2
Study Design: Population-Based Studies

Janet Cade[1] and Jayne Hutchinson[2]

[1] *University of Leeds*
[2] *University of York*

Key messages

- Ecological studies may be the first step in generating hypotheses concerning diet and disease relationships, but they are limited by the 'ecological fallacy'. This occurs when relationships that are observed for groups are assumed to hold for individuals.
- In population studies the researcher has no control over the exposure of interest (the diet).
- Confounding of diet–disease relationships is a possibility in observational studies; relationships seen between diet and disease can change considerably when confounders are included. Confounders are variables that affect both the exposure (diet) and the outcome (disease).
- Cross-sectional studies measure exposure and disease at the same point in time and so cause and effect cannot be determined.
- Case-control studies are subject to recall bias due to problems in reporting past diet.
- Cohort studies are often large, long-term studies in which recall bias is avoided if exposure data is collected before outcome data. However, they are expensive and not particularly useful for rare diseases.

2.1 Introduction

This chapter will discuss population-based, observational studies. The methods used are based on epidemiological approaches; epidemiology is the study of diseases in populations. The key consideration in population-based studies is that the researcher has no control over the exposure of interest (e.g. diet). Study types include ecological, case-control and cohort studies. They are useful for generating hypotheses and exploring associations between diet and health outcomes. These study designs can help to build up evidence to support a suggested effect of a particular dietary factor on a certain disease, but they cannot categorically show cause-and-effect association, which is required for proof of a link between a dietary factor and a disease. Since these methods do not use randomisation to select participants, they are more prone to bias than are randomised controlled trials (RCTs). Bias is a systematic error resulting in an estimated association between exposure and outcome that deviates from the true association in a direction that depends on the nature of the systematic error. Selection bias can result in systematic differences between characteristics of participants in different exposure or outcome groups within a study, which can lead to confounding of the results. Non-response bias at the start of a study and non-random attrition (dropping out of participants) during a study are other forms of selection bias. Recall bias and social desirability reporting bias are forms of measurement bias; the systematic differences in recall and reporting between exposure or outcome groups who have dissimilar characteristics can lead to confounding of the results.

Confounding variables can provide alternative explanations for an apparent association between a dietary exposure and a disease/health outcome in observational studies. Confounders are associated with both the exposure of interest (diet) and the outcome variable (disease), but are not on the causal pathway between exposure and outcome. Confounders can be dealt with in a number of ways depending on the study design: during the design of the study by matching or by restricting study members; or

Nutrition Research Methodologies, First Edition. Edited by Julie A Lovegrove, Leanne Hodson, Sangita Sharma and Susan A Lanham-New.

Companion Website: www.wiley.com/go/nutritionsociety

through data analysis by stratification (e.g. age standardisation), restriction or adjustment in regression models. Most analyses of disease risk control for age, since disease risk increases with age and age is often associated with dietary intake. Confounders are discussed further in Section 2.6.

This chapter will consider ways to minimise problems. However, its overall aim is to provide an overview of different methods used in observational epidemiology.

2.2 Ecological studies

The focus of this type of study is on characterising population groups rather than on linking individuals' exposures to health outcomes. Ecological studies of diet and health explore associations between population or group indicators of diet or nutritional status and population or group indices of health status. Two population-based measures are needed for this type of study, one for the exposure of interest (the diet) and the other for the health outcome (the disease). The individuals in the populations used to describe the dietary exposure may or may not be the same as those providing data for health outcomes. In nutritional epidemiology, ecological studies have predominantly been used to explore geographical or temporal relationships between diet and health: for example, exploring country differences in dietary intakes and health, or comparing changes in diet in populations over time.

There are occasions when ecological studies may be the only feasible research method available to explore the association between diet and disease. This would occur when exposure data are not available at the individual level, such as for fluoride in drinking water.

Methods

In the simplest study, two population-based measures are required, one for the exposure of interest and the other for the health outcome.

Indices of dietary intake

Estimates of population dietary intake can be made from survey data collected for the purpose of the study in a population or from pre-existing dietary data, which will be less costly although it may not sufficiently reflect consumption.

