Cohort Studies

Anthony B. Miller, David C. Goff Jr., Karin Bammann, Pascal Wild

5.1 Introduction

This chapter summarises our basic understanding of cohort studies, a type of observational epidemiology study that some have also called longitudinal, or prospective. A cohort study evaluates the risk of disease or disease-related outcome in a population that is characterised in terms of relevant risk factors or exposures, placed under observation, and followed for some time until disease develops or not. In contrast to its classical counterpart, the case-control study (cf. Chap. I.6 of this handbook), cohort studies can relate multiple diseases to the exposure or exposures identified. On the other hand, cohort studies are frequently restricted to a limited number of exposures and potential confounders that can be included in the study, if historical data is used.

The chapter is organised as follows: First, a brief historical perspective on cohort studies is given, showing the importance of this study design by giving examples from the past and from today. Second, conceptual features of cohort studies are presented, where the two basic types of cohort studies, concurrent and non-concurrent historical cohort studies are summarised, and the basic concepts of data analysis in cohort studies are described. These concepts include the description of outcome events in the cohort, the comparison with external data and the analysis of effects of exposure. The chapter then deals with key concerns of cohort studies, like selection of the study population, and on the important question of how to determine exposure and outcome events in the framework of a cohort study. A review on ethical issues, mainly raised through the potential future use of specimens, is given.

5.2 A Brief Historical Perspective on Cohort Studies

Cohort studies have been used for over a century to study determinants of disease. Since the early days of epidemiology, they have been used as a powerful tool to study a broad range of exposures like infections, nutritional factors, occupational exposures, and lifestyle factors as the following examples illustrate.

The classical study on the London cholera epidemic of 1849 conducted by John Snow is an example of a cohort study on infectious diseases (Snow 1855; Sutherland 2002). Previous reports from the Registrar General had drawn attention to the possibility that differences in water supply were associated with differences in cholera rates across sections of London. Two different water companies (the Lambeth and the Southwark & Vauxhall) supplied households within various regions of London, and frequently these two water companies supplied adjacent households. The companies differed in one important feature, the location of the water intake. The Lambeth had moved their water intake upstream from the sewage discharge point in 1849; whereas, the Southwark & Vauxhall continued to obtain water downstream of the sewage discharge point. Dr. Snow classified households according to their exposure to the two water sources and showed a substantial difference in cholera mortality, 315 versus 37 cholera deaths per 10,000 households served by the Lambeth and Southwark & Vauxhall companies, respectively.

Cohort studies continue to be an important tool in the investigation of infectious diseases. For example, McCray (1986) used a cohort design to quantify the risk of developing the acquired immunodeficiency disorder (AIDS) among healthcare workers exposed to blood and body fluids of AIDS patients.

Joseph Goldberger employed a variety of epidemiological approaches, including cohort methods, to study pellagra, a systemic disease endemic in the southeast of the United States in the late 19th and early 20th century (Terris 1964). In one investigation, Goldberger examined the dietary exposures of households in relation to the occurrence of pellagra and demonstrated that a cornmeal subsistence diet was associated with pellagra. Subsequent trials showed that pellagra could not be transmitted from person to person, as might be expected for an infectious disease, but could be prevented by the "pellagra preventive factor" later determined to be niacin. More recently, Oomen and colleagues studied the association of trans-fatty acids, a hydrogenation product of oils containing polyunsaturated fatty acids, and heart disease among men in the Netherlands (Oomen et al. 2001). They found a relative risk of 1.28 of heart disease for an increase of 2% of energy from trans-fatty acids intake at baseline.

Occupational epidemiology is another classical field of application of cohort studies. Typically workers exposed to a putative harmful substance are compared to other workers in the industry or to the general population. Occupational cohorts were used to study, for example, the association between exposure to dyes and urinary bladder cancer (Case et al. 1954), exposure to mustard gas and respiratory cancer (Wada et al. 1968), and exposure to benzene and leukaemia (Rinsky et al. 1987). The health effects for workers exposed to asbestos continue to be examined. Ulvestad and colleagues (2004) conducted a cohort study of members of the Norwegian Trade Union of Insulation Workers hired between 1930 and 1975 and followed through 2002, demonstrating relative increases in risk of mesothelioma and lung cancer when compared with the experience of the general population.

In addition to diet, other lifestyle exposures have attracted the attention of epidemiologists, including physical activity, tobacco and alcohol use. Morris and colleagues (1953a, b) demonstrated that British bus drivers had approximately twice the risk of heart disease in comparison to the more active conductors (who went up and down the stairs to collect tickets). This result was confirmed in a comparison of postmen with telephonists and clerks (Morris et al. 1953a,b). In 1951, Doll and Hill (1954) initiated a cohort study of British physicians by collecting data on tobacco use via questionnaire. By collecting death certificate data, they were able to demonstrate a 10-fold increased risk of lung cancer death for smokers compared to non-smokers (Doll and Peto 1976). Doll and colleagues also reported on the association of alcohol consumption with mortality among British doctors (Doll et al. 1994a) demonstrating a u-shaped relationship, with greater mortality among abstainers and heavy drinkers and the lowest mortality among moderate drinkers, defined as 1–2 drinks per day on average. Concerns persist that the increased risk described in abstainers may be falsely elevated by the experience of former drinkers who may have quit drinking due to health decline. This concern has been addressed by Eigenbrodt and colleagues using cohort methodology within the Atherosclerosis Risk in Communities (ARIC) study (Eigenbrodt et al. 2001). Eigenbrodt and colleagues measured perceived health status and alcohol consumption behaviour longitudinally and were able to identify changes in health status that preceded changes in drinking behaviour. They demonstrated that perceived health decline predicted cessation of drinking, thereby providing evidence that the risk among abstainers may have been inflated in studies that failed to distinguish between lifelong abstainers and former drinkers.

Despite disadvantages regarding cost and complexity, cohort studies remain until today of substantial public health importance as indicated by several of the previously cited examples and by such evidence as was recently provided by the National Institutes of Health (NIH). The NIH is considering the establishment of a 500,000-person cohort study to examine genetic and environmental influences on common diseases in the United States (National Institutes of Health 2004). The large sample size under consideration for this study would enable the examination of gene-gene- and gene-environment interaction in the general population and in subgroups of interest. Therefore, a sound understanding of cohort methodology is of substantial importance to the modern epidemiologist.

5.3 Conceptual Foundations

5.3.1 Types of Cohort Studies: Concurrent and Non-concurrent Approaches

The central feature of a cohort study is the collection of exposure data in a defined population and the subsequent surveillance of possible outcome events regarding health, morbidity, and mortality. For this purpose, healthy members of a defined population (the cohort) are classified according to their exposure status (e.g. exposed vs. unexposed) and followed over a longer period with respect to their health status. Then, the question can be answered if incidence of outcome events is associated with former presence or absence of exposure, which would indicate a possible causal relationship.

Within this framework, cohort studies can be classified in two major categories depending on the timing of follow-up period relative to the time of study conduct. In *concurrent cohort studies*, sometimes referred to as prospective cohorts (Fig. 5.1), a defined population is assembled and possibly screened to eliminate persons with disease. Then, information on exposure, possible confounders, and other important factors is gathered. The cohort members are subsequently followed for a specified period into the future recording outcome events of interest. In *non-concurrent* or *historical cohort studies* (Fig. 5.1), a population is assembled

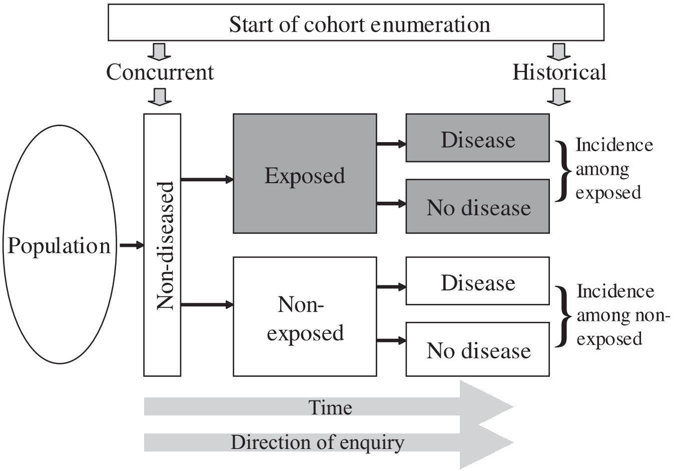


Figure 5.1. Design of a cohort study

from available data records, for example from company files. Exclusion of persons with disease and assessment of exposure and other factors is based on the available data from the past. Cohort members are monitored for outcome events through existing documents and data systems (e.g. vital statistics files or disease registries) to some point in the past. As in concurrent studies, outcome rates may be compared across exposure categories within the cohort, or, if all members of the cohort are assumed to be exposed, outcome rates may be compared between the cohort and the general population, assumed to be unexposed. A combined approach is also possible, with the cohort assembled and followed initially through historical documents or other data sources such as data from registries and subsequently followed using concurrent methods. The distinction between these two major categories of cohort studies has important implications regarding data collection.

In concurrent studies, the methods for cohort assembly and data collection can more easily be controlled; whereas, in non-concurrent studies, the investigators must rely on data recorded in historical records almost always for reasons other than medical research. This notable disadvantage of the non-concurrent approach is compensated by the ability to study exposures, such as occupational exposures, that meet one or more of the following key conditions: (1) the exposure can be attributed to selected employed populations based on individual records of job descriptions or other employment data, (2) the exposure is relatively rare in the general population outside the occupations of interest, (3) the induction period is long, and (4) the health concern is substantial, making the continued exposure required for a concurrent study undesirable from a public health perspective.

