeconomic and occupational groups have also provided important insights into the epidemiology of cancer.

Example 11.7 shows marked associations between mortality and educational level. The gradient was positive for breast and colon cancers but negative for stomach and cervical cancers (Table 11.2). These gradients are similar to those found in western countries.

In Example 11.8, men engaged in outdoor occupations had the highest proportional incidence of lip cancer (Table 11.3). These findings suggest that sun exposure may be important in the etiology of this cancer.

People with different occupations tend to have different lifestyles as well as different occupational exposures. Thus, differences in their cancer risks might also give etiological clues in relation to factors that are not directly related to the occupation (as in Example 11.9).

Table 11.4 shows a strong social-class gradient in the risk of married women dying from cervical cancer in England and Wales. Wives of men in social class I (highest social class) experienced a mortality rate which was only 34% the rate for all married women. In contrast, the mortality of wives of men in social class V (lowest social class) was 81% higher than that of all married women. Within each social class, however, the highest rates were among wives of men in occupations involving travel

Example 11.8. *Cancer registries in England and Wales collect data on the occupation of the cancer cases at the time of their diagnosis. These data were examined to identify occupational groups associated with high incidence of lip cancer (OPCS, 1978). The results are shown in Table 11.3.*

Table 11.3.
Lip cancer incidence in men aged 15–74 years, England and Wales, 1968–69, by occupation, for those occupations associated with statistically significant increases in proportional incidence ratios.[a]

Occupation	Observed number of cases	Proportional incidence ratio (PIR)[b]
Farmers, farm managers, market gardeners	23	2.18
Agricultural workers	35	5.26
Bricklayers, tile setters	14	2.39
Construction workers	16	2.35

[a] Data from OPCS (1978)

[b] Age-adjusted. $P < 0.01$ for all the PIRs shown.

(See Example 11.14 for illustration of calculation of proportional incidence ratios).

and absence from home for long periods, a situation known to be associated with high risk of venereal diseases among men. This pattern supported the hypothesis that cervical cancer might be a sexually transmitted infection long before any infectious agent was identified (Beral, 1974).

In interpreting differences in incidence or mortality between socioeconomic groups, it must be remembered that health itself may determine entry into a specific group. For instance, people with poor health are usually forced to work in physically less demanding jobs and more demanding jobs selectively include only those in good health.

11.1.4 Time trends

Information on changes in cancer risk over time can generate etiological hypotheses or support suspected associations between risk factors and disease. Moreover, while the existence of geographical variation in incidence between populations may be explained by genetic differences, changes in incidence in single populations imply the introduction or disappearance of environmental risk factors much more clearly.

In Example 11.10 the incidence of papillary carcinoma of the thyroid increased steadily in the younger age groups in Connecticut during the 40-year study period, resulting in a peak at ages 25–44 years (Figure 11.6). This sudden increase in incidence in the USA followed the widespread use of radiation therapy for 'enlarged thymus' and other benign conditions of the head and neck among children and adolescents between the early 1920s and the late 1950s (Pottern *et al.*, 1980). No similar increase in the incidence of this cancer was observed in populations where the use of this therapy had not been common. This example provides a good illustration of a large increase in the incidence of a particular cancer due to the introduction of a specific exposure.

Example 11.9. *Routinely collected mortality data for England and Wales were used to examine risk of mortality from cervical cancer among married women by husband's occupation and social class (Beral, 1974). (In this country, death registrars are required to enter the husband's occupation on the death certificate of a married woman or widow; social class for married women is determined by the husband's occupation.) The results are shown in Table 11.4.*

Table 11.4.
Cervical cancer mortality among married women by husband's social class and occupation, England and Wales, 1959–63.[a] (Only data for occupations with the lowest and highest mortality levels are shown in the table to illustrate the range of risks within each social class.)

Social class	Occupation of husband	Standardized mortality ratio (SMR, %)[b]
I	All occupations	34
	Clergymen	12
	Scientists	17
	Civil engineers	60
II	All occupations	64
	Teachers	30
	Senior government officials	40
	Publicans and innkeepers	120
	Lodging house and hotel keepers	150
III	All occupations	100
	Clerks of work	40
	Clerks	64
	Crane and hoist operators	159
	Drivers of road goods vehicles	168
IV	All occupations	116
	Shopkeepers and assistants	71
	Gardeners and groundsmen	98
	Fishermen	257
	Deck and engine-room ratings, barge and boatmen	263
V	All occupations	181
	Office and window cleaners	95
	Labourers	222

[a] Data from Beral (1974)

[b] Age-specific rates for all married women in England and Wales taken as the standard.

In Example 11.11, there was no consistent change in the risk of testicular cancer for successive generations of men born between 1880 and 1920. However, for generations born since then, the risk increased steadily, except for men born in Denmark, Norway and Sweden during 1930–45 (the years around the Second World War). The reasons for the marked increase in incidence are not known. The rise has occurred mainly in industrialized countries and in upper socioeconomic groups, suggesting that lifestyle factors associated with affluence may be responsible. Paradoxically, the small

Example 11.10. *Data on all thyroid cancer cases incident between 1935 and 1975 were extracted from the Connecticut Tumor Registry (USA) to examine changes over time in the rates of this cancer and its four main histological types (Pottern* et al.*, 1980). The temporal changes for papillary carcinoma, one of the four histological types, are shown in Figure 11.6.*

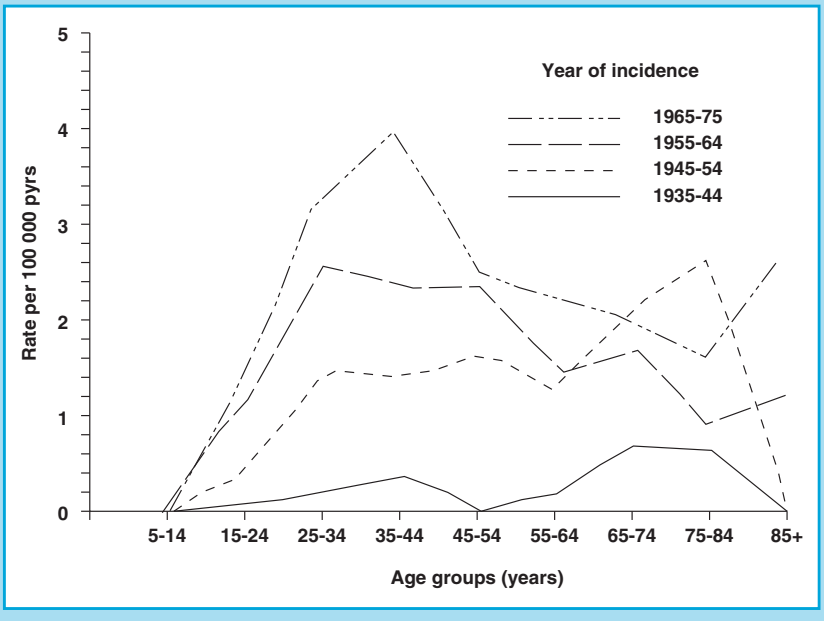


Figure 11.6.
Age-specific incidence rates for papillary carcinoma of the thyroid in women by calendar period, Connecticut (USA), 1935–75 (reproduced with permission from Pottern *et al.*, 1980).

decline in risk for generations born during the Second World War occurred in Denmark, Norway and Sweden, which were apparently less affected by the war than Poland, the former East Germany and Finland.

Time trends can also be used to assess and monitor the effectiveness of cancer control activities such as mass screening programmes. In Example 11.12, there was a close relationship between the decline in incidence of cervical cancer in each country and the degree of coverage by organized mass screening programmes. The decline in incidence was most marked in Finland and Iceland, where national screening programmes were initiated in the early 1960s. The fall was less marked in Sweden, where the programme was introduced more gradually, and in Denmark where only 40% of the population lived in areas with organized mass screening. There was no obvious decline in the incidence of cervical cancer in Norway, the only country which did not have an organized programme (except in one county).