National food supply

An important source of internationally available food data comes from the Food and Agriculture Organization (FAO) food balance sheets, available at http://faostat3.fao.org/faostat-gateway/go/to/home/E. These provide a comprehensive picture of the pattern of a country's food supply for a particular time point. For each food item, they show the total quantity produced and imported and link this to utilisation, including export, amounts fed to livestock and used for seed, and losses during storage and transport. From this the amount of each food available for human consumption can be estimated. This type of data has been used to assess trends in dietary intakes; however, it may overestimate dietary intakes (Pomerleau, Lock and McKee 2003).

Household budget surveys

These studies collect data on food availability at a household level. Participants record food purchases and other food coming into the home. This type of data is used to generate consumer price indices, which are used as measures of inflation. A household expenditure survey, now called the Living Costs and Food Survey, has been conducted annually in the UK since 1957, making it a useful tool for monitoring changes in family food behaviour over time.

Individual survey data

Nutrition and health population-based surveys were used to estimate mean fruit and vegetable intake for the Global Burden of Disease study (Lim *et al*. 2012). Ecological analysis has been undertaken using diet and health information collected from a range of European countries included in the European Prospective Investigation into Cancer (EPIC) cohort study.

Indices of health outcomes

Routine measures of mortality and morbidity

Measures of mortality or morbidity at a national level are usually available through government reports or World Health Organization (WHO) publications. National mortality data and Global Burden of Disease data can all be found here: http://www.who.int/healthinfo/statistics/en/

A classic example

Ecological studies are generally the first step in exploring whether there is a differential distribution of disease among people with different risk profiles. For example, ecological comparisons showed that economically developed countries with a higher intake of dietary fat had much higher coronary heart disease (CHD) rates than countries with lower dietary fat consumption. This evidence was based on an early study analysing diets from groups of men in seven different countries (Keys *et al*. 1986); see Figure 2.1. These results have been challenged over the years because of difficulties in characterising the

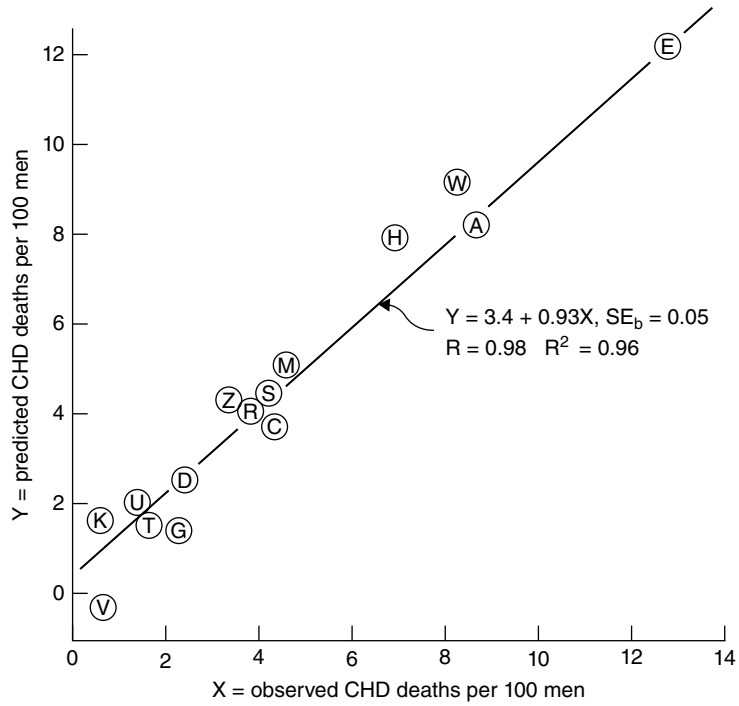

Figure 2.1 Observed 15-year death rates per 100 men compared with death rates from coronary heart disease (CHD) predicted from the multiple regression of the ratio of monounsaturated to saturated fatty acids in the diet, adjusting for age, body mass index, systolic blood pressure, serum cholesterol, and number of cigarettes smoked daily in the Seven Countries Study. Keys, A. *et al.* (1986) The diet and 15-year death rate in the Seven Countries Study, *American Journal of Epidemiology*, **124** (6), 903–915, by permission of Oxford University Press.

dietary intakes of the different country populations. Other types of study are needed to show causation.