Because many of the non-infectious diseases tend to be multi-factorial in causation, a crucial point in the validity of cohort studies is the inclusion of data on possible confounders at baseline. This is a problem in historical cohort studies, that will be discussed in the section on determining exposures below.

Two modern extensions of cohort studies that try to integrate the advantages of cohort and case-control studies are designed to have nearly all the power of classic cohort studies, but utilise relatively economically detailed exposure information from questionnaires, biomarkers or other biological measurements determined from the collection of biological specimens at the time the study is initiated. These analytic designs, i.e. nested case-control studies and case-cohort studies, are discussed in detail in Chap. I.7 of this handbook and will not further be considered here.

5.3.2 Description of Outcome Events in the Cohort

In contrast to case-control studies, cohort studies with their straightforward design allow direct comparisons of exposed and unexposed persons and can provide measures of effects for various outcome events, like e.g. different endpoints (morbidity, mortality, pre-morbidity) and/or different diseases. Nevertheless, analysis of cohort data requires reasonable care especially in the steps of data preprocessing for description and analysis. The often necessary change of perspective from persons at risk to person-time at risk needs special attention to ensure that unbiased results can be obtained. This subsection will refer mainly to disease incidence; however other measures can principally be treated in the same manner.

The results from a cohort study can be presented as shown in Table 5.1.

Table 5.1. 2 × 2 table summarising the results of a cohort study

		Second observe		Total
		Disease contracted	No disease	
First	Exposed	a	b	$a + b = n_E$
select	Non-exposed	c	d	$c + d = n_{\overline{E}}$
Total		$a + c$	$b + d$	$a + b + c + d = N$

The easiest way to describe outcome events in a cohort is by counting the number of persons experiencing the event of interest and to relate this number to the crude number of persons at risk in the cohort. Disease incidence, for example, can be described by the cumulative incidence or risk, which is calculated by dividing the number of incident cases by the number of persons at risk at baseline:

$$\widehat{\text{Risk}} = \text{number of incident cases/number of persons at risk}\,, \tag{5.1}$$

that can be calculated as

$$\widehat{\text{Risk}} = (a + c)/N \tag{5.2}$$

and accordingly for the exposed and unexposed study populations as

$$\widehat{\mathrm{Risk}}_E = a/(a+b) = a/n_E$$

$$\widehat{\mathrm{Risk}}_{\overline{E}} = c/(c+d) = c/n_{\overline{E}}\,.$$

The cumulative incidence or risk is unit-free and represents an individual risk of developing the disease. It is a proportion, not a rate and it does not account for possible different periods of disease-free follow-up time of cohort members, but assumes a fixed cohort. In cohort studies on acute diseases with short induction periods and a short time of follow-up, like outbreaks, the risk of disease can be estimated directly using the cumulative incidence, given a fixed cohort with fixed period of follow-up and a low fraction of drop-outs. In cohort studies on chronic diseases with their long follow-up periods, however, the use of the cumulative incidence is not appropriate because usually disease-free follow-up periods differ strongly among cohort members. In this case, outcome events are preferably described by rates, that represent the number of outcome events divided by the cumulated duration of event-free follow-up periods of all cohort members at risk. For further analysis, all rates presented in the following can be used to determine rate ratios and rate differences as described in Chap. I.2 of this handbook. Disease incidence can be expressed as incidence rate (I):

$$\widehat{I} = \text{number of incident cases/person-time at risk}\,, \tag{5.3}$$

where each cohort member is contributing the time from entry into the study to either development of disease or end of follow-up to the denominator of the incidence rate, thus accounting for different times at risk of the cohort members to develop the disease. The incidence rate is sometimes called incidence density and should not be confused with the above mentioned cumulative incidence. Assuming total person-time of follow-up of t, with t_E and $t_{\overline{E}}$ follow-up of exposed (E) and unexposed ($\overline{E}$) populations, (5.3) results in

$$\widehat{I} = (a+c)/(N \times t)\,, \tag{5.4}$$

where $N \times t$ denotes the person-time at risk. Calculating the incidence rates separately for the exposed and unexposed study populations gives

$$\widehat{I}_E = a/\left[(a+b) \times t_E\right] = a/(n_E \times t_E)$$

$$\widehat{I}_{\overline{E}} = c/\left[(c+d) \times t_{\overline{E}}\right] = c/(n_{\overline{E}} \times t_{\overline{E}})\,.$$

Measures of risk and incidence of disease may provide important information regarding the public health burden of the outcome or disease of interest.

Since incidence rates often vary considerably by e.g. age, sex, calendar year, and race, the calculation of specific incidence rates instead of crude incidence rates may be desirable. For this purpose, different strata (for one group variable) or cells (for two or more group variables) have to be defined over the group variables' range.

The individual contributions of the cohort members to numerator and denominator of the incidence rate have to be assigned to the respective stratum or cell. Usually, each cohort member will contribute to more than one stratum or cell as he/she moves through the cohort during follow-up. Age- and calendar-specific incidence rates can be approximated well enough on the base of calendar year data if more precise information on months and days is not available (see Breslow and Day 1987).

A simple example demonstrates the principle steps for the calculation of specific incidence rates for the age groups 30–39 years, 40–49 years, and 50–59 years. Table 5.2 shows the data of a fictitious cohort, for which we will calculate age-specific incidence rates. Since exact dates in terms of months and days are not available in our example, age and follow-up time will be approximated by full and half years. The contribution of the year at entry into the study and the year of diagnosis is approximated as half a year (see Fig. 5.2).

The cohort consists of 10 persons who were followed for 20 years resulting in a total of 155 person-years of follow-up, deaths and drop-outs accounted for the lacking 45 person-years. Three cases of the disease of interest occurred in the cohort during follow-up, resulting in a crude incidence rate of 3/135 = 0.022 cases/person-year. The difference between the total of 155 observed person-years and the 135 person-years in the denominator of the incidence rate results from 20 years of cumulated follow-up time after diagnosis in the three cases. A useful general way in which to think of cohort data is to separate person-time at risk and person-time under observation.

A subject is “at risk” at a given moment if the event of interest can happen. Thus if a subject gets a thyroid surgery, she/he is no longer at risk of getting a thyroid cancer. If on the other hand the event of interest were a pregnancy, a woman would not be “at risk” of becoming pregnant if she already is pregnant or during spells of abstinence. In this case, however, the woman is “at risk” again from the moment on she desires another child. In the example above, a subject is no longer considered

Table 5.2. Data from a fictitious cohort

No.	Age at entry	Years of follow-up	Age at end of follow-up	Age at diagnosis	Person-years at risk
1	34	15	49		15
2	39	20	59	54	15
3	31	12	43		12
4	36	17	53	41	5
5	38	9	47		9
6	38	16	54	51	13
7	41	11	52		11
8	32	20	52		20
9	39	18	57		18
10	42	17	59		17
Total		155			135

"at risk", after diagnosis of the disease. Of course no subject is "at risk" from the moment of his/her death. Being at risk depends only on the endpoint studied.

On the contrary, being under observation, (i.e. being followed up), depends on the precise definition of the cohort and the method of follow-up considered in the epidemiological study. A subject is under observation at a time *t*, if, were the event of interest to occur at this moment, it would be recorded. Thus for example if the cohort definition were "all subjects employed in a given factory with at least one year of employment", the follow-up would start only at the moment the subject satisfies this criterion. In this case, all the person-time in the first year must be ignored. If the event of interest occurred in this year, it would not satisfy the inclusion category. Similarly a subject would be dropped from the follow-up at a time *t* if no information as to his/her disease status could be retrieved from time *t* on (e.g. the subject moves abroad), the subject is then considered "lost to follow-up". A subject contributes person-time to the study at any moment *t* if and only if at this moment he/she is "at risk" and "under observation".

Coming back to the example, each incident case is assigned to the age group he/she belonged to until diagnosis. In the same manner, the disease-free time of follow-up of each cohort member is allocated to the three age groups yielding the age-specific incidence rates presented in Table 5.3.

Incidence rates are commonly re-scaled e.g. to cases per 100,000 person-years underlining their reference to populations rather than to individuals. The crude

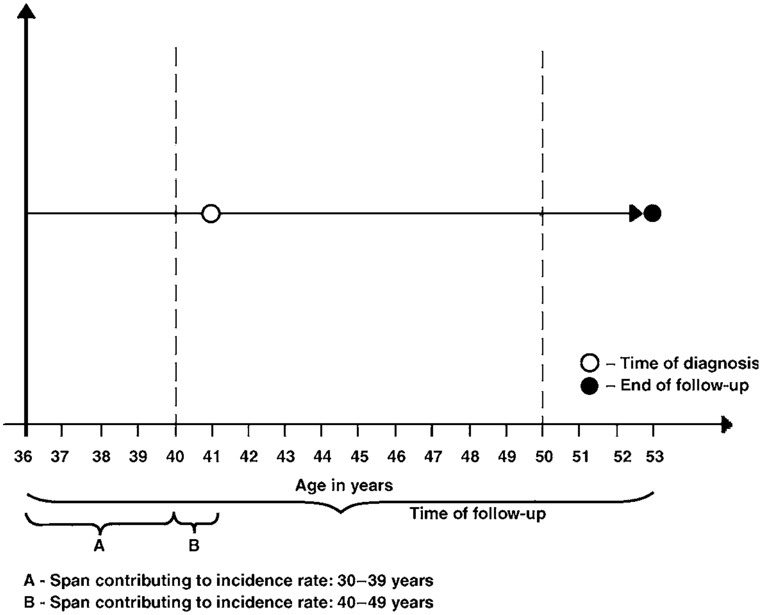


Figure 5.2. Follow-up time of cohort member No. 4 of the fictitious cohort

Table 5.3. Age-specific incidence rates for fictitious cohort data

Age group	Incident cases	Disease-free follow-up time	Age-specific incidence rate
30–39	0	5.5+0.5+8.5+3.5+1.5+1.5+0+7.5+0.5+0 = 29	0/29 = 0
40–49	1	9.5+10+3.5+**1.5**+7.5+10+8.5+10+10+7.5 = 78	1/78 = 0.013
50–59	2	0+**4.5**+0+**0**+0+**1.5**+2.5+2.5+7.5+9.5 = 28	2/28 = 0.071

incidence rate of 0.022 cases/person-year of the fictitious cohort, for example, would then be expressed as 2222/100,000 person-years.