11.1.5 Record linkage studies

Linkage of cancer registry records with records from other sources such as census data, mortality data, company records, hospital admission records, etc., has been undertaken to investigate risk factors for a large

Example 11.11. *Data from six European countries (Denmark, Finland, the former German Democratic Republic, Norway, Poland and Sweden) where cancer registration dates back to the 1940s or 1950s were used to examine long-term time trends in the incidence of testicular cancer. These registries cover a total population of about 76 million people. A total of 30 908 incident cases of testicular cancer diagnosed during 1945–84 in men aged 20–84 years were identified (Bergström* et al.*, 1996). Figure 11.7 shows cohort trends (i.e., trends by year of birth of the men) in each country.*

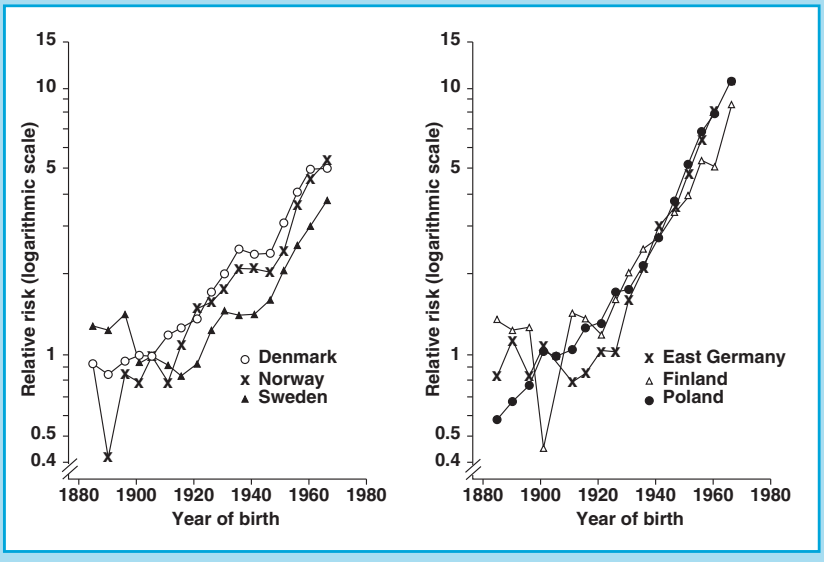


Figure 11.7.
Relative risk (RR) of developing testicular cancer by country and year of birth. (Men born in the study countries between 1900 and 1909 were taken as the baseline category: RR=1) (reproduced with permission from Bergström *et al.*, 1996).

number of cancers such as occupational and reproductive-related cancers (see Section 2.9.3).

In Example 11.13, a cohort of 73 917 men in Denmark was identified retrospectively from hospital discharge and pathology registries as having had vasectomy during 1977–89. They were passively followed from the time of their operation to 31 December 1989, when their vital and migration status was assessed by linkage with the population registry. The occurrence of cancer among cohort members was ascertained by linkage with the national cancer registry.

11.1.6 Analysis

Studies based on routine data conducted at an individual level may be regarded as cohort studies in which a group of people, or cohort, is followed up in time. For instance, comparison of cancer incidence or mortality across different occupational groups may be regarded as a study of groups of people with different occupational 'exposures' who were followed up in time and their cancer experience compared. Thus, the analysis of these studies is similar to the analysis of any other cohort study and the methods are basically those described in Chapters 4 and 8. The analysis is based on calcula-

Example 11.12. *Data from five Nordic national cancer registries were extracted to examine time trends in the incidence of cervical cancer in each country in relation to their screening activities (Hakama, 1982).*

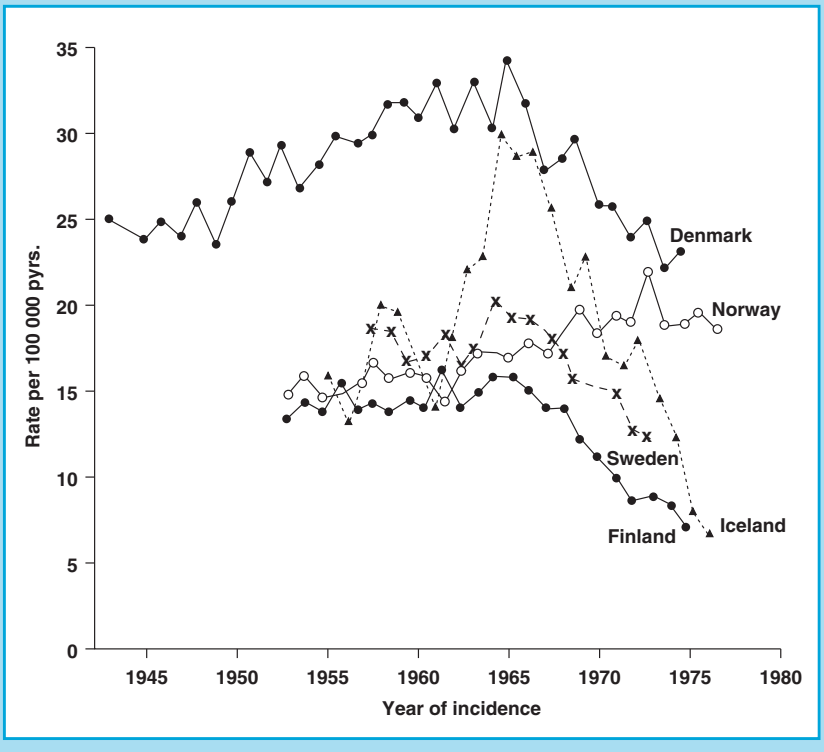


Figure 11.8.
Trends in age-adjusted incidence of cervical cancer in the five Nordic countries, 1943–78 (reproduced with permission from Hakama (1982)).

tion of *rates* as measures of disease occurrence, and of *rate ratios* and *rate differences* as measures of effect. The analysis of time trends by calendar period and birth cohort was considered in Section 4.3.2.

Calculation of rates requires information on *person-time at risk* (i.e., denominators) for each of the groups of interest. Population estimates to allow these calculations are not always available, however. For instance, although national statistical offices in many countries are able to provide population estimates for the whole country and for relatively large geographical areas, they may be unable to provide data for smaller geographical areas. Moreover, cancer registrations and death certificates may contain information on many variables for which data are not collected by the census (or any other population enumeration system). For instance, cancer registries may collect information on ethnicity, but proper denominators will not be available if information on that variable is not collected by the census. In other situations, denominators are available from the census but case-finding is so incomplete (for instance, if on a large proportion of death certificates, data on occupation of the deceased is missing) that the numerator data are not comparable with the available denominator data.

Example 11.13. *A study was conducted in Denmark to assess whether vasectomy increases the risk of cancers of the testis and prostate. The study was based on a computerized record linkage between four population-based registries. The linkage was done by using the unique personal identification number allocated to every resident in Denmark.*

The Danish Hospital Discharge Registry has recorded all hospital admissions in the country since 1977, recording dates of admission and discharge, diagnosis, and operations. Some hospital departments, however, regarded vasectomy as an outpatient service and did not inform the registry.

Pathology registries were set up in some counties to record all specimens analysed in hospital pathology departments within the county. As the tissue removed at vasectomy is routinely sent for pathological examination, in counties where pathology registries exist it is possible to identify all vasectomized men (whether treated as inpatients or as outpatients) from their files.

The Danish Central Population Registry records information about vital status and date of emigration or death of everyone who was alive in 1968 or who was born in or immigrated into Denmark thereafter.

The Danish Cancer Registry covers all cases of cancer in Denmark diagnosed since 1943, based on notifications from hospital departments, specialists, necropsy reports and death certificates.

In the present study, the Hospital Discharge Registry and the Pathology Registry were searched to identify all men who underwent vasectomy from 1977 to 1989. A total of 73 917 men were identified and their records were then linked with the Central Population Registry to provide information on their vital status on 31 December 1989 and dates of emigration or death. Cancer occurrence among the cohort members was identified by record linkage with the Danish Cancer Registry (Møller et al.*, 1994).*

The traditional answer to situations where suitable denominators are not available has been to calculate *proportional incidence or mortality ratios* (see Section 4.3.5). This is illustrated in Example 11.14.