A recent example

Diet features very strongly as a risk factor for top adverse health outcomes in the recently published Global Burden of Disease Study 2010 (Lim *et al.* 2012); see Figure 2.2. This study used published and unpublished secondary sources of data to calculate the relationships between 67 different risk factors in 21 regions and linked them with deaths or disease burden for each region between 1990 and 2010. Out of the top 20 leading risk factors contributing to the burden of disease in 2010, 6 are dietary factors (diet low in fruit, nuts and seeds, whole grains, vegetables, seafood and omega-3 fatty acids, and high in sodium) and another 7 are directly linked to diet (high blood pressure, high body mass index, high fasting plasma glucose, childhood underweight, iron deficiency, suboptimal breastfeeding and high total cholesterol). An ecological approach was employed to link risk factors to disease outcomes, using data collected via different epidemiological methods. The data do not directly link individual exposures to risk factors with the diseases of interest. Limitations include variable quality of exposure data across countries and the possibility of residual confounding (see Section 2.6), meaning that some associations could be the result of other factors that have not been considered or taken into account in the analysis.

Analysis of ecological data

The most straightforward analysis would be the calculation of a correlation coefficient between the exposure of interest and the outcome. This is a measure of the strength and direction of the linear relationship between two different continuous variables, for example energy intake and body mass index. The correlation coefficient, denoted by 'r', can have values between +1 (a perfect positive linear relationship) and –1 (a perfect inverse linear relationship). A value of 0 indicates no linear relationship between the two variables. An ecological analysis of 21 wealthy countries (Pickett *et al.* 2005) found that income inequality was positively correlated with the percentage of obese men ($r = 0.48$, $p = 0.03$). The relationship was even stronger for obese women in these countries, with a positive correlation coefficient of 0.62 ($p = 0.003$).

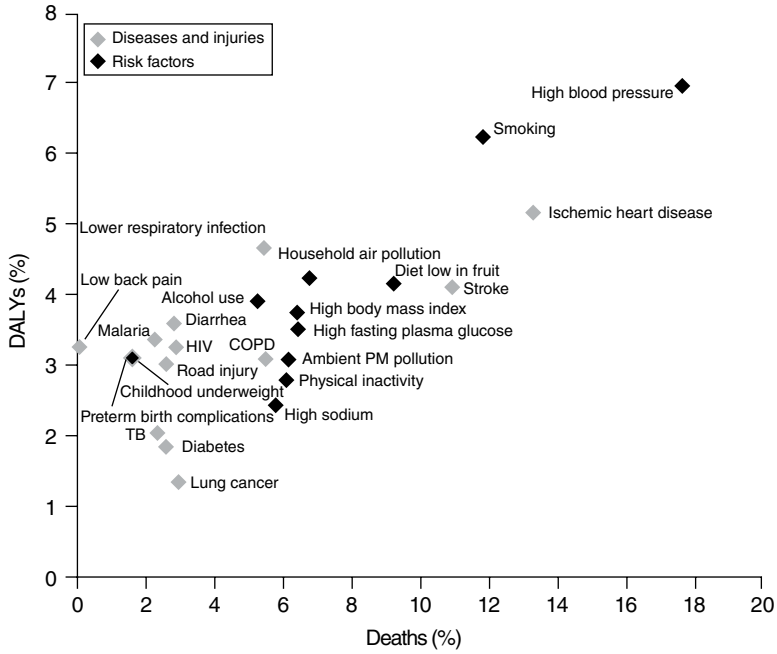

Figure 2.2 The 10 leading diseases and injuries and 10 leading risk factors based on percentage of global deaths and disability-adjusted life years (DALYs), 2010. http://www.healthmetricsandevaluation.org/gbd/publications/policy-report/global-burden-disease-generating-evidence-guiding-policy.

Further analysis of ecological studies could include multiple regression modelling to estimate the magnitude of associations, taking into account other factors of relevance that may otherwise confound the analysis. Confounding factors may include age and other lifestyle factors. Regression modelling can be undertaken using continuous variables as the dependent variable or outcome, such as height or weight. In this case linear regression modelling would be undertaken. When the outcome is categorical or dichotomous, such as the presence or absence of a disease, then logistic regression is appropriate. A study of routine data from South Australia used logistic regression analysis to assess factors that might affect food security, a dichotomised outcome (Foley *et al.* 2010). Food insecurity was highest in households with low levels of education or limited capacity to save money, and in Aboriginal households and those with three or more children.