In Fig. 5.2 the follow-up time of cohort member No. 4 is depicted schematically with respect to age. The first three and a half years, denoted with A, of the five years of disease-free follow-up time (41 years at time of diagnosis – 36 years at entry into the study) are contributing to the denominator of the incidence rate of the first age group (30–39 years), the next one and a half year, denoted with B, contribute to the numerator of the incidence rate of the second age group (40–49 years).

To quantify the frequency of exposure in the population under study the prevalence of exposure may be considered:

$$\widehat{P}_E = (a + b)/N = n_E/N \,. \tag{5.5}$$

The various quantities presented here can be used to derive measures of association accordingly (see Sect. 5.3.4).

5.3.3 External Comparisons

One important task in cohort studies is the comparison of the cohort with external data, preferably from the general population. Irrespective of the existence of internal comparison groups, external comparisons always give valuable insights by putting the cohort data in a broader context. For external comparisons either age-, sex- and calendar year-specific incidence or mortality rates or cumulative measures can be used. Standardised incidence rates can be calculated from specific incidence rates by weighting them with the age-, sex- and calendar year-distribution of the external comparison data (direct standardisation). However, cumulative measures have to be interpreted cautiously since they can mask underlying differences in specific disease patterns, like e.g. an unusually high incidence rate among younger persons in the cohort. With d_i denoting the number of cases in the age group i, n_i denoting the disease-free person-years accumulated in the age group i and w_i denoting the proportion of persons in the age group i in the standard population, the directly age-standardised incidence rate $\widehat{I}_W$ calculates as:

$$\widehat{I}_W = \sum_{i=1}^{I} w_i d_i / n_i \,, \tag{5.6}$$

Indirectly standardised measures requiring morbidity or mortality rates of the standard population are the standardised morbidity or incidence ratio (SIR) and the standardised mortality ratio (SMR). Since morbidity data is not routinely available in most countries the standardised mortality ratio is used much more frequently. The SMR compares the observed numbers of deaths in the cohort with the expected numbers, given the age structure of the cohort and the age-specific mortality rates λ_i of a reference population. With d_i denoting the number of deaths in the age group and n_i denoting the person-years accumulated in the age group, the SMR is estimated as

$$\widehat{\mathrm{SMR}} = \sum_{i=1}^{I} d_i / \sum_{i=1}^{I} n_i \lambda_i \,, \tag{5.7}$$

where $\sum_{i=1}^{I} d_i$ represents the total number of observed deaths in the cohort under investigation and $\sum_{i=1}^{I} n_i \lambda_i$ the expected number of deaths that are obtained by applying age-specific incidence rates of the reference population to the cohort under investigation. A SMR above 1 indicates a larger mortality in the cohort, a SMR below 1 a smaller mortality in the cohort compared to that of the reference population. Statistical testing of a single SMR can be done with a simple χ^2-test (observed vs. expected) with one degree of freedom. Assuming that the number of observed cases $D = \sum_{i=1}^{I} d_i$ follows a Poisson distribution with expectation $\gamma = E(D)$, confidence limits for the SMR ($\widehat{\mathrm{SMR}}_L$, $\widehat{\mathrm{SMR}}_U$) can be obtained by finding confidence limits $\widehat{\gamma}_L$, $\widehat{\gamma}_U$ for the number of observed cases:

$$\widehat{\mathrm{SMR}}_L = \widehat{\gamma}_L / \sum_{i=1}^{I} n_i \lambda_i \quad \text{and} \quad \widehat{\mathrm{SMR}}_U = \widehat{\gamma}_U / \sum_{i=1}^{I} n_i \lambda_i \,. \tag{5.8}$$

The confidence limits for γ can be determined as:

$$\widehat{\gamma}_L = (1/2)\chi^2_{2D,\alpha/2} \quad \text{and} \quad \widehat{\gamma}_U = (1/2)\chi^2_{2(D+1),1-\alpha/2} \,, \tag{5.9}$$

where $\chi^2_{2D,\alpha/2}$ denotes the $100(\alpha/2)$th percentile of the χ^2-distribution with $2D$ degrees of freedom, and $\chi^2_{2(D+1),1-\alpha/2}$ denotes the $100(1-\alpha/2)$th percentile of the χ^2-distribution with $2(D+1)$ degrees of freedom (see e.g. Sahai and Khurshid 1996).

If the age-specific rates of the standard population are just estimations of the exact rates, as is often the case with morbidity data, calculation of confidence intervals for the SMR can be performed by the method described in Silcocks (1994). A method for estimating the SMR where information on vital status is complete but information on cause of death is partly missing as may be the case in historical cohort studies can be found in Rittgen and Becker (2000).

Comparison of rates by direct standardisation has poor statistical properties, especially due to large variances of age-specific rates in small cohorts. Therefore, indirect standardisation is usually preferred (see Chap. I.2 of this handbook).

5.3.4 Summary Effects of Exposure

The main goal of cohort studies is to compare morbidity and/or mortality in exposed and non-exposed subjects or between different exposure groups of the cohort, and to investigate dose-effect relationships between exposure and disease. If the exposure is constant and can be determined at entry into the cohort, internal comparisons can be performed by calculating specific incidence rates for each exposure category separately as if each group were a separate cohort. Cumulative rates can be used, again provided the subgroups do not differ in important determinants of disease, like e.g. age.

In the simple case of a single dichotomous exposure several measures of association of exposure with disease can be estimated from results provided by a cohort study (see Table 5.1). In the following, the most important ones will be briefly introduced. A detailed discussion of their properties and examples for their calculation can be found in Chap. I.2 of this handbook.

The perhaps most popular measure of association is the risk ratio (RR), also known as relative risk, that compares the experience of exposed and unexposed populations. With the notation given in Table 5.1 and the risks for the exposed and unexposed subjects calculated according to (5.2) it can be estimated as

$$\widehat{\mathrm{RR}} = \widehat{\mathrm{Risk}}_E / \widehat{\mathrm{Risk}}_{\overline{E}} = [a/(a+b)] \,/\, [c/(c+d)] = (a/n_E)/(c/n_{\overline{E}})\,. \tag{5.10}$$

The incidence ratio (IR) compares the incidence rates in the exposed and unexposed study populations. According to (5.4) its estimator is given as

$$\widehat{\mathrm{IR}} = \widehat{I}_E/\widehat{I}_{\overline{E}} = \left\{a/\left[(a+b) \times t_E\right]\right\} / \left\{c/\left[(c+d) \times t_{\overline{E}}\right]\right\} = [a/(n_E \times t_E)] / [c/(n_{\overline{E}} \times t_{\overline{E}})] \tag{5.11}$$

The RR and IR provide estimates of the relative strength of the association between the exposure of interest and the outcome or disease of interest.

The absolute difference in risk (AR) between the exposed and unexposed groups provides an estimate of the impact of the exposure on the risk of disease in absolute terms. This measure is not to be confused with the absolute risk, which is the absolute probability that a disease-free individual will develop a given disease over a specific time-interval (Benichou 1998). Using the above formulas for the risks among exposed and unexposed it can be obtained from a cohort study as

$$\widehat{\mathrm{AR}} = \widehat{\mathrm{Risk}}_E - \widehat{\mathrm{Risk}}_{\overline{E}} = [a/(a+b)] - [c/(c+d)] = a/n_E - c/n_{\overline{E}}\,. \tag{5.12}$$

Based on the attributable risk several other measures can be derived. The so-called attributable fraction (AF) can be interpreted as the proportion of risk due to exposure in exposed individuals. It may be useful for quantifying the degree to which risk can be reduced at the individual level if the exposure (and its effects) can be eliminated. It may, therefore, be a sensible measure for counselling individuals:

$$\widehat{\mathrm{AF}} = \widehat{\mathrm{AR}}/\widehat{\mathrm{Risk}}_E = \left\{[a/(a+b)] - [c/(c+d)]\right\} / [a/(a+b)] = \left(a/n_E - c/n_{\overline{E}}\right) / \left(a/n_E\right) . \tag{5.13}$$

The population attributable risk (PAR) reflects the absolute level of risk of the outcome in the population due to the exposure. It can be used to estimate the public health impact, in absolute terms, of elimination of the exposure, at least with respect to the outcome of interest. Based on the attributable risk and the prevalence of exposure (see (5.5)) it is given as

$$\widehat{\mathrm{PAR}} = \widehat{\mathrm{AR}}/\widehat{P}_E = \left\{[a/(a+b)] - [c/(c+d)]\right\} / [(a+b)/N] = (a/n_E - c/n_{\overline{E}})/(n_E/N) . \tag{5.14}$$