Proportional incidence (or mortality) ratios are usually age-adjusted by using the indirect method of standardization in a way similar to that used in the calculation of standardized incidence (or mortality) ratios, except that a set of age-specific proportions rather than age-specific rates is taken as the standard in the calculation of expected numbers (see Section 4.3.3).

It can be shown that *odds ratios* may be more appropriate than proportional incidence (or mortality) ratios in situations where no proper denominators are available (Miettinen & Wang, 1981). Odds ratios can be calculated to estimate the risk of a particular cancer ('cases') relative to other cancer sites ('controls') in a population group A ('exposed') compared with another population group B ('unexposed').

Example 11.14. *Suppose that an investigator is interested in examining the incidence of lung cancer among men working in the printing industry. Information on the occupation of all new lung cancer cases that occurred in a particular region A during the years 1970–74 can be obtained from the local population-based cancer registry but no population estimates by occupational group are available from the census. In this instance, it is not possible to calculate rates since there are no denominator data. Instead, proportional incidence measures are calculated.*

Suppose that a total of 10 000 male incident cancers occurred in region A during the years 1970–74, of which 2000 were lung cancers. Thus, the proportion of male lung cancers in this region was

$$2000/10\,000 = 0.20$$

Suppose also that the total number of incident cancers among male printers during the same period (1970–74) was 100, of which 40 were lung cancers. The proportion of male lung cancers among printers was

$$40/100 = 0.40$$

We can calculate the ***number of lung cancer cases we would have expected (E) among the printers if they had the same proportion of lung cancers as the whole population of the region****. This would be equal to*

$$100 \times 0.20 = 20.$$

The ***proportional incidence ratio (PIR)*** *can be obtained by dividing the observed number of lung cancers among the printers (O) by the expected number (E):*

$$PIR = O/E = 40/20 = 2.0$$

(or, equivalently, the PIR can be calculated by dividing the proportion of lung cancers among the printers by the proportion of lung cancers in the region (i.e., PIR = 0.40/0.20 = 2.0).)

Thus the number of lung cancers observed among the printers was twice the number we would expect if they had the same proportion of lung cancers (out of all cancers) as the whole region.

In Example 11.15, the mortality data were examined as a series of case–control analyses. In each of these analyses, all cancers except the particular one under investigation were taken as the 'controls' and people born in England and Wales as the 'unexposed'.

Example 11.15. *In England and Wales, information on country of birth of the deceased has been entered on death certificates since 1969, and this information has been coded and included in national mortality files by the Office of Population Censuses and Surveys (OPCS). Information on ethnicity is not recorded in the death certificate but, for deaths from 1973 to 1985, OPCS undertook ethnic-origin coding for persons born in the Indian subcontinent. The coding was based on name analysis, exact place of birth and other items on the death certificate, and it separated the Indian-born into those of Indian ethnic origin, British ethnic origin and others. Data on ethnicity were not collected by the census at that time and therefore no proper denominators were available. Thus, age-adjusted odds ratios were calculated to estimate relative risks of mortality from different cancer sites in the migrants compared with people born and resident in England and Wales (Swerdlow* et al.*, 1995). Table 11.5 shows results from this analysis for certain cancer sites.*

Table 11.5.
Age-adjusted odds ratios of cancer mortality for selected sites in male migrants from the Indian subcontinent to England and Wales, by ethnic group, compared with males born and resident in England and Wales, 1973–85.[a]

Cancer site (ICD-9 code)	England and Wales born[b]		British ethnic, born in India		Indian ethnic, born in India	
	No.	Odds ratio	No.	Odds ratio (95% CI)[c]	No.	Odds ratio (95% CI)[c]
Oral (141, 143–145)	4564	1.0	24	1.7 (1.1–2.5)	30	2.2 (1.5–3.1)
Pharynx (except nasopharynx) (146, 148,149)	3440	1.0	22	2.1 (1.4–3.1)	55	5.5 (4.2–7.2)
Liver (155)	6177	1.0	33	1.7 (1.2–2.5)	88	5.0 (4.0–6.2)
Lung (163)	296 012	1.0	759	0.7 (0.7–0.8)	581	0.6 (0.6–0.7)

[a] Data from Swerdlow *et al.* (1995).

[b] Taken as the 'unexposed' baseline group.

[c] CI = confidence interval

Thus, for example, for lung cancer in Indian ethnic migrants, a 2 × 2 table was constructed as follows:

	Indian ethnic, born in India	England and Wales-born
Lung cancer ('cases')	*a* (581)	*b* (296 012)
Other cancer ('controls')	*c* (1737)	*d* (466 756)

OR = (*a*/*b*)/(*c*/*d*) = (581/296 012)/(1737/466 756)=0.5

(The methods described in Chapter 14 were used to adjust for age.)

A similar approach was used in Example 11.7 to examine educational differentials in cancer mortality in São Paulo (Brazil). Although data on educational level were collected by the census, the quality of these data was likely to differ from that in the death certificates since the information on the census forms was provided by the individuals themselves, whereas that on the death certificate was given by the relatives of the deceased. To overcome the lack of comparability between numerator and denominator data, odds ratios were calculated instead of rates.

A discussion of the relative merits of odds ratios versus proportional ratios is beyond the scope of this chapter. Suffice it to say that it can be shown that the odds ratio equals the rate ratio (the measure that would have been calculated if suitable denominators had been available) provided that the total incidence rate of 'other cancers' is similar in the two population groups, in other words, that the total incidence rate of 'other cancers' in the analysis is unrelated to the 'exposure' (Miettinen & Wang, 1981). No similar relationship exists between proportional ratios and rate ratios. Despite this advantage of the odds ratios, it should be noted that the calculation of odds ratios does not solve one important problem of the proportional ratios—the 'borrowing effect' occurring in populations with high incidence of a particularly common cancer, which inevitably leads to lower proportional incidence for other cancers. In other words, the proportion (or odds) for an individual cause may be high because the incidence for that cause is high or because the incidence for other major causes is low. These two situations can be distinguished only when proper denominator data are available (see Section 4.3.5).

11.1.7 Interpretation

Routine surveillance systems usually cover large catchment populations, sometimes of millions of people followed up over long periods of time. Thus, one of the main advantages of routine-data-based studies is that they allow the study of a very large number of people at a very low cost.

A major limitation, however, is the restricted range of variables collected by these systems, that often tend to be just proxy measures of more biologically relevant exposures. For instance, country of birth may act just as marker of environmental factors (e.g., diet, reproductive variables) for which data are not available in routine data systems.

Data quality

Potential data artefacts need to be considered when interpreting results from these studies. The observation of differences in recorded cancer incidence between populations does not necessarily reflect true underlying variations in cancer risks. Differences in recorded incidence may arise because of differences in health service access (including screening), diagnosis, and registration practices. In addition, they may be due to variations in the accuracy of enumeration of the population. Some of these issues were discussed briefly in Section 2.9 and Appendix A2.2.

An example is the difficulty in assessing rising trends in the recorded incidence of prostatic cancer in many western countries due to uncertainty about the effect of changes in diagnostic and registration practices. Strong parallels have been observed in the USA between time trends in the recorded incidence of localized prostate cancer and time trends in the use of transurethral resection of the prostate (Potosky *et al.*, 1990; Severson, 1990), a surgical procedure performed to relieve commonly occurring symptoms of benign pro-

static hyperplasia. An increasing use of this medical procedure seems to have led to an increased detection of clinically silent tumours that would otherwise not have been diagnosed. Inter-country variations in the use of this technique also seem to account for some of the difference between recorded incidence rates in Japan and the USA (Shimizu *et al.*, 1991).

Cancer incidence data are not available in many countries, and, if available, the registration scheme may not cover the whole population or the data collected may be incomplete. Mortality data are available for a much larger number of countries and for much longer periods. Mortality data, however, have the disadvantage of not directly reflecting incidence, particularly for cancers with a good prognosis. This is because mortality depends on both incidence and case fatality. Death rates closely parallel incidence rates only if the disease is fatal and if death occurs shortly after diagnosis. For example, death rates are a good indication of the magnitude of the incidence rates for lung cancer because this tumour has a high and early fatality, but not for testicular cancer, for which survival is relatively good. Moreover, the accuracy of the mortality data provided in death certificates is influenced by geographical differences and trends in diagnostic and certification practices as well as by changes in the frequency with which post-mortem examinations are carried out.