Problems with ecological analyses

The 'ecological fallacy' is the major trap for the unsuspecting researcher. This occurs when relationships that are observed for groups are assumed to hold for individuals. For example, ecological analysis has shown that countries with more fat in the diet have higher rates of breast cancer, suggesting that women who eat fatty foods would be more likely to develop breast cancer. This assumption is only weakly supported by case-control and cohort data. Correlations found in ecological analyses may be due to confounding by other related factors that have not been controlled for, some of which may be difficult to measure at the population level. Age standardisation often needs to be undertaken, since countries may have very different age profiles. This process adjusts disease rates to a standard population, allowing comparisons to occur. When disease rates are age standardised, any differences in the rates over time or between geographical areas will not simply reflect variations in the age structure of the populations. This is important when looking at disease rates because some conditions, such as cancer, can predominantly affect the elderly. So if rates are not age standardised, a higher disease rate in one country may simply reflect the fact that it has a greater proportion of older people. Additionally, the quality of diagnostic data can differ widely between countries and over time.

2.3 Cross-sectional studies

A cross-sectional survey is a type of observational or descriptive study. The information in this type of survey represents a snapshot about the population at one point in time and it is not possible to determine whether the exposure and the outcome are causally related. Cross-sectional surveys are also known as prevalence surveys,

since they can be used to estimate the prevalence of disease in a population. The prevalence is the number of cases of a disease in the population at a particular point in time usually expressed as a rate.

A recent example

A cross-sectional analysis of data from older people in the Singapore Longitudinal Ageing Study found that higher measures of fasting homocysteine and low folate were negatively associated with measures of performance-oriented mobility and activities of daily living (Ng *et al*. 2012). Although these results are suggestive of a relationship in the direction of poorer nutrition to poorer physical function in older people, it is not possible to claim causality, primarily because temporal relationships between exposure and disease were not examined. It is equally plausible that older people with poorer physical functioning have a poorer diet and therefore a worse nutritional status. In order to prove cause-and-effect relationships a different type of study, a randomised controlled trial, would be needed.

Methods

Describing population characteristics

The major nutrition survey conducted in the UK is the National Diet and Nutrition Survey (NDNS). It is a rolling programme that began in 2008 and collects nationally representative dietary data from 1000 individuals per year aged 18 months and over from private households. The National Health and Nutrition Examination Survey (NHANES) is a major rolling programme of survey data collection in the USA that began in the early 1960s. About 5000 individuals are surveyed each year. The sample is selected to represent the US population of all ages. To produce reliable statistics, NHANES over-samples people 60 and older, African Americans, Asians and Hispanics.

There are two major aspects of national nutrition surveys that are important with respect to data collection: cost and organisation. Data should be as *nationally representative* as possible and also be as *accurate and complete* as possible (Stephen *et al*. 2013). In the NDNS, national representation in terms of age, gender and region is achieved by randomly selecting postcodes and addresses from the UK population as a whole (Figure 2.3).