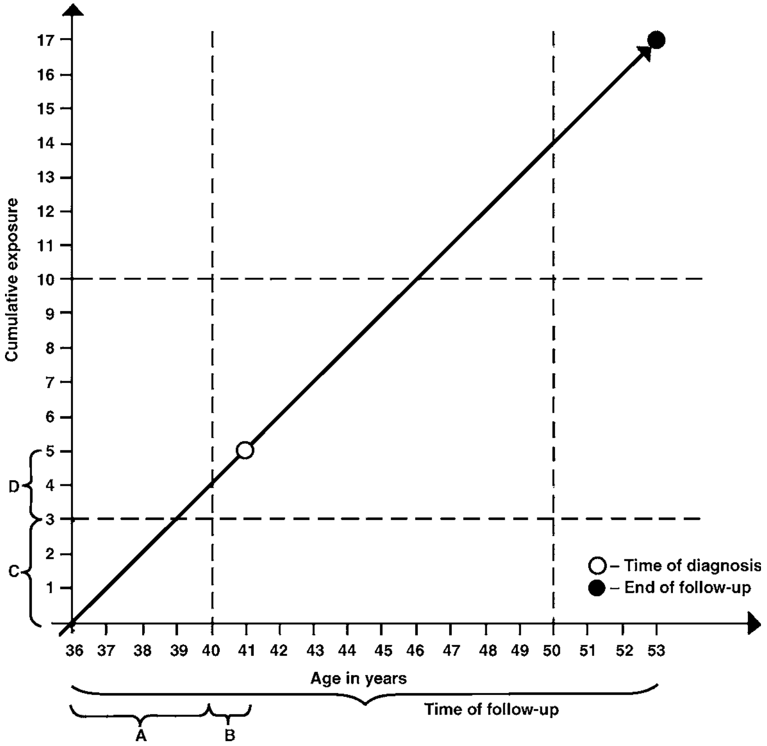
The last measure to be mentioned here may be used to estimate the proportion of all events of interest that could be prevented in the overall population if the exposure (and its effects) can be eliminated. The population attributable fraction (PAF) is defined as the proportion of all events of interest that occur in the population due to the exposure:

$$\widehat{\mathrm{PAF}} = \widehat{\mathrm{PAR}}/\widehat{\mathrm{Risk}} = (a/n_E - c/n_{\overline{E}})/\left\{(n_E/N)\,[(a+c)/N]\right\} . \tag{5.15}$$

Internal Modelling of the Effects of Exposure 5.3.5

The situation is more complicated, if cohort members continuously add exposure over follow-up time. Simple categorisation on the basis of cumulative exposure would lead to biased results. Person-years accumulated shortly after entry into the study of cohort members with high cumulative exposure would wrongly be allocated to a high exposure category, although the cumulative exposure at that time-point was still low for these cohort members, resulting in underestimation of high exposures and overestimation of low exposures. Therefore, the disease-free person-time of each subject has to be subdivided and assigned to the respective age- and sex-specific exposure category the cohort member belongs to as he or she moves through the cohort, meaning that most cohort members contribute to different age-exposure-categories. In the same manner, the incident cases have to be assigned to the categories where they occurred.

In Fig. 5.3 the follow-up time of cohort member No. 4 is again depicted schematically, this time with respect to age and cumulative exposure assuming that the exposure starts at the beginning of the follow-up and that it is constant over time. For age- and exposure-specific incidence rates, the disease-free follow up time is assigned to the groups according to the squares in the figure that are defined by the categorisation of the group variables, resulting in a contribution of cohort member No. 4 of two and a half year to the denominator of the incidence rate of category A × C (30–39 years of age and < 3 units of cumulative exposure), one year to the denominator of the incidence rate of category A × D (30–39 years of age and 3



A - Span contributing to incidence rate : 30 – 39 years
B - Span contributing to incidence rate : 40 – 49 years
C - Span contributing to incidence rate : 0 – 2
D - Span contributing to incidence rate : 3 – 10

Figure 5.3. Follow-up time of cohort member No. 4 of the fictitious cohort

to smaller than 10 units of cumulative exposure), and one and a half year to the denominator of the incidence rate of category B × D (40–49 years of age and 3–< 10 units of cumulative exposure). The case itself contributes to the nominator of the incidence rate of category B × D, since this is the category in which he/she was diagnosed.

This procedure can be extended in several ways. The exposure may have started before beginning of follow-up or may start later. It can vary over time, it can even vary from individual to individual or can be lagged to account for induction time. Several measures of exposure (e.g. time since first exposure and and/or confounders) can be considered simultaneously and possible confounders can be included in the analyses as additional variables. Figures 5.4, 5.5 and 5.6 illustrate some of these features. For simplicity no half-years are considered in these examples.

In Fig. 5.4, a subject is followed up from age 23 but has been exposed from age 19 on, he/she is exposed until age 27 followed by an unexposed 5 year period. He/she is again exposed until age 39 at which time his/her person-time at risk ceases either

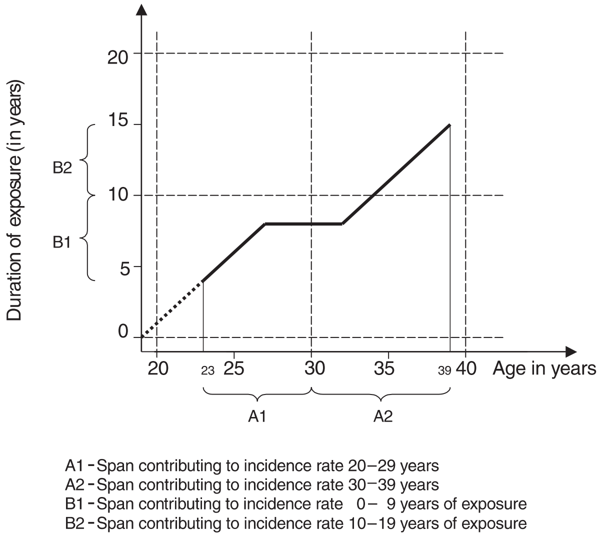


Figure 5.4. Person-time classification with varying duration of exposure

because of disease diagnosis or because of end of follow-up. This subject would contribute 7 years (from age 23 to age 30) to the A1 × B1 group (20–29 years of age, 0–10 years exposure) 4 years (from age 30 to 34) to the A2 × B1 (30–39 years of age, 0–10 years of exposure), 5 years (from age 34 to 39) to A2 × B2 ((30–39 years of age, 10–19 years of exposure).

Fig. 5.5 presents the same subject assuming that the first exposure spell was twice as intensive (e.g. 20 ppm of a given chemical) than the second exposure (10 ppm). The unit of cumulative exposure *y*-axis is now in ppm.years. The subject would contribute 1 year to group A1 × B1, (his cumulative exposure is then 100 ppm.years) then 5 years to group A1 × B2 (at age 30 his cumulative exposure is 160 ppm.years), then 6 years (from age 30 to age 36 at which he reaches 200 ppm.years) in group A2 × B2) and finally 3 years in group A2 × B3.

Fig. 5.6 considers the same subject again but this time the exposure is lagged by 10 years, say, to account for disease induction time. The first period would then be a non-exposed period. The rationale is that, were the disease to occur in these first 10 years, it would not be attributable to exposure. Applying the same rationale as before, the subject would contribute 6 years in group B0 × A1, then 1 year in group B1 × A1, finally 9 years in group B1 × A2, the lagged cumulative exposure at end of follow up (i.e. at age 39) is 160 ppm−years.

Another exposure can occur during the follow-up, e.g. the preceding subject starts smoking at age 25. In this case a further splitting of the time periods would be done separating periods in which the subject was a non-smoker and periods in which he/she smoked.

This splitting of person-time into age and exposure groups must be done for each subject of the cohort and gets more complex with a growing number of group

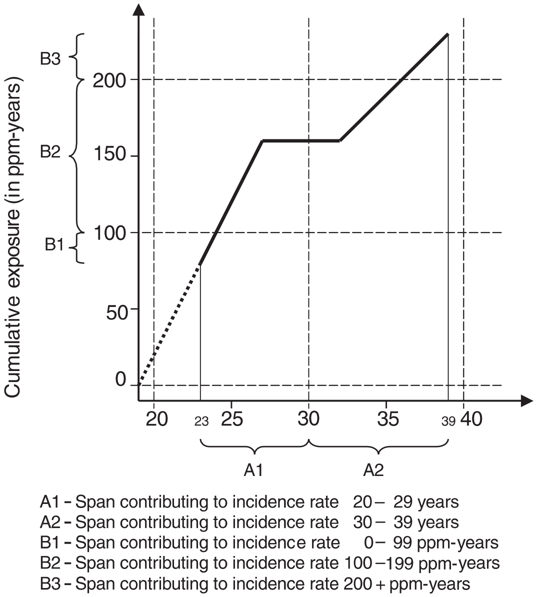


Figure 5.5. Person-time classification with varying cumulative exposure

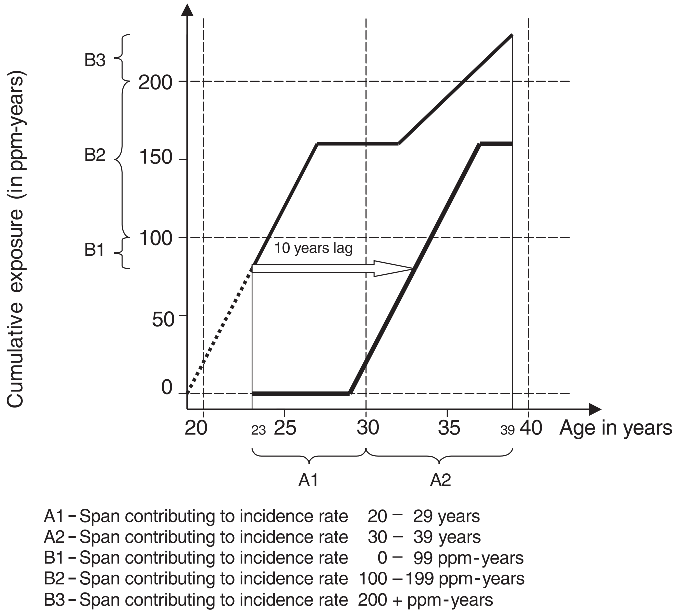


Figure 5.6. Person-time classification with varying lagged cumulative exposure

variables. Specialised software packages exist (e.g. Coleman et al. 1986) to perform these computations but they are usually limited in the complexity they can handle. Interestingly, these restrictions do not apply to some more general packages as Stata (version 7 or later – StataCorp. 2001) or Epicure (Preston et al. 1993) in which the statistical modelling procedures of such data are furthermore included. The end result of the calculations carried out in these packages can then be presented as a data table with each line corresponding to a separate combination of age and exposure classes (other classifications like calendar periods might also be included) and containing the following variables: the value of each age and exposure group, the number of person-years n_i accumulated in this category over the entire cohort and the number d_i of events of interest falling in this category.