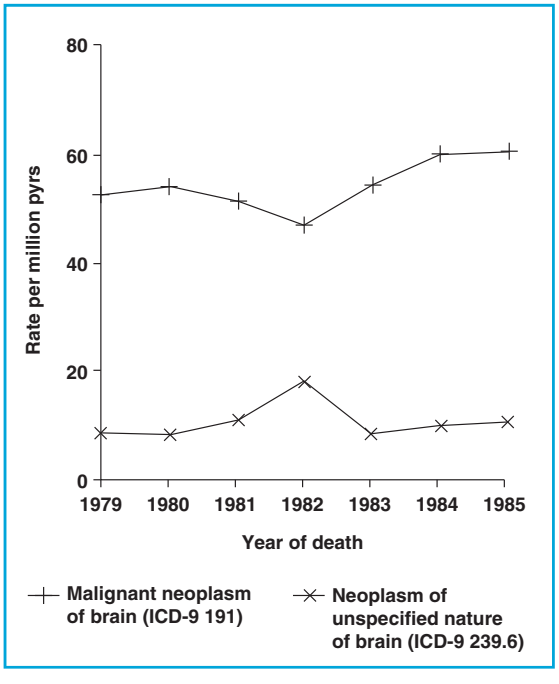


Figure 11.9.
Mortality from brain neoplasms, malignant and of unspecified nature, in males, England and Wales, 1979–85. Death rates age-standardized to the 1981 England and Wales population (reproduced by permission of Churchill Livingstone, from Swerdlow (1989)).

Routine mortality statistics in many countries are based on a single underlying cause of death. The rules for determining the underlying cause of death are primarily those of the *International Classification of Diseases* (see Appendix A.2.2), but national vital statistics offices may, on occasions, superimpose their own rules. For instance, the Office of Population Censuses and Surveys in England and Wales changed one of these rules (rule 3) in 1983, so that cancer is now more likely to be coded as the cause of death even if it is only recorded as a contributory cause in the death certificate (Ashley & Devis, 1992). This change in coding affected secular trends in recorded cancer mortality. It has been estimated that close to 50% of the increase in prostatic cancer mortality at ages 75–84 years in England and Wales from 1970 to 1990 could possibly be explained by changes in coding practices (Grulich *et al.*, 1995).

Many other factors, some of them less apparent, may affect recorded incidence and mortality data. Mortality rates for malignant neoplasms of the brain decreased greatly in England and Wales during the years 1981–82, but it is clear from examination of the data for neoplasms of unspecified nature of the brain that this decrease was largely, or entirely, an artefact. Most of the unspecified cases are allocated to malignancies of the brain after further enquiries are sent to the certifiers requesting clarification of the cause of death. The artefact shown in Figure 11.9 occurred because the number of enquiries referred to certifiers was reduced during a strike by registrars in 1981–82.

Comparability of the numerator and denominator data

Calculations of rates requires estimates of person-years at risk, according to the variables under study. These usually rely upon census data. It is essential that the variables in the census are defined, classified and coded in a similar way to the variables in cancer registration and mortality databases. However, even when the same criteria are used, individuals may still be classified differently in the two databases if the way the information was obtained differs. For instance, the information in the census is provided by the individuals themselves, whereas that on the death certificates is given by informants (usually relatives of the deceased). Informants may not be able to provide accurate information on many items such as occupation of the deceased, or may deliberately report a more prestigious occupation (OPCS, 1978).

Selection bias

People who migrate are not generally a representative group of their population of origin. For instance, migrants tend to be healthier than those who stay in the home country. This 'healthy migrant effect' will affect comparisons with home population risks. A similar bias may occur in analyses by occupation (the 'healthy worker effect') and this will affect comparisons with the general population (see Sections 8.2.2 and 13.1.1).

Another source of bias, that is more difficult to detect, results from changes in 'exposure' which are related to the disease event itself. For example, migrants may return to their country of origin soon before death or people move jobs as a consequence of being diagnosed with a particular condition. In these situations, risks in the host country or in the job held before diagnosis will be under-estimated.

Confounding

One of the main limitations of routine-data-based studies is that information on important confounding factors (with the exception of age and sex) is generally not available. For instance, the high risks of lung cancer found in some occupational groups are not due to occupational exposures but to high prevalence of smoking. Unfortunately, data on smoking habits are rarely collected by routine surveillance systems.

Final remarks

In summary, the study of variations in cancer incidence and mortality by place of residence and birth, ethnicity, socioeconomic status and over time is a valid and useful exercise, provided that the investigator has a thorough knowledge of the way data are collected and processed so that all possible sources of data artefacts, bias and confounding are considered in the interpretation of the findings.

11.2 Aggregated level (ecological studies)

The routine-data-based studies considered thus far share the characteristic that the observations made pertain to individual subjects. It is, how-

ever, possible to conduct research at a group level rather than at the individual level. Studies which involve investigating the frequency of disease (or of any other outcome of interest) in relation to the level of exposure in several groups of individuals (or in the same group over different periods of time) are called ecological studies. In this type of study, it is not possible to link the exposure of a particular individual with his or her outcome. Thus, the group rather than the individual is the unit of observation and analysis. The groups may be defined in a large number of ways, according to place of residence, place of birth, socioeconomic status, occupation, etc.

Ecological studies are frequently used as a first step in investigating a possible exposure–outcome relationship, because they can be performed quickly and inexpensively by using readily available information. Exposure data may be available from governmental and private organizations which routinely collect data on demographic, environmental and lifestyle variables. Disease rates may be available from vital statistics offices, surveillance programmes or disease registries (e.g., cancer registries).

In Example 11.16, it is not possible to link the ovarian cancer mortality experience of any individual woman with her family size because the only pieces of information available were an estimate of the average family size and an estimate of the average level of mortality from ovarian cancer for each country included in the analysis. Thus, the country rather than the individual was the unit of study.

Similar comparisons can be performed between changes over time in the average exposure level and changes in the disease rate for a single population (as in Example 11.17).

Ecological studies may be the most appropriate design to study exposures that are easier to define and measure at a population rather than at an individual level. This is the case with many environmental exposures such as air pollution, water quality and ultraviolet radiation (Example 11.18).

Ecological studies are also useful for monitoring the effectiveness of population interventions such as health education campaigns (e.g., anti-smoking campaigns), immunization programmes and mass screening programmes.

11.2.1 Analysis

Ecological studies differ from individual-based epidemiological studies in that the ‘exposed’ and ‘unexposed’ individuals in each of the populations are not actually identified. Thus, it is not possible to measure the strength of an association between exposure and outcome by using any of the approaches described in Chapter 5 for individual-based studies.

Moreover, in contrast to other epidemiological studies, the outcome measure in an ecological study is usually a quantitative variable (e.g., mortality rate) rather than a binary one (such as ‘diseased’ versus ‘non-diseased’). The

Example 11.16. *Routinely collected data on ovarian cancer mortality in twenty countries were examined in relation to their average family size as estimated from various demographic surveys. There was a clear inverse relationship between mortality from this tumour and average completed family size, i.e., total number of children per woman at the end of her reproductive life (Figure 11.10) (Beral* et al.*, 1978). This finding suggested that pregnancy protected against ovarian cancer, a hypothesis that has been confirmed in many case–control and cohort studies.*