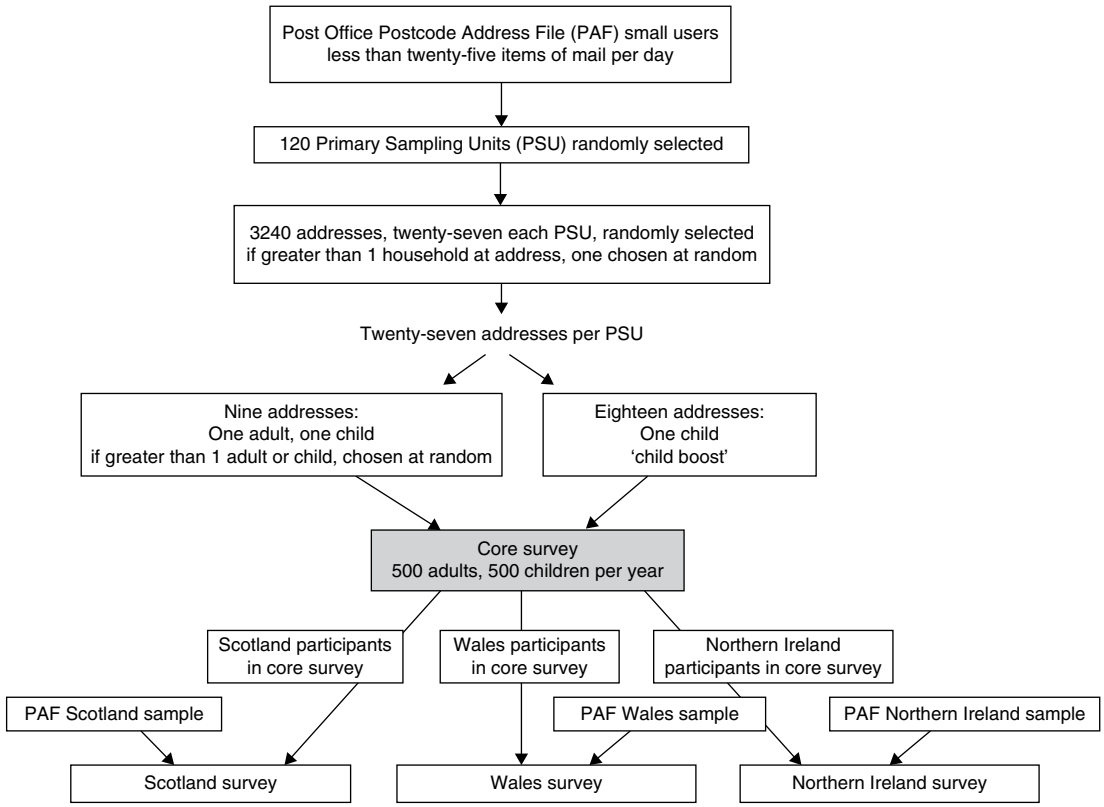

Figure 2.3 Sampling process to ensure national representation in the NDNS survey. Stephen, A.M., Mak, T.N., Fitt, E. *et al*. (2013) Innovations in national nutrition surveys, *Proceedings of the Nutrition Society*, **72** (1), 77–88. Reproduced with permission of Cambridge University Press.

The NDNS currently uses the four-day estimated diary to assess diet. This is a compromise between detail and respondent burden. Respondent burden is particularly important to consider in large-scale surveys of this kind. A high level of low-energy reporting has been found in a previous national survey of older British adults that used four-day weighed diaries, which was considered to be a result of the weighed intake method and reluctance to report consumption of unhealthy food.

Three large cross-sectional data sets from the USA, including NHANES, were used to explore causes of changing energy intake in children from 1977 to 2010. Changes in the number of eating/drinking occasions per day and portion size per eating occasion were the major contributors to changes in total energy intake per day (Duffey and Popkin 2013).

Prevalence surveys

Demographic and Health Surveys (DHS) are nationally representative household surveys that provide data for a wide range of monitoring and impact evaluation indicators in the areas of population, health and nutrition. More than 300 surveys have been conducted in over 90 countries, and survey data and results can be found at http://www.measuredhs.com/. Among the nutrition topics included and reported is the prevalence of anaemia in children and women, as well as the percentage breast fed and anthropometric indicators. High response rates, national coverage, interviewer training and standardised data-collection procedures across countries as well as consistent content over time enable comparisons to be made across populations cross-sectionally and temporally (Corsi *et al.* 2012).

Migrant studies

Cross-sectional analyses of migrants, comparing populations migrating from rural to urban areas or migrating between countries, have been undertaken to explore the associations between genetic background and environmental exposures in relation to risk of disease. Rural–urban migrants experience rapid environmental changes associated with urbanisation, enabling epidemiological transitions to be examined. Changes seen in migrants over relatively short time periods may therefore provide insights into wider population health changes. The Indian Migration Study (Bowen *et al.* 2011) explored the impact of migration to urban areas on dietary patterns, comparing migrants with their rural siblings. Migrant and urban participants reported up to 80% higher fruit and vegetable intake than rural participants ($p = 0.001$) and up to 35% higher sugar intake ($p = 0.001$). Meat and dairy intake were higher in migrant and urban participants than in rural participants ($p = 0.001$); see Figure 2.4.

Analysis of cross-sectional data

As with ecological analyses, cross-sectional data can be analysed using correlations between exposures and outcomes. In addition, regression modelling can be used to explore the influence of one continuous variable on another, while taking into account potential confounding factors.