In epidemiological cohort studies the standard model for analysing such data is the Poisson model which is a statistical model of the disease rates. Basically the Poisson model assumes that the number of events d_i in each category i (combination of age category j and the kth combination of exposure variables) follows a Poisson distribution with parameter $n_i\lambda_i$. The standard (multiplicative) model would then assume that

$$\ln(\lambda_i) = \alpha_j + \beta_k \tag{5.16}$$

where λ_i are the unknown true disease rates, the α_j are nuisance parameters specifying the effects of age and (possibly) other stratification variables like calendar periods and β_k the parameters that describe the effects of primary interest. As usual in regression models $\beta_0 = 0$ would be a baseline category. $\exp(\beta_k)$ is then an estimate, adjusted on the nuisance parameters, of the relative risk of the kth exposure category vs. the baseline category assuming absence of interaction between exposure. The full modelling strategy of the Poisson regression is beyond the scope of this chapter but is not different from any regression modeling (see Chaps. II.3 and II.4 of this handbook). A comprehensive account of Poisson modeling is given by Breslow and Day (1987, Chap. 4).

An alternative way of analysing event history data (another denomination of cohort data focussed on events), is by using Cox' proportional hazard model. This model acknowledges that the categorisation of continuous data always implies a loss of information and therefore a loss in statistical power. Moreover, there is no need to explicitly estimate the effects of nuisance parameters if it can be avoided.

The first step in proportional hazard model is the choice of one of the time variables considered. This basic time variable can either be age as was implicit at the beginning of this chapter, but in some settings, this variable can be the calendar time or even the time since the beginning of follow-up. Once this special time variable has been fixed, its effects are estimated nonparametrically.

The key idea of Cox's regression is that no information is lost when considering only the time points t_i at which an event of interest occurs. At each such time point a “risk set” is set up including all members of the cohort contributing person-time (at risk and under observation) at this time point. If one wants to use a Cox model, the first step is thus to identify all risk sets. Then, one must obtain the value at each time t_i of all variables to be included in the model for all members of the corresponding risk set. The statistical analysis is then similar (in fact the same software can be used) to a conditional logistic regression analysis, in which the matching variable is the indicator of the risk set. As in the logistic regression, the exposure at time t_i of the case, i.e. the subject experiencing the event at time t_i, and the exposure at time t_i of the other members of the risk set are compared. Again, the full modelling strategy of the Cox proportional hazard model and its various extensions are beyond the scope of this chapter (see Chaps. II.3 and II.4). A comprehensive account of this model is given by Breslow and Day (1987, Chap. 5). As for Poisson models, both Stata and Epicure provide easy to use software, but once the risk sets and the corresponding exposure variables have been computed for each risk set, any logistic regression package (e.g. Proc PHREG in SAS) can be used.

5.3.6 Internal Versus External Comparisons

In Sect. 5.3.3 the event rate (morbidity or mortality) of a cohort is compared to the rates of an external population. This is done by comparing the observed number of deaths in the cohort with the expected numbers, given the age structure of the cohort and the age-specific mortality rates λ_i of a reference population. The ratio of observed to expected (the SMR) is then interpreted as a rate ratio between the cohort and the general population taken as a reference.

If the cohort is set up for investigating a specific risk factor, as would be the case in an occupational cohort, one can be tempted to interpret the SMR as a risk ratio due to the risk factor under investigation. However, this interpretation would only be valid if the cohort were comparable to the general population for all factors except for the risk factor under investigation. This is obviously only rarely the case. The general population consists of all subjects including the very ill and very poor, which would rarely be included in the same proportion in a cohort. Thus the mortality in the general population is usually higher than in any (unexposed) cohort. In occupational cohorts, this phenomenon has been termed the "*Healthy Worker Effect*" (see e.g. Li and Sung 1999; Goldberg and Luce 2001). Other factors, like regional differences, owing to social, behavioural, nutritional and environmental factors, might cause the mortality of a regionally based cohort to be different from a nationwide general population. In summary, the SMR is a biased estimate of the effect of any risk factor.

This bias can be reduced by choosing a reference population which is as comparable as possible (except for the risk factor of interest) to the cohort under investigation. This implies to carefully select the reference population and in the end to compare the cohort to another reference cohort. In this case, however, the computation of the confidence interval of the SMR is no longer valid as it assumes that, because of the large number of subjects in the reference population, the disease rates and hence the expected numbers are observed without any sampling error. In this case, the only statistically valid methods are those presented in the preceding section, although the confidence intervals of the risk ratio become wider. The choice between an external comparison and an internal comparison is thus the choice between accepting an (often small) bias and accepting a larger variance, which implies a lower power. Such a choice can only be made in the context of each study and, if possible, both approaches should be tried. Finally, methods have been proposed including external reference rates to stabilise internal comparisons (e.g. Breslow and Day 1987, p 151) that might be used as reasonable compromise.

5.4 Key Concerns in Cohort Studies

5.4.1 Selection of the Study Population

Usually, vital statistics data of the general population, or data derived from national disease registries are used as a reference for the calculation of expected cases. However, they can only be regarded as valid for deriving an expectation of mortality and disease rates if the cohort under investigation is a representative sample of the general population. Indeed, many cohorts are convenience samples, derived from a group that happens to be accessible. Representative cohorts can for example be derived from national censuses, utilizing the data collected for the specific census. Obtaining access to census data is generally not easy, since most censuses guarantee confidentiality to participants. Exceptions to that rule are for example, a Swedish occupational census-based sample or a 10% sample of the Canadian labour force, derived from data collected from Canadian having a social insurance number that is required for all who are employed in an active occupation (Howe and Lindsay 1983). These types of population samples are very valuable, because subsets among them chosen for specific analysis can be regarded as comparable to the general population apart from the characteristics that caused them to enter, or be selected for, that subset.

Occupational cohorts (cf. Chap. III.2 of this handbook) are usually identified by company files or sometimes by workers' union files. Access to these cohorts is usually granted, if the company or union is interested in determining whether a suspected increase in disease rates has occurred, or there is concern that exposure to a potential hazard bears an increased risk of disease. Many carcinogens have been confirmed in humans, after first evidence from animal studies, by investigations of specific cohorts (Tomatis et al. 1990). This mechanism is still being used, as exhibited by a tri-utility study of electrical and magnetic field exposures (Theriault et al. 1994), and a study of Motorola employees on the potential risks of exposure to radiofrequency fields (Morgan et al. 2000). It is very helpful, if employment records indicate exposure to specific agents. This is the case when routine measurements are taken for safety reasons, as for most workers exposed to radiation. In their absence estimation of exposures may be required, as discussed further below.

So-called multi-purpose cohorts identified for study, however, have to be recruited by some mechanism that provides the opportunity for potential subjects to volunteer. For example, much has been learnt from an ongoing study of American nurses, who were given the opportunity to volunteer for the study by completing a questionnaire of dietary and other lifestyle factors (Willett et al. 1992). Similar studies were initiated in Canada by providing self-administered questionnaires to women already participating in a mammography screening trial (Howe et al. 1991) and in Sweden by approaching women who participated in a routine mammography screening programme (Wolk et al. 1998). In Europe, a large multi-centre cohort study was initiated in 10 countries using different approaches (Riboli and Kaaks 1997). Some used population registers as the basis for mailing invitations to participate. The response proportions were good in most countries, but still tended to include more health conscious and more highly educated people than the general population as is often the case in volunteer studies (cf. Chap. I.10 of this handbook).

Another recent feature of cohort studies has been the attempt to bring many together and analyse them almost as a multicentre study to enable the investigators to identify risks which none of them individually were capable of demonstrating. The Pooling Project is a case in point, originally funded to evaluate further uncertain associations between diet and breast cancer, it has proven a very useful source of additional knowledge because of the ability of cohort studies to identify multiple endpoints. Thus it has already been extended to lung cancer (Smith-Warner et al. 2003), with findings similar to the EPIC study (Miller et al. 2004), and other diseases will follow.

When a truly representative cohort cannot be obtained, because the mechanism used involves the opportunity to volunteer, and to refuse to participate, comparisons with the general population in terms of mortality and disease rates may not be valid. Thus the cohort may lack external validity. However, provided that the recruitment mechanism is unbiased with regard to the exposure of interest, and the data obtained on exposure enables the investigators to stratify their population into exposed and unexposed subgroups, the estimation of the association between the exposure and the outcome will be valid (internal validity).