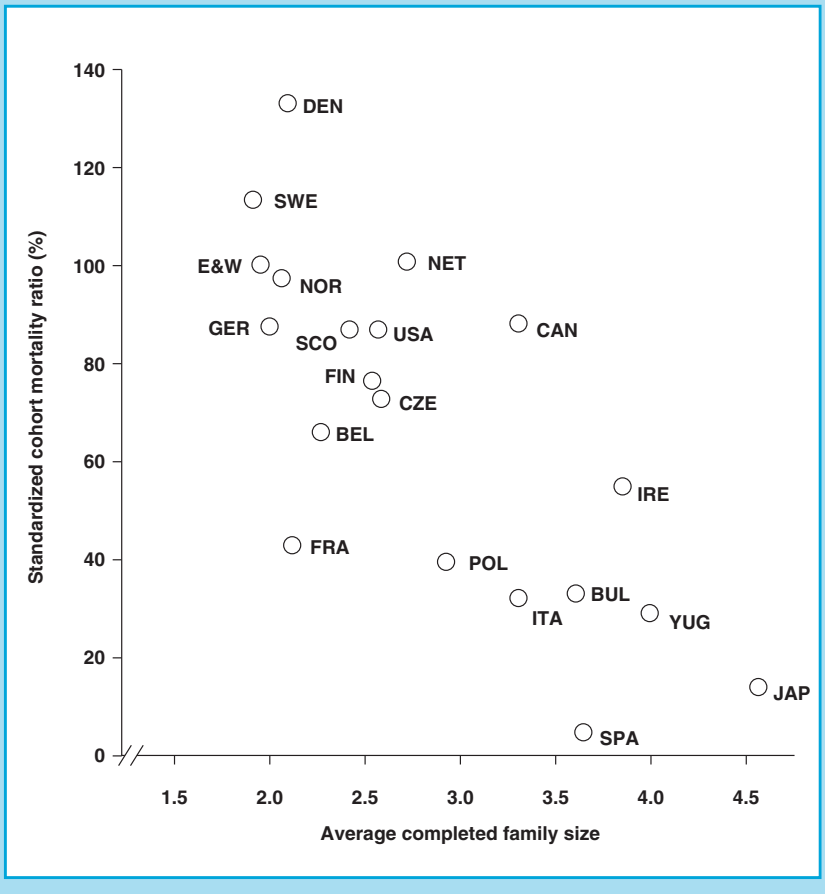


Figure 11.10.
Scattergram showing the relationship between age-standardized mortality ratios from ovarian cancer and average completed family size for women born around 1901 in 20 countries. The age-specific mortality rates in England & Wales were used as the standard, i.e. the mortality ratio for England & Wales = 100. (Reproduced with permission from Beral *et al.*, 1978. © by The Lancet Ltd, 1978). BEL=Belgium; BUL=Bulgaria; CAN=Canada; CZE=former Czechoslovakia; DEN=Denmark; E&W=England and Wales; FRA=France; FIN=Finland; GER=former West Germany; IRE=Ireland; ITA=Italy; JAP=Japan; NET=Netherlands; NOR=Norway; POL=Poland; SCO=Scotland; SPA=Spain; SWE=Sweden; USA=United States of America; YUG=former Yugoslavia.

exposure variable also tends to be measured on a quantitative scale. Even qualitative variables become quantitative when averaged for a population: sex is a binary variable, but the proportion of a population that is male (or female) is quantitative.

A statistical measure called a *correlation coefficient*, denoted by *r*, has been widely used to measure the strength of associations between exposure and outcome in ecological studies. Another, more appropriate, approach is to fit a *regression* line which predicts incidence or mortality as a function of the exposure level. These two statistical methods are briefly introduced below.

Introduction to regression and correlation

When investigating the relationship between two quantitative variables, the first step should always be to plot the data (see Section 3.3.3). The required plot is called a *scattergram* (or *scatter diagram*), in which the vertical (or y) axis refers to the outcome variable and the horizontal (or x) axis to the exposure variable (as in Figures 11.10–11.12). In such a graph, each unit under consideration is represented by a point.

Example 11.17. *In the study on ovarian cancer mortality and family size described in Example 11.16, it was shown that mortality from ovarian cancer among successive generations of women born in England and Wales, and in the USA was closely related to their completed family size (Figure 11.11).*

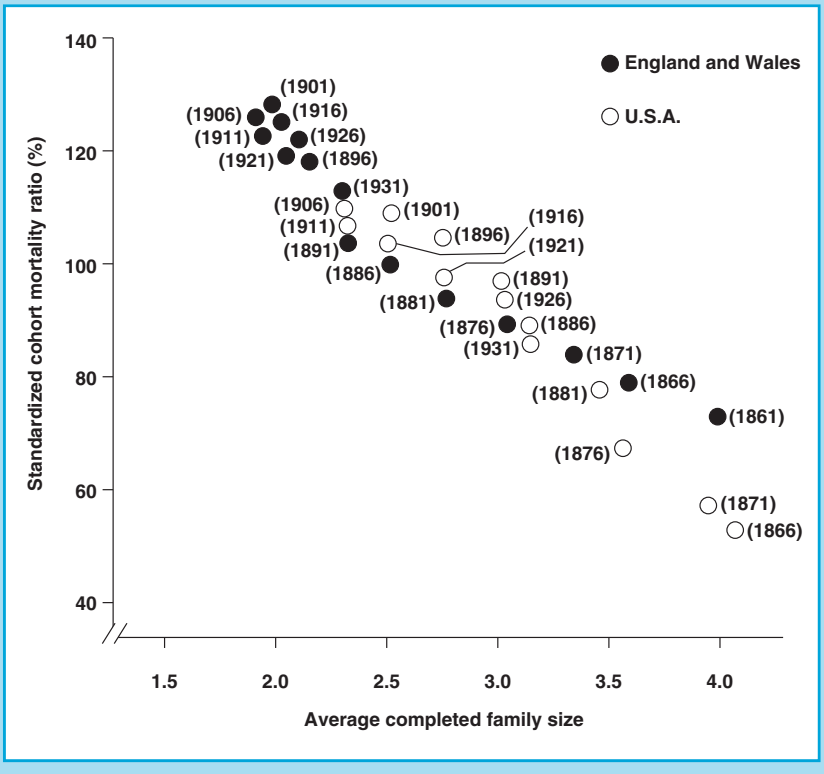


Figure 11.11.
Scattergram showing the relationship between age-standardized mortality ratios from ovarian cancer and average completed family size for different generations of women born in England and Wales and the USA. The average age-structure of the combined population of England & Wales and the USA from 1931–73 was taken as the standard. Thus, a ratio of 140 means that the ovarian cancer mortality of the cohort is 40% higher than the average for women in England & Wales and the USA. (The mid-year of birth of each generation is shown in brackets) (reproduced with permission from Beral *et al.*, 1978. © by The Lancet Ltd, 1978).

Consider the scattergram in Figure 11.13(*a*). The points on the scattergram show a clear trend, upwards and to the right; there is said to be a *positive* relationship between the two variables. High values of one variable are associated with high values of the other, and low with low. To summarize the relation between the two variables, so as to be able to predict the value of one variable when we only know the other variable, we could just draw a straight line through the scatter of points. Any straight line drawn on a graph can be represented by an equation. In the above example, this relationship could be summarized as $y = x$, because each

Example 11.18. *Routinely collected data from a large number of population-based cancer registries and published measurements of ambient solar ultraviolet light were obtained to assess whether the geographical distribution of squamous-cell carcinoma of the eye was related to solar ultraviolet light. The analysis was based on data from 47 populations: 3 in Africa, 9 in Australasia, 20 in Europe, 12 in North America, 2 in South America and 1 in the Middle East. The study period covered by each registry varied, but most encompassed the 1980s. The results (Figure 11.12) were consistent with the hypothesis that exposure to solar ultraviolet light increases the risk of squamous-cell carcinoma of the eye (Newton* et al.*, 1996).*

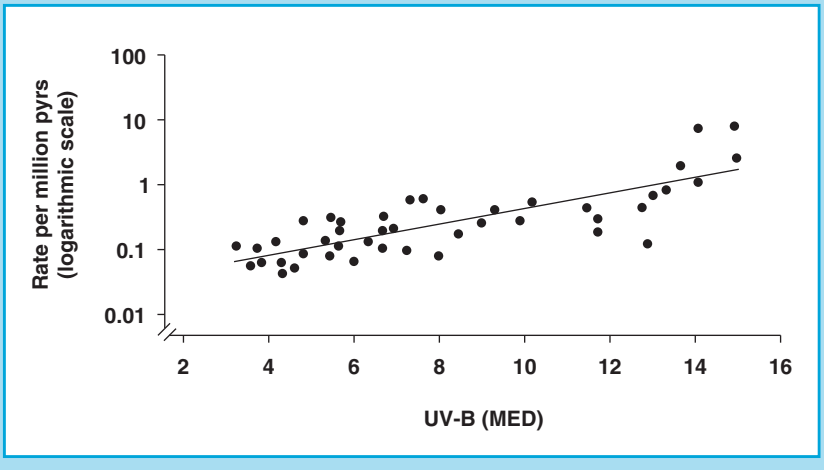


Figure 11.12.
Age- and sex-adjusted incidence rates of squamous-cell carcinoma of the eye in relation to measurements of ultraviolet B radiation expressed as minimum erythemal dose (MED), a unit which reflects more closely the biologically effective dose (reproduced with permission from Newton *et al.*, 1996. © by The Lancet, Ltd, 1996).

value of y is exactly equal to the corresponding value of x. If the value of y was always 0.9 times the value of x , we could express the relationship as $y = 0.9x$.