Problems with cross-sectional studies

The main disadvantage of cross-sectional studies is that, since the exposure and disease or outcome are measured at the same time, it is not possible to say which is cause and which is effect. For example, an analysis of questionnaire data recording women's use of vitamin C supplements and irritable bowel syndrome (IBS; Hutchinson *et al.* 2011) could not be certain whether the supplementary vitamin C had been taken to prevent or manage symptoms of disorders or whether vitamin C had caused them, due to the cross-sectional nature of the data. Associations observed with IBS could have been due to abdominal pain and diarrhoea caused by taking large doses of vitamin C. However, since the associations occurred at any dose of vitamin C, rather than at high doses specifically, a plausible explanation is that very health-conscious women who take supplements may also be prone to anxiety, which might cause IBS.

Others have suggested that using cross-sectional datasets like NHANES to draw conclusions about short-lived environmental chemicals and chronic complex diseases is inappropriate since a one-off snapshot of intakes cannot adequately characterise the relevant exposures. Furthermore, snapshots may be inadequate at capturing exposure detail from people with acute fatal diseases who have a short illness between diagnosis and death, for example pancreatic cancer.

2.4 Case-control studies

In case-control studies, people with a disease (cases) are compared to people without the disease (controls). Both groups have past exposure to the dietary factors of interest measured and they are compared to estimate the risk of disease associated with the risk factor. Case-control studies are quicker to conduct and cheaper than longer-term, larger-scale cohort studies; they are also useful for rare conditions. This study design potentially leads to greater statistical power as well as rapid and cost-effective management of the study. However, challenges arise with regard to the choice of appropriate controls and obtaining an unbiased measure of previous dietary exposure.

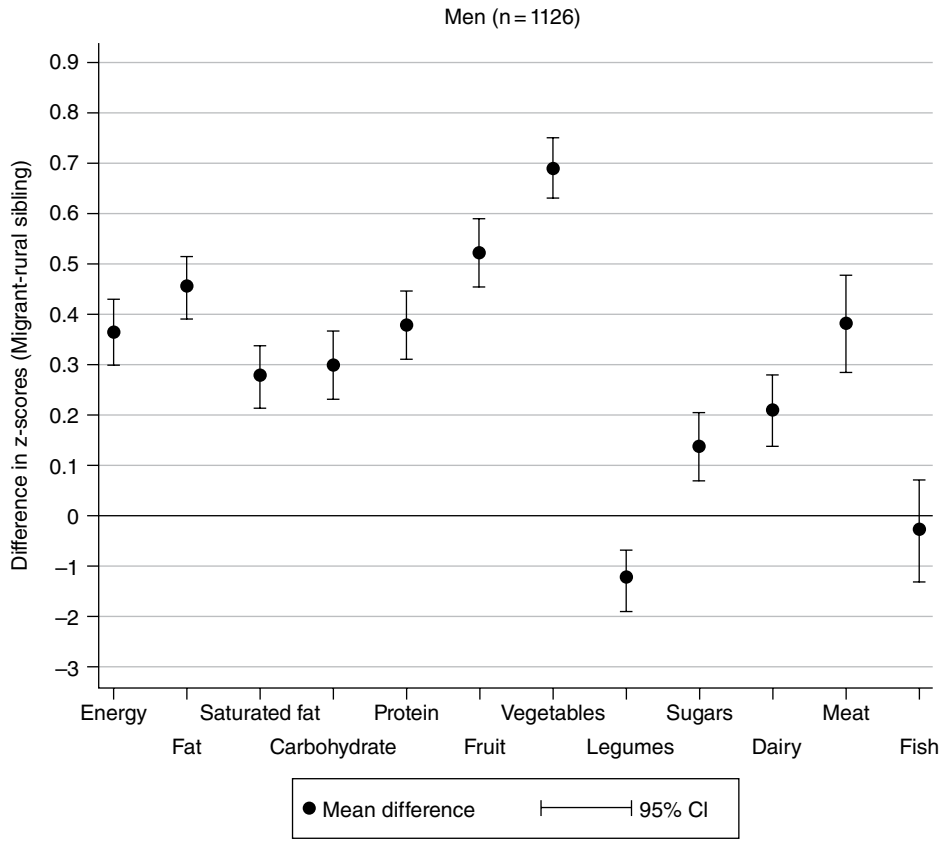

Figure 2.4 Differences in food intake z-scores between migrant and rural siblings in India. Bowen, L. *et al.* (2011) Dietary intake and rural-urban migration in India: a cross-sectional study, *PLoS One*, **6** (6), e14822.