Tables 5.4 and 5.5 demonstrate the effects of different participation patterns (selection) on estimates that can be obtained from cohort studies. In the presence of a fair sample, all of the measures of disease occurrence and association will be unbiased (Table 5.4). In the presence of over-representation of exposed persons (Table 5.5), the prevalence of the exposure will be overestimated and the risk of the outcome will be over- or under-estimated depending on whether the exposure is positively or negatively associated with disease. Nevertheless, the estimates of the relative risk and the attributable risk will be unbiased. Since the estimate of the prevalence of exposure is biased, estimates of the public health impact will be biased. Other participation patterns that can theoretically introduce selection bias including over-representation of diseased individuals and participation rates

Table 5.4. Effects of a fair sampling process on the measures of disease occurrence and association

Target Population

		Disease		
		Yes	No	
At Risk	Yes	40	160	200
	No	60	740	800
		100	900	1000

→

Selection Weights

50	50
50	50

→

Study Population

		Disease		
		Yes	No	
At Risk	Yes	20	80	100
	No	30	370	400
		50	450	500

Target Population		Study Population
10%	Prevalence of Disease	10%
20%	Prevalence of Risk Factor	20%
2.67	Relative Risk	2.67
3.08	Odds Ratio	3.08
125/1000	Attributable Risk	125/1000

Table 5.5. Effects of oversampling of exposed individuals on the measure of disease occurence and assiociation (positive association between exposure and outcome)

Target Population

		Disease		
		Yes	No	
At Risk	Yes	40	160	200
	No	60	740	800
		100	900	1000

→

Selection Weights

100	100
10	10

→

Study Population

		Disease		
		Yes	No	
At Risk	Yes	40	160	200
	No	6	74	80
		46	234	280

Target Population		Study Population
10%	Prevalence of Disease	16%
20%	Prevalence of Risk Factor	71%
2.67	Relative Risk	2.67
3.08	Odds Ratio	3.08
125/1000	Attributable Risk	125/1000

that differ by both, exposure and disease status, are unlikely to affect cohort studies due to the customary exclusion of persons with the outcome of interest at baseline. This assurance is only relative, relying on the degree to which persons with prevalent disease can be excluded from the cohort. In general, selection bias can be minimized by avoiding the use of volunteers (or using volunteers exclusively) and by minimizing non-participation. The potential for selection bias can be assessed by evaluating non-participants for study characteristics, if possible.

5.4.2 Exposure and Confounders in Cohort Studies

As already indicated, some cohorts will have exposure data readily available, especially those derived from occupational groups where exposure was routinely collected for safety monitoring purposes. It is the strength of such cohorts that they offer the possibility to report the exposure before the disease occurs. However, for population-based cohorts, the investigators will have to collect data specifically for the study, or to refine existing data.

Because most cohorts will be very large, the collection of exposure data is not a simple task. If exposure data is to be collected by questionnaires, the scale of the effort required will generally mean that neither personal nor telephone interviews are feasible, as would normally be planned for case-control studies. This means that the exposure data will generally be collected by mailed self-administered questionnaires, often linked to the recruitment mechanism of the cohort, with response to the questionnaire qualifying the individual for inclusion in the study. Inevitably, the amount of data that can be collected by self-administered questionnaire is limited. The degree of detail for a given variable that can be obtained by such instruments is also restricted (cf. Chap. I.11 of this handbook), so that in addition to the problems of the ability of the respondent to recall accurately the exposure he/she has experienced, the data will be potentially subjected to major misclassification.

The extent of misclassification in cohort studies has only recently been appreciated, probably explaining the fact that the results of many cohort studies, especially when diet was the exposure of interest, have been negative (Day and Ferrari 2002). Thus although many of the questionnaires used in cohort studies have been subject of validation studies, and correlation with other assessment methods seemed reasonable, these validation studies have served to reassure the investigators, but probably have not protected them from reporting negative, or very weak results. Even for smoking, the information obtained in cohort studies cannot be regarded as precise as investigators would have wished.

Misclassification of exposure can be differential or non-differential with respect to the outcome of interest; that is, the degree of misclassification of the exposure can differ, or not, by outcome status. In cohort studies, non-differential misclassification is the more typical form of misclassification due to the customary exclusion of persons with prevalent disease at baseline. It is unlikely that the measurement of exposure at baseline will be influenced by the development of an outcome sometime in the future. Differential misclassification is potentially a much greater problem in case-control and cross-sectional studies. Non-differential misclassification always introduces a bias toward a null finding (a finding of no association) if the exposure

Table 5.6. Non-differential misclassification of exposure

		True Cohort (no error)					Observed Cohort (error)		
		MI	No	MI			MI	No	MI
Cigarettes	Yes	60	300	360	Cigarettes	Yes	54	270	324
	No	30	330	360		No	36	360	396
		90	630	720			90	630	720

2.00	Rate Ratio	1.83

status is dichotomized; whereas, differential misclassification can introduce a less predictable bias. Table 5.6 shows the impact of a 10% non-differential error rate in classifying smokers. In this example, 90% of exposed individuals were correctly classified regarding exposure and 100% of unexposed individuals were correctly classified. Assuming a true relative risk of 2.0, the observed relative risk would be 1.8. With greater degrees of misclassification, the bias towards the null would increase. This bias can be minimized through the use of standardized and validated procedures for exposure assessment.

Another issue that affects cohort studies differently than case-control studies is the effect of change in exposure with time. In case-control studies detailed exposure biographies that include changes in exposure patterns, e.g. change in intensity of smoking, or cessation of smoking, or even measures taken to affect dietary change, can be retrieved using just one survey, with the problem of uncertainty, and possibly differential error, in recall. The concurrent cohort study with its prospective data collecting does offer the possibility of assessing changes in exposure while they happen. To assess changes in exposure patterns, a mechanism has, however, to be set up specifically e.g. by re-administering the questionnaire on a regular basis. This could be done as part of the follow-up mechanism adopted, though some loss to follow-up will be inevitable. An alternative to incorporating this new information into the analysis is shown in the Nurses Health Study (Willett et al. 1992). The follow-up period with regard to the time from the first exposure information to the second was used as a separate cohort from the follow-up period subsequent to the second exposure information. This is justifiable as blocks of person-time in different periods are statistically independent, regardless of the extent they are derived from the same people (Rothman and Greenland 1998). However, sometimes cohorts are analysed with regard to the exposure determined at baseline, and although that may seem distant from the period when many endpoints are determined, for those with a long induction period from exposure to outcome, as for many cancers, this has not always been regarded as a major disadvantage.

Exposure assessment by questionnaires always depends on subjects' accuracy of recall and their willingness to participate, and many efforts have been made to introduce more objective measures of exposure determination. For radiation exposure, cohorts with occupations that require wearing film badges provide cumulative, and in some instances, peak measurements of exposure. For uranium and other hard rock miners, measures of the radiation exposure in mines were often made for safety reasons to limit the length of exposure of those at risk and these measurements can be assigned to the job history of the individual.

However, in many instances, exposure has to be estimated simply from the type of occupation at a certain time since no further information is available, and misclassification of exposure assessment cannot be avoided. In occupational studies, attempts have been made to refine exposure assessment by developing a job exposure matrix (cf. Chap. I.11). Often using data from hygiene assessments performed in the past, a matrix can be constructed with the different job tasks in the rows, and columns indicating the probability and/or intensity of exposure within that job to the agents (chemical or physical) of interest. The approach was for example used in a study of electrical and magnetic field exposures in electric utility workers in Canada and France (Theriault et al. 1994). Extension of the work upon a sample of workers wearing portable electric and magnetic field exposure meters, and using historical data of electrical usage in the province enabled the investigators to identify strong associations of leukaemia and non-Hodgkin's lymphoma risk with high electric field exposure (Miller et al. 1996; Villeneuve et al. 1998).

Another source of exposure data collected in cohort studies is gained from biological material of the cohort members. Historically, rather simple parameters were under study, like blood pressure or cholesterol levels, derived from blood samples that were collected in the framework of large cohort and intervention studies on cardiovascular disease. Now, there is increasing interest in the study of disease aetiology by biomarkers of exposure and/or of genetic factors, as e.g. in the European Prospective Investigation of Diet and Cancer (EPIC) (Riboli and Kaaks 1997).

The findings of cohort studies regarding the effects of exposure can be strengthened if it is possible to evaluate a dose-response relationship. This requires the assessment of intensity of exposure that can be quantified as peak, average, or cumulative exposure. Sometimes duration of exposure is used as a surrogate for cumulative exposure. However, using duration in this way is problematic if the exposure is associated with an early, perhaps toxic effect. Then it could be anticipated that these workers would tend to change their employment and could not cumulate long durations of exposure. If such workers represented a particularly susceptible subgroup, perhaps for genetic reasons, it is possible that in this subgroup a relatively brief exposure results in the same incidence of disease than in subgroups with a longer duration of exposure that are less susceptible. The absence of a dose-response relationship without appropriate statistical control for the genetic background might then be incorrectly interpreted as indicator that the exposure is not causal for the disease (Blair and Stewart 1992).

The treatment of potential confounding factors is the major challenge of the analysis of cohort studies. This is in part because the basic data set may not contain information on all relevant confounders, particularly not in historical cohort studies, but also because the data available on confounders may not be assessed with sufficient precision to take account of their effect. An example is the possible confounding effect of cigarette smoking with fruit and vegetable consumption and lung cancer. Although two large cohort studies (one multicentre and one the result of a pooled analysis) which fully adjusted for the effects of cigarette smoking in the opinion of the investigators were available (Miller et al. 2004; Smith-Warner et al. 2003) a working group of the International Agency for Research on Cancer (IARC) was not convinced that there was not residual confounding of fruit consumption by smoking with lung cancer, and therefore judged the evidence to be limited rather than sufficient (IARC 2003).