More generally, the equation for a straight line is expressed as

$$y = a + bx$$

where y refers to values of the outcome variable and x to values of the exposure variable, a is the intercept of the line on the y axis, and b is the slope of the line, the increase (or decrease) in y per unit increase in x (Figure 11.14). In the special case in which the line passes through the origin (O) of the two axes, as in Figure 11.13(*a*), and each value of y is exactly equal to the corresponding value of x, the equation reduces to $y = x$, since $a = 0$ and $b = 1$.

In Figure 11.13(*b*), there is also a perfect association between x and y, but the trend is downwards to the right; there is then said to be an *inverse*, or *negative*, relationship between the two variables. Such a relationship can also be expressed by the equation $y = a + bx$, except that in this case the coefficient b has a negative value, indicating that as one variable increased, the other decreased.

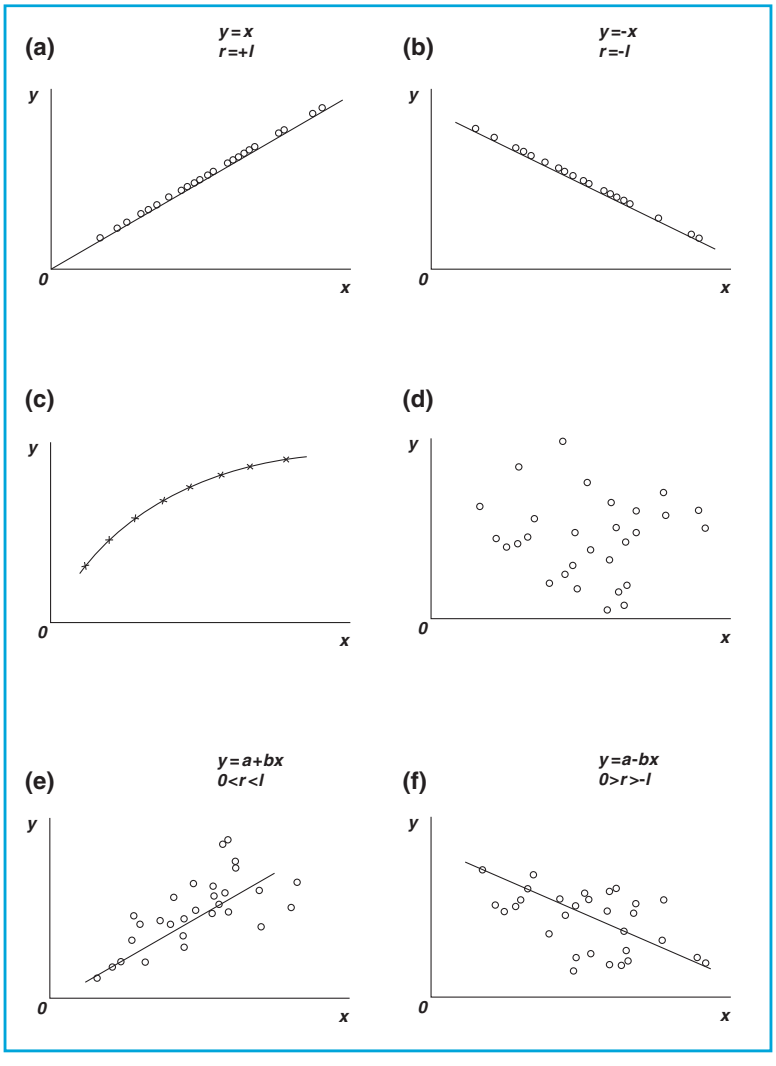


Figure 11.13.
Scattergrams, regression equations and correlation coefficients.

In Figures 11.13(*a*) and (*b*), the points on the scattergram lie on a straight line and the relationship is said to be *linear*. In Figure 11.13(*c*), there is still a clear relationship between the two variables, but it is *non-linear* in form, and the equation is more complex. Finally, in Figure 11.13(*d*), there is no apparent relationship between x and y. High values of x are associated with both high and low values of y. The points in the scattergram show no particular trend and, in the absence of any relationship between the two variables, the use of an equation is inappropriate.

The relationships illustrated in Figure 11.13(*a*), (*b*) and (*c*) are perfect, in that all the points on the scatter diagram lie on a line. In most real situations, however, the points are scattered around it (as in Figures 11.13(*e*) and (*f*). In these circumstances, we use a straight line that gives the 'best' prediction of y for any value of x. We could just draw a line 'by eye', but such a subjective method is unlikely to yield the best line. An alternative is to use the statistical method called *regression analysis*, to find the best line that '*fits*' the data. The equation of the straight line obtained by this method is called the *regression equation* (see Armitage and Berry (1994) for illustration of calculations).

This statistical method was applied to the ovarian cancer mortality data shown in Figure 11.11 (using the data from both countries) and yielded the following regression equation (Beral *et al.*, 1978):

$$y = 182 - 30x$$

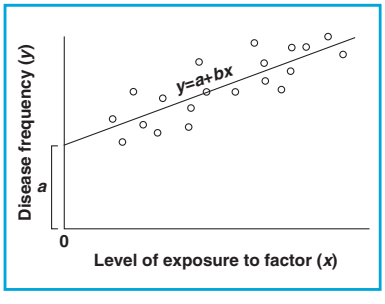


Figure 11.14.
Scattergram and regression line.

This equation means that for any given value of x (i.e., average family size), an associated value for y (i.e., ovarian cancer mortality) can be calculated. Thus, the age-standardized mortality ratio from ovarian cancer in any particular birth cohort of women can be predicted from this equation. For example, the mortality ratio in a cohort with a mean family size of three children can be predicted by substituting $x = 3$ into the regression equation

$$y = 182 - 30 \times 3 = 92$$

which is the value we would have estimated by examining the graph. This predicted value can be interpreted as the *average* value of y associated with a given value of x.

If all the values in the scattergram lay on the regression line, the predicted and observed values of y would be identical. The regression equation would describe exactly the relationship between average family size and ovarian cancer mortality. In other words, the variation in ovarian cancer mortality would be completely explained by the variations in family size. In practice this is not the case. Ovarian cancer mortality can vary independently of variations of family size, so that two populations with the same average family size may have different ovarian cancer mortality risks. Thus, the regression equation can only measure the *average* relationship between the two variables.

If the points on the scattergram lie close to the regression line, this suggests that the observed values for y do not differ markedly from the predicted values represented by the regression line. Thus, most of the variation in y can be explained by the variation in x. If, on the other hand, there is a wide scatter of points around the regression line, a considerable amount of the variation in y is not explained by the variation in x.

To quantify the degree of scatter around the regression line, we can calculate a measure called a correlation coefficient, r. The value of this coefficient always lies between –1 and +1:

(*a*) For perfect *positive* correlations (Figure 11.13(*a*)), $r = +1$.

(*b*) For perfect *negative* correlations (Figure 11.13(*b*)), $r = -1$.

(*c*) If there is some scatter about the regression line (Figure 11.13(*e*) and (*f*)), r lies between 0 and +1 (or between 0 and –1). The less the scatter, the closer r is to 1 (or –1).

(*d*) If there is *no linear* relationship between y and x, r is close to 0. This implies that either there is no relationship at all between the two variables (Figure 11.13(*d*)) or the relationship is non-linear (Figure 11.13(*c*)).