Determining Outcome Events 5.4.3

A limiting factor for cohort studies is that most diseases are relatively rare, with rates determined in the population per 100,000 persons. Therefore to accrue sufficient cases of the disease the size of the cohort has to be large, and/or the follow-up time has to be long. Another factor affecting the length of follow-up relates to the long induction period from the beginning of many exposures to the occurrence of disease. For many cancers, for example, the induction period exceeds ten, often 20 years. One example for the importance of a long enough follow-up period is the British Doctors' Study that showed much higher lung cancer risks of cigarette smoking after 40 years of follow-up than in the ten- and twenty-year reports of this study (Doll et al. 1994b). The reason for this was a dominant effect of duration of smoking compared to intensity of exposure on the risk of lung cancer (see also Flanders et al. 2003). It seems probable that this is not the only example of this phenomenon – it may particularly affect exposures with a long induction period from initiation of exposure to effect. The possibility of such an effect should encourage investigators to maintain the follow-up of well documented cohorts for as long as proves feasible, and granting agencies will agree to provide the necessary funds. If grants are limited it may be useful to store the necessary data and extend the follow-up after a certain time lapse. It is unusual for cohort studies to start from the first exposure and the possible initiation of disease, covering the whole spectrum of exposure in a subject's lifetime. Attempts have to be made to determine or to estimate past exposure, with all the error and potential misclassification of such inquiries. Nevertheless, a major advantage of cohort studies over case-control studies is that exposure is determined prior to the diagnosis of disease, thus avoiding a major bias of concern in case-control studies, the recall bias.

As already indicated, the follow-up of cohorts enables multiple endpoints to be determined, e.g. different types of cardiovascular disease and/or different cancer sites. In determining endpoints in cohort studies, it is essential that ascertainment bias is avoided. Ascertainment bias relates to the possibility that the surveillance of cohort members, by virtue of the fact that they are in a study, may result in greater efforts to make a diagnosis than would occur in the general population. Special surveillance mechanisms in a cohort study are valid if internal comparisons of exposed versus unexposed within the cohort are planned, but would invalidate external comparisons with general population data. Orencia and colleagues (1995) provided an example of this bias in a non-concurrent cohort study examining the association of mitral valve prolapse (MVP) with stroke. Using the database of the Mayo Clinic, they assembled a cohort of persons with MVP, followed them for the occurrence of stroke, and compared the rate of stroke with the rate in the general population of Olmsted County, Minnesota. The overall standardized mortality ratio was 2.1, indicating a risk of stroke twice of that of the general population. However, Orencia noted that MVP can be diagnosed by auscultation or as a serendipitous finding during an echocardiogram conducted for other medical reasons (e.g. following myocardial infarction, chronic heart failure, atrial fibrillation) often associated with risk of stroke. When the cohort was further subdivided according to method of diagnosis, the auscultatory group demonstrated no increase in risk. The increased risk was confined to the group identified serendipitously during a cardiac evaluation motivated by other medical concerns associated with risk of stroke.

In some cohort studies, annual or less frequent contact by mail, generally with the cohort member directly, or sometimes with his or her designated physician, will identify the probable occurrence of a study endpoint, or death from a cause unrelated to the disease of interest. However, these processes are costly, and also pose the risk of losing an increasing proportion of cohort members with time. Further, if the participant has died, family members may not always be willing to collaborate in providing the required information. Hence, in many studies, other mechanisms are used for follow-up, and indeed may have to be used also for subjects lost if the basic mechanism of follow-up is by mail. Losses to follow-up lead to a loss of power due to the resultant loss of sample size and can introduce bias in a manner similar to the selection processes described previously. Losses that do not differ by either exposure or disease status result in a picture similar to that shown in Table 5.4, that is, no bias, but a loss of power. Losses that differ by exposure (but not outcome) status introduce the same bias as that described in Table 5.5. More problematic are losses that differ by outcome status (Table 5.7) and those that differ by both exposure and outcome status (Table 5.8). In these situations, estimates of the relative risk may be biased in unpredictable directions.

Apart from special surveillance mechanisms, including screening for the disease of interest, there are many sources of routinely collected data for endpoints in cohort studies. These include medical records of physicians, health maintenance organizations and hospitals, vital statistics systems and disease registries. The process to determine whether a particular record relates to a cohort member involves some form of record linkage, determining whether the identifying data in the study file of a cohort member corresponds with the identifying data on the medical or other record of endpoint information. In the past, much of this linkage used to be done manually. Increasingly some form of computerised record linkage is performed. Although such linkages are easier if both sets of records contain the same (national) identifying number, computerised record linkage can still be extremely efficient, and less costly than individual-based follow-up. If record linkage is planned to determine endpoints in a cohort study, great care should be taken at the time of recruitment to collect sufficient identifying information for record linkage purposes, this includes full name, full date of birth, place of birth, mothers maiden name, social security number, other identifying number (if available), and current address. Further, the name and address of friends or relatives of the cohort member should also be collected, to facilitate tracing an individual if other means of tracing them have failed, or if record linkage to another data source has resulted in an uncertain linkage.

In many countries, in addition to disease registries, such as cancer registries, there are other data sources that have been developed to facilitate record linkage for cohort studies and large scale trials. These include the National Health Service Central Register in the UK, the Canadian National Mortality Data Base, the National Death Index in the USA, and similar national registers in the Scandinavian countries. Relatively new in this context are the population-wide registries of genetic data, like the registry already established in Iceland or the one planned in Estonia. Record linkage using these national data bases overcomes many of the issues regarding confidentiality of data, as confidentiality procedures are readily available for such systems. In Canada, what is returned to the investigator is generally anonymous (i.e. stripped of personal identifiers), unless the subjects have signed a prior consent form that specifically permitted record linkage. This was

Table 5.7. Effects of losses to follow-up that differ by outcome status on estimates of disease occurence and assiociation

Target Population

		Disease Yes	No	
At Risk	Yes	40	160	200
	No	60	740	800
		100	900	1000

→

Selection Weights

100	10
100	10

→

Study Population

		Disease Yes	No	
At Risk	Yes	40	16	56
	No	60	74	134
		100	90	190

Target Population		Study Population
10%	Prevalence of Disease	53%
20%	Prevalence of Risk Factor	29%
2.67	Relative Risk	1.60
3.08	Odds Ratio	3.08
125/1000	Attributable Risk	260/1000

Table 5.8. Effects of losses to follow-up that differ by both exposure and outcome status on estimates of disease occurence and assiociation

Target Population

		Disease Yes	Disease No	
At Risk	Yes	40	160	200
	No	60	740	800
		100	900	1000

→

Selection Weights

50	100
100	100

→

Study Population

		Disease Yes	Disease No	
At Risk	Yes	20	160	180
	No	60	740	800
		80	900	980

Target Population		Study Population
10%	Prevalence of Disease	8%
20%	Prevalence of Risk Factor	18%
2.67	Relative Risk	1.48
3.08	Odds Ratio	1.54
125/1000	Attributable Risk	36/1000

the case, for example in a cohort study that was linked to a large multi-centre trial of breast screening (Howe et al. 1991).

5.5 Ethical Issues

It is now generally accepted that studies on humans should be carried out with informed consent. This principle, originally developed in relation to controlled clinical trials, has generally now been extended to observational epidemiology studies, including cohort studies.

In the past, if a cohort was recruited that involved the subjects participation in providing data, their agreement to supply the data (e.g. respond to a questionnaire) was generally regarded as implied consent. However, now, in addition to providing information on questionnaires, for many cohorts, biological specimens (e.g. blood, buccal cells) are requested, and then it becomes mandatory that the respondent provide consent for the future use of such specimens for research purposes. However, at the time the specimens are provided, it is impossible to know the precise use the investigators may wish to apply to this material. An example relates to the fact that the majority of participants in the sub-cohorts of the European Prospective Investigation of Diet and Cancer (EPIC; Riboli and Kaaks 1997) provided blood specimens in the early 1990s; a few without signing a consent form, the majority did so. However, now that genetic studies are commonplace on such specimens, it has become apparent that some of the consent forms did not specifically mention genetic analyses as potential research usages. This has led to difficulties in obtaining approval for such sub-studies from human experimentation committees, some of which wanted new consent forms to be signed, specific to the genetically-associated sub-study planned. Obtaining new consent, however, will become increasingly difficult as time goes on, and a number of subjects with the endpoint of interest may have died. In the United States, potential restrictions upon studies such as these have caused difficulties. In Europe, especially Scandinavia, there has been a more relaxed view of the ethical acceptability of studies on stored specimens, many such collections having been originally made without a formal informed consent process, but for which studies conducted with full preservation of confidentiality have been deemed to be ethically acceptable.

The issue as to whether respondents whose stored specimens have been tested should be informed of the results of such tests is also controversial. The European view tends to be that as the testing is being conducted as part of research, it may be impossible to interpret the results of tests for individuals, until this particular research track reaches agreed conclusions. Thus, it is not necessary, indeed possibly unethical, to inform the respondent of the results. Some consent forms specifically state this as a policy. In the United States, however, the opposite viewpoint tends to hold, say, it being regarded as ethically inappropriate for investigators to take a decision on whether or not a subject receives information on themselves. The difficulty with a universal application of such a principle is that for some, the test results may come too late for any possibility of benefit, but, especially in the case of genetic-related information, this may not preclude the test result having implications for the relatives of the subject, and such knowledge is not always a blessing. However, all would agree that if a test reveals information of potential benefit to a subject, they should be informed.

The question of consent for historical cohort studies in general does not arise, though again, there may be issues on informing subjects of the findings of the research. In general, as the research is unlikely to harm the individuals, and providing confidentiality is maintained, human experimentation committees will approve such studies.

One further ethical issue has already been mentioned in Sect. 5.4.3, and that relates to the use of record linkage in obtaining outcome data. In general, providing full confidentiality is maintained, this should not cause difficulties in obtaining approval from human experimentation committees. For further discussions of ethical aspects we refer to Chap. IV.7 of this handbook.

Conclusions 5.6

Cohort studies are a critical method for evaluating causality in epidemiology, and may also be used in evaluating screening (see Chap. III.10 of this handbook).