Use of correlation and regression in the analysis of ecological studies

The correlation coefficient is quite often used in the analysis of ecological studies as a measure of the strength of the association between exposure and disease. It is not, however, the most appropriate approach. First, its magnitude depends on the range of the exposure variable; if this is wide, the correlation will be greater than if it is narrow. Second, the value of the correlation coefficient cannot be translated into any of the conventional measures of relative effect and it is therefore difficult to interpret in epidemiological terms.

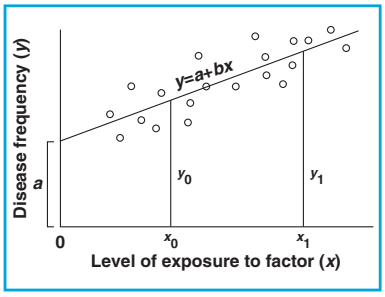


Figure 11.15.
Illustration of the association between disease frequency (y) and level of exposure (x) to the factor being studied in various populations.

A more appropriate method to quantify the effect of exposure in an ecological study is to fit a regression line to the data, which predicts incidence or mortality as a function of the level (or prevalence) of exposure. In contrast to the correlation coefficient, a regression line remains unaffected by changes in the range of the exposure variable. Moreover, a measure of relative effect can be estimated from the slope and the intercept of the regression line (Beral *et al.*, 1979).

Let us assume that in n different populations the level of disease (y) is linearly related to the average level of exposure (x) to the factor being studied. This relationship can be expressed by a regression equation fitted to the data from n populations:

$$y = a + bx$$

We can consider two of these populations with different average levels of exposure x_0 and x_1 and associated disease frequencies y_0 and y_1, respectively (Figure 11.15). The relative risk between the two populations can be defined as the frequency of disease in individuals with average exposure x_1, relative to that in individuals with average exposure x_0, or simply y_1/y_0.

Thus, if

$$y_0 = a + bx_0 \qquad \text{and} \qquad y_1 = a + bx_1$$

the relative risk estimate (RR) can be calculated as

$$\text{RR} = \frac{y_1}{y_0} = \frac{a + bx_1}{a + bx_0}$$

For instance, in our ovarian cancer mortality example, the regression equation was found to be $y = 182 - 30x$. Thus, the relative risk of ovarian cancer in a population with an average family size of two children compared to that with an average of four children can be estimated as:

$$\text{RR} = \frac{182 - 30 \times 2}{182 - 30 \times 4} = 2.0$$

If the exposure is measured in terms of the *proportion* of people in the population exposed to the factor of interest (e.g., proportion of cigarette smokers) rather than in terms of the *average level* of exposure (e.g., mean number of children) as in the above example, the relative risk can be estimated as:

$$\text{RR} = \frac{\text{slope}}{\text{intercept}} + 1 = \frac{b}{a} + 1$$

This formula is just a special case of the previous equation obtained by setting x_0=0, x_1=1, which are the range of values a proportion can take.

Certain assumptions underlie this approach. First, it assumes that the *y* variable (disease frequency) is linearly related to the *x* variable (exposure). Secondly, it presumes that the frequency of the disease in each population group is entirely determined by the level of exposure. Thirdly, it assumes that exposure is measured without error. As we shall see below, most often these assumptions are not satisfied, limiting the use and interpretation of these methods.

11.2.2 Interpretation of ecological studies

Ecological fallacy

The observation that there is a relationship at a population level between two variables does not necessary imply that the same relationship will hold at an individual level. This is known as the 'ecological fallacy'. For instance, the previous analyses on family size and ovarian cancer suggested that pregnancy might confer protection against ovarian cancer. We do not know, however, if the women who died from ovarian cancer in each population group were really those with few or no children.

Example 11.19. *Approximately 500 Finnish municipalities were grouped into various categories (five or four) according to various socioeconomic indicators such as average monthly income per inhabitant in 1968, percentage of inhabitants belonging to the two highest social classes (social classes 1 and 2) in 1970, percentage of people with secondary education in 1970, number of television licences per 1000 inhabitants in 1970, percentage of people living in dwellings with more than two inhabitants per room in 1950 (crowdedness), and percentages of dwellings with running water, central heating and electricity in 1950. The incidence of breast and cervical cancers was then calculated for each of these groups of municipalities (Hakama* et al.*, 1982). The results are shown in Figure 11.16.*

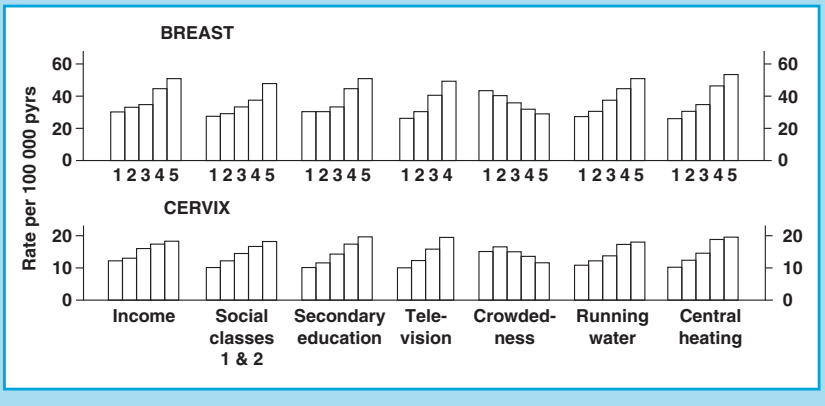


Figure 11.16.
Age-standardized incidence rates of breast and cervical cancers in groups of municipalities defined according to the levels of various socioeconomic characteristics of each municipality, Finland, 1955–74. The average Finnish population in 1955–74 was taken as the standard population. (Reproduced with permission from Hakama *et al.*, 1982).

In an ecological study, exposure levels represent the average levels for each population group. Since the true interest is in individual risk as determined by individual exposure, ecological studies are justified only in the ideal situation in which all individuals within a group have the same level of exposure. This may be the case with environmental exposures such as air pollution, natural radiation and water quality.

Most often, however, exposure is heterogeneous within a group, with some individuals not exposed at all and those exposed likely to be exposed in different levels. In such situations, the exposure–response relationship observed at the group level need not reflect the exposure–response relationship at the individual level. Thus, the finding of a linear relationship

Example 11.20. *In the study described in Example 11.19, analyses were also carried out at an individual level by individually linking all breast and cervical cancer registrations with the 1970 census forms. For each cancer patient, occupational and educational data were extracted from the census tapes, and the patients were grouped into socioeconomic and educational classes on the basis of the original census codes (Hakama* et al.*, 1982). The results of this individual analysis are shown in Table 11.6.*

Table 11.6.
Age-standardized incidence ratios (SIR) of breast and cervical cancer by socioeconomic status and educational level in women aged 30–69 years. Finland, 1971–75.[a]

	Breast cancer			Cervical cancer		
	Observed (O)	Expected[b] (E)	SIR (%) (100×O/E)	Observed (O)	Expected[b] (E)	SIR (%) (100×O/E)
Socioeconomic status[c]						
Employees	172	148.9	116	24	35.0	69
Farmers	562	705.1	80	108	165.9	65
Other self-employed	127	128.2	99	26	30.2	86
Managerial	343	229.3	150	24	53.8	45
Clerical	1061	842.7	126	177	197.3	90
Skilled workers	797	890.6	90	271	208.8	130
Unskilled workers	357	412.5	87	146	96.9	151
Pensioners	1169	1228.2	95	313	299.3	105
Educational level						
Primary	3011	3356.1	90	897	797.1	113
Secondary	975	847.5	115	150	200.0	75
High school	255	186.7	137	26	44.0	59
College/ university	356	206.8	172	17	48.9	35

[a] Data from Hakama *et al.* (1982)

[b] The breast and cervical cancer age-specific rates for the whole Finnish female population aged 30–69 years were taken as the standard.

[c] In Finland, socioeconomic status is based on occupation. For economically inactive women (e.g., housewives), it was defined as that of the head of the household (usually the husband).

between average exposure and disease frequency in an ecological study does not imply that such a linear relationship will be present at the individual level.