There are, however, several needs if they are to be valid. You need skilled investigators being familiar with the peculiarities of the planning and the conduct of cohort studies, a sensible source for cohort recruitment, evaluable hypotheses to consider, a validated questionnaire for use at enrolment, unbiased mechanisms to administer the questionnaire as well as for follow-up, quality controlled procedures to collect biological material if relevant for the question under research, facilities for data entry and of course the expertise as well as the facilities for analysis and interpretation.

Cohort studies are often rated at a higher level than case-control studies, largely because the latter are susceptible to recall bias. However, both are usually regarded as "level II" evidence (level I are randomised controlled trials) and there are potential deficiencies in cohort studies that may be less intrusive than in case-control studies, especially a greater propensity for measurement error. Both, however, continue to have an important role in disease epidemiology.

References

Blair A, Stewart PA (1992) Do quantitative exposure assessments improve risk estimates in occupational studies of cancer? Am J Ind Med 21:53–63

Benichou J (1998) Absolute risk. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics. Wiley, NewYork

Breslow NE, Day NE (1987) Statistical methods in cancer research. Volume II: The design and analysis of cohort studies. IARC Scientific Publications No. 82. International Agency for Research on Cancer, Lyon

Case RA, Hosker ME, McDonald DB, Pearson JT (1954) Tumours of the urinary bladder in workmen engaged in the manufacture and use of certain dyestuff intermediates in the British chemical industry. Br J Ind Med 11:75–104

Coleman M, Douglas A, Hermon C, Peto J (1986) Cohort study analysis with a FORTRAN computer program. Int J Epidemiol 15:134–137

Day NE, Ferrari P (2002) Some methodological issues in nutritional epidemiology. In: Riboli E, Lambert A (eds) Nutrition and lifestyle: Opportunities for cancer prevention. IARC Scientific Publications No. 156. International Agency for Research on Cancer, Lyon, pp 5–10

Doll R, Hill AB (1954) The mortality of doctors in relation to their smoking habits: A preliminary report. Br Med J 1:1451–1455

Doll R, Peto R (1976) Mortality in relation to smoking: 20 years' observations on male British doctors. Br Med J 2:1525–1536

Doll R, Peto R, Hall E, Wheatley K, Gray R (1994a) Mortality in relation to consumption of alcohol: 13 years' observations on male British doctors. Br Med J 309:911–918

Doll R, Peto R, Wheatley K, Gray R, Sutherland I (1994b) Mortality in relation to smoking: 40 years' observations on male British doctors. Br Med J 309:901–911

Eigenbrodt ML, Mosley TH, Hutchinson RG, Watson RL, Chambless LE, Szklo M (2001) Alcohol consumption with age: a cross-sectional and longitudinal study of the Atherosclerosis Risk in Communities (ARIC) study, 1987–1995. Am J Epidemiol 153:1102–1111

Flanders WD, Lally CA, Zhu BP, Henley SJ, Thun MJ (2003) Lung cancer mortality in relation to age, duration of smoking, and daily cigarette consumption: Results from Cancer Prevention Study II. Cancer Res 63:6556–6562

Goldberg M, Luce D (2001) Selection effects in epidemiological cohorts: Nature, causes and consequences. Rev Epidemiol Sante Publique 49:477–492

Howe GR, Friedenreich CM, Jain M, Miller AB (1991) A cohort study of fat intake and risk of breast cancer. J Natl Cancer Inst 83:336–340

Howe GR, Lindsay JP (1983) A follow-up of a ten-percent sample of the Canadian labor force. I. Cancer mortality in males, 1965–1973. J Natl Cancer Inst 70:37–44

IARC Working Group (2003) IARC handbooks on cancer prevention, vol 8: Fruits and vegetables. International Agency for Research on Cancer, Lyon

Li CY, Sung FC (1999) A review of the healthy worker effect in occupational epidemiology. Occup Med 49:225–229

McCray E (1986) Occupational risk of the acquired immunodeficiency syndrome among health care workers. N Engl J Med 314:1127–1132

Miller AB, To T, Agnew DA, Wall C, Green LM (1996) Leukemia following occupational exposure to 60-Hz electric and magnetic fields among Ontario electric utility workers. Am J Epidemiol 144:150–160

Miller AB, Altenburg H-P, Bueno de Mesquita, B, Boshuizen HC, Agudo A, Berrino F, Gram IT, Janson L, Linseisen J, Overvad K, Rasmuson T, Vineis P, Lukanova A, Allen N, Amiano P, Barricarte A, Berglund G, Boeing H, Clavel-Chapelon F, Day NE, Hallmans G, Lund E, Martinez C, Navarro C, Palli D, Panico S, Peeters PH, Quiros JR, Tjonneland A, Tumino R, Trichopoulou A, Trichopoulos D, Slimani N, Riboli E (2004) Fruits and vegetables and lung cancer: Findings from the European Prospective Investigation into Cancer and Nutrition. Int J Cancer 108:269–276

Morgan RW, Kelsh MA, Zhao K, Exuzides KA, Heringer S, Negrete W (2000) Radiofrequency exposure and mortality from cancer of the brain and lymphatic/hematopoietic systems. Epidemiology 11:118–127

Morris JN, Heady JA, Raffle PA, Roberts CG, Parks JW (1953a) Coronary heart disease and physical activity of work. Lancet 265:1053–1057

Morris JN, Heady JA, Raffle PA, Roberts CG, Parks JW (1953b) Coronary heart disease and physical activity of work. Lancet 265:1111–1120

National Institutes of Health (2004) Request for information: design and implementation of a large-scale prospective cohort study of genetic and environmental influences on common diseases. (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-04-041.html) Accessed May 11, 2004

Oomen CM, Ocke MC, Feskens EJ, van Erp-Baart MA, Kok FJ, Kromhout D (2001) Association between trans fatty acid intake and 10-year risk of coronary heart disease in the Zutphen Elderly Study: a prospective population-based study. Lancet 357:746–751

Orencia AJ, Petty GW, Khandheria BK, Annegers JF, Ballard DJ, Sicks JD, O'-Fallon WM, Whisnant JP (1995) Risk of stroke with mitral valve prolapse in population-based cohort study. Stroke 26:7–13

Preston DL, Lubin JH, Pierce DA, McConney ME (1993) Epicure user's guide. HiroSoft International Corp, Seattle

Riboli E, Kaaks R (1997) The EPIC project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. Int J Epidemiol 26 Suppl 1:S6–14

Rinsky RA, Smith AB, Hornung R, Filloon TG, Young RJ, Okun AH, Landrigan PJ (1987) Benzene and leukemia. An epidemiologic risk assessment. New Engl J Med 316:1044–1050

Rittgen W, Becker N (2000) SMR analysis of historical follow-up studies with missing death certificates. Biometrics 56:1164–1169

Rothman KJ, Greenland S (1998) Modern epidemiology, 2nd edn. Lippincott Williams & Wilkins, Philadelphia

Sahai H, Khurshid A (1996) Statistics in epidemiology. Methods, techniques, and applications. CRC Press, Boca Raton New York London Tokyo

Silcocks P (1994) Estimating confidence limits on a standardised mortality ratio when the expected number is not error free. J Epidemiol Community Health 48:313–317

Smith-Warner SA, Spiegelman D, Yaun SS, Albanes D, Beeson WL, van den Brandt PA, Feskanich D, Folsom AR, Fraser GE, Freudenheim JL, Giovannucci E, Goldbohm RA, Graham S, Kushi LH, Miller AB, Pietinen P, Rohan TE, Speizer FE, Willett WC, Hunter DJ (2003) Fruits, vegetables and lung cancer: a pooled analysis of cohort studies. Int J Cancer 107:1001–1011

Snow J (1855) On the mode of communication of cholera. Churchill, London

Sutherland J (2002) EXTRACTS from appendix (A) to the Report of the General Board of Health on the Epidemic Cholera of 1848 & 1849. Int J Epidemiol 31:900–907

StataCorp (2001) Stata statistical software: Release 7.0. Stata Corporation, College Station

Terris M (ed) (1964) Goldberger on pellagra. Louisiana State University Press, Baton Rouge

Theriault G, Goldberg M, Miller AB, Armstrong B, Guenel P, Deadman J, Imbernon E, To T, Chevalier A, Cyr D (1994) Cancer risks associated with occupational exposure to magnetic fields among electric utility workers in Ontario and Quebec, Canada, and France: 1970–1989. Am J Epidemiol 139:550–572

Tomatis L, Aitio A, Day NE, Heseltine E, Kalder J, Miller AB, Parkin DM, Riboli E (eds) (1990) Cancer: Causes, occurrence and control. IARC Scientific Publications No. 100. International Agency for Research on Cancer, Lyon

Ulvestad B, Kjaerheim K, Martinsen JI, Mowe G, Andersen A (2004) Cancer incidence among members of the Norwegian trade union of insulation workers. J Occup Environ Med 46:84–89

Villeneuve PJ, Agnew DA, Corey PN, Miller AB (1998) Alternate indices of electric and magnetic field exposures among Ontario electrical utility workers. Bioelectromagnetics 19:140–151

Wada S, Miyanishi M, Nishimoto Y, Kambe S, Miller RW (1968) Mustard gas as a cause of respiratory neoplasia in man. Lancet 1:1161–1163

Willett WC, Hunter DJ, Stampfer MJ, Colditz G, Manson JE, Spiegelman D, Rosner B, Hennekens CH, Speizer FE (1992) Dietary fat and fiber in relation to risk of breast cancer. An 8-year follow-up. J Amer Med Assoc 268:2037–2044

Wolk A, Bergstrom R, Hunter D, Willett W, Ljung H, Holmberg L, Bergkvist L, Bruce A, Adami HO (1998) A prospective study of association of monounsaturated fat and other types of fat with risk of breast cancer. Arch Intern Med 158:41–45