In the ecological analysis illustrated in Example 11.19, the risk of both breast and cervical cancer increased with increasing average socioeconomic level of the municipalities. This positive gradient is what we would expect for breast cancer but not for cervical cancer (see Examples 11.7 and 11.9). When analyses were conducted at an individual level, however, the socioeconomic and educational trends in risk were different for breast and cervical cancers (Example 11.20). Women with better jobs and higher education had a higher risk of breast cancer but a lower risk of cervical cancer (Table 11.6).

The difference in the cervical cancer results between the individual and the ecological approaches may be due to the fact that, for instance, women from the poorest socioeconomic groups (e.g., prostitutes) tend to be concentrated in urban municipalities, which are also the ones with a higher proportion of residents of higher socioeconomic status. This could potentially account for the positive relationship between high socioeconomic level and cervical cancer risks observed in the ecological analysis.

Confounding

A second major limitation of ecological studies is the lack of ability to control for the effects of potential confounding factors. For example, in a study of average *per caput* daily intake of fat in 24 countries in relation to their breast cancer incidence in women aged 35–64 years, there was a positive relationship between these two variables (Armstrong & Mann, 1985), suggesting a possible association between fat intake and risk of developing breast cancer. However, increased fat consumption may merely be acting as a marker for other factors that are related to elevated risk of breast cancer, such as higher socioeconomic status, lower fertility and later age at first birth. Data on known or suspected confounders are not generally available in ecological studies and, even if available, it would be difficult to adjust for them at a population level.

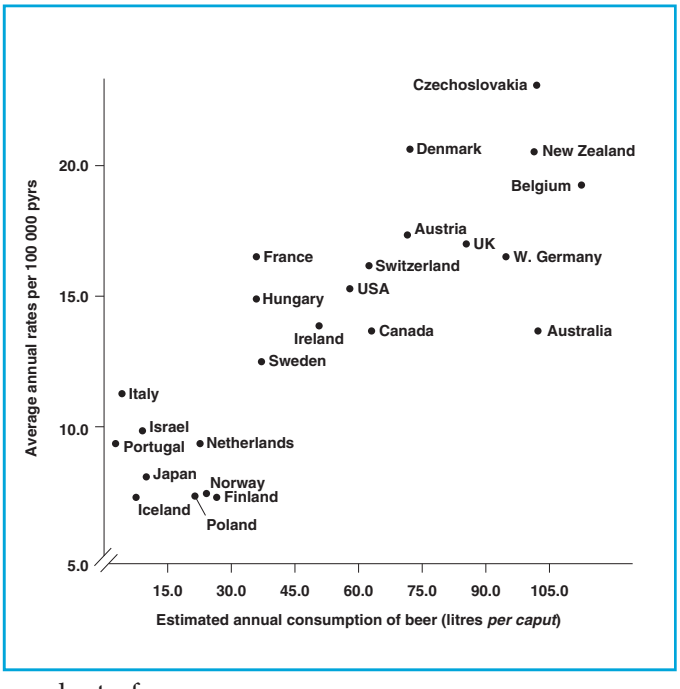


Figure 11.17.
Scattergram showing relationship between estimated average annual age-adjusted incidence rates for rectal cancer among males in 24 countries and *per caput* beer consumption, 1960–62 (approx.) (reproduced, by permission of Oxford University Press, from Breslow & Enstrom, 1974).

It should be noted that risk factors which are independent of exposure at the individual level may become correlated with it, and therefore become confounders, when aggregated at the population level. For example, in an investigation of the relationship between the proportion of woodworkers and lung cancer across geographically defined areas, smoking will induce confounding if cigarette consumption changes with the proportion of woodworkers in each area, even if the two factors in ques-

tion are independent at the individual level. Conversely, a confounding variable at the individual level may not be a confounder at the ecological level. For example, although the risk of lung cancer is much greater for men than women, sex, as an ecological variable (e.g., percent male), will not be associated with disease rate across geographically defined groups if the proportion of males is similar in all groups.

Measurement errors

Exposure is most often estimated from data collected for other reasons, which generally provide only an indirect measure of possible risk factors. For example, data on smoking and alcohol are often based on sales, which only partially reflect consumption, because losses and unregistered imports are not taken into account. Moreover, since the data are not collected by the investigators themselves, it may be difficult to assess their quality adequately.

Measurement of cancer incidence and mortality can also be affected by errors, as discussed in Section 11.1.7.

Latent period

Most ecological studies compare exposure measured at one point in time with disease measured at another (or the same) point in time. This is illustrated in Figure 11.17, where data on both the disease rate and the exposure of interest (*per caput* beer consumption) refer to approximately the same period (1960–62).

Ideally, an appropriate time-lag period should be incorporated into the analysis, so that exposure data refer to the relevant etiological period (e.g., 10–20 years before the development of cancer). Data on past exposures are not always available and, quite often, we are forced to rely on data from a period far too recent. This constitutes a serious problem when exposures are likely to have changed markedly over time (e.g., smoking and alcohol consumption), particularly if the rate of change has been different in the different groups. Even when relevant past exposure data are available, the populations on which exposure and outcome are measured may not be the same, as a result of dynamic changes introduced by births, deaths and migrations.

A special situation arises when birth cohort changes in exposure are related to birth cohort changes in disease risk or when both the exposure and the disease data refer to a specific birth cohort of individuals. Examples 11.16 and 11.17 illustrate this point. The average family size summarized the reproductive experience of each generation of women, that is the total number of children achieved by the end of their reproductive lives. The ovarian cancer cohort risks used in these examples also summarized the mortality experience of each generation of women (see Beral *et al.* (1978) for details of the calculations). In such situations, there is no need to build any time lag into the analysis.

Final remarks

Despite their limitations, ecological studies have been useful in describing differences in populations. Even if confounded by unknown or uncontrollable factors, such differences at least signal the presence of effects worthy of further investigation.

Ecological studies are particularly helpful in identifying factors responsible for risk differences *between* populations rather than risk variations *within* the same population. For instance, international comparisons have shown a strong relationship between fat intake and breast cancer risks. However, most individual-based studies conducted within populations have failed to observe such a relationship. It has been suggested that a possible reason for this difference in results is that between-population variability in levels of fat intake is much higher than the inter-individual variation within populations.

Box 11.1. Key issues

- Routine-data-based studies make use of routine surveillance systems to obtain data on both the exposure(s) and outcome(s) of interest. Thus, this type of study can be conducted without establishing contact with any of the study subjects.

- Routine-data-based studies can be carried out at an *individual* level (if the individual is the unit of study) or at an *ecological (aggregated)* level (if the group is the unit of study).

- The main advantages of routine-data-based studies are:

For individual level and ecological studies

1. They are very economical and rapid, since they generally use existing data on exposure and outcome, with no costs involved in collection.

2. They allow the study of very large numbers of people and, therefore, small increases in risk can be investigated.

3. They may include populations with a very wide range of exposure levels (more than can be found in a single population used for conventional cohort or case–control studies).

For ecological studies

4. They may be the only practical method if exposure level is relatively homogeneous in a population, but differs between populations (e.g., water quality), or when individual measurements of exposure are impossible (e.g., air pollution).

Further reading

* The book by Estève *et al.* (1994) gives a comprehensive (although statistically elaborate) review of methods used in the analysis of routine-data-based studies.

* A fuller discussion on ecological studies can be also found in papers by Greenland & Morgenstern (1989), Walter (1991a, b) and Greenland & Robins (1994).

* A discussion of the rationale and methodology of migrant studies in cancer epidemiology is also given in Parkin & Khlat (1996).

Box 11.1. (Contd) Key issues

• The main disadvantages of routine-data-based studies are:

For individual level and ecological studies

1. The number of variables on which data are available is limited.
2. It is difficult to assess errors in the measurement of exposure and outcome variables, since the data are not collected by the investigators themselves.
3. Data on potential confounding variables (except sex and age) are rarely available.

For ecological studies

4. They are prone to the *ecological fallacy*.
5. It is difficult to control for confounding even when data on potential confounders are available.
6. It is usually difficult to incorporate an appropriate time-lag period.