Example case-control study

INTERHEART is one of the largest case-control studies that has been undertaken. It was designed to assess the importance of risk factors for coronary heart disease worldwide and 15 152 cases and 14 820 controls were enrolled from 52 countries, representing all inhabited continents. Specific objectives were to determine the strength of associations between various risk factors and acute myocardial infarction in the overall study population and to ascertain if this association varied by geographical region, ethnic origin, sex or age. Daily fruit and vegetable intake was found to reduce risk with an odds ratio of 0.70 (95% confidence interval [CI] 0.64 to 0.77). This means that people who ate fruit and vegetables every day had a 30% reduced risk of acute myocardial infarction compared to people who did not eat fruit and vegetables every day (Yusuf *et al.* 2004). Obesity doubled the risk, with an odds ratio of 2.24 (95% CI 2.06–2.45).

Methods

The research question to be studied needs to be formulated in order to ensure that the best population is chosen with an adequate supply of cases and suitable controls. In the study population there should also be a diversity of exposure to the dietary risk factors being studied. This is particularly important, since the error associated with dietary measurements tends to obscure potential associations with disease. If those in the study population are all similar with regard to dietary behaviour and the range of food or nutrient intakes is small from the lowest to highest values, it will be difficult to demonstrate an effect of the food or nutrient on the risk of disease due to measurement error. This may be larger than the real differences in intakes between cases and controls.

Selection of cases

The study population must be large enough with a high enough incidence of the disease of interest to provide

enough cases over the course of the study. Cases may be incident or newly diagnosed during the recruitment period. Prevalent, either existing or fatal, cases are sometimes used as an alternative to incident cases, since they may be more common and easier to find. However, associations made with prevalent or fatal cases need to be interpreted carefully, since the effect of the diet might be on survival rather than on the development of the disease. For example, if a case-control study found that vitamin D status was associated with an increased risk of pancreatic cancer mortality, this effect might not have occurred because vitamin D actually increases the risk of the disease. It could also be that lower vitamin D is associated with a higher cure rate or longer survival once a tumour is present.

Cases can be identified from hospital or general practice records or alternatively from case-finding in the general population. Using cases from hospitals or general practice would lead to missing undiagnosed people from the general population. Factors that determined earlier diagnosis might also be linked to differences in diet: for example, people who consult their doctor more frequently than others might also eat more healthily. If this is the case, then spurious associations could arise when studying only newly diagnosed patients.

The specificity of diagnosis is also important in the selection of cases. For example, it may be important to know the particular type of stomach cancer being linked to dietary behaviour, since different cancer types may have a different relationship. Intake of fruit and vegetables has been associated with an overall decreased risk of gastric cancer, but dietary intake seems to have a clearer effect on the intestinal type of stomach cancer compared to diffuse types.

Selection of controls

Selection of controls is one of the most difficult aspects of establishing a case-control study, as it is prone to bias. Controls should be selected from the same population, or one that represents the source population from which cases were drawn. Not only does this help to balance confounders between case and control groups, it is also important that selection of controls is independent of exposure status and is representative of the source population in terms of exposure. For instance, if cases are selected from screening clinics, then controls should also be selected from these clinics to avoid self-selection bias, since those attending clinics are likely to be more health conscious than the general population and may eat more healthily. They may also be in a specific age range or genetically more at risk than the general population, depending on the type of clinic. When cases are selected from hospitals, other than via screening clinics, then the selection is less straightforward. Patients with other diseases can be used as controls. Ideally, both cases and controls should be blind to the purpose of the study to avoid explanations for the disease under study being provided through their responses to the questionnaire. Additionally, selection of controls with a range of diseases may reduce the bias relating to exposure. Alternatively, controls can be selected from the general population, whose exposure may be more representative of the population at risk of becoming cases. Nevertheless, bias can occur because of differences between responders and non-responders.

More than one control from the study can be matched to each case to increase the power to detect associations. Cases and controls can be matched on variables such as age and sex, which are often related to disease and exposure. The use of siblings as controls can be useful, since shared genetic, socio-economic and environmental factors can be controlled for that otherwise may be difficult to measure or define. However, over-matching causing selection bias or reduced efficiency to detect associations should be avoided. This occurs when a factor is used to match cases and controls that is not a confounder of the exposure–disease association. For example, if a case-control study of fat intake in relation to type 2 diabetes matched cases and controls on body mass index (BMI), this could be considered over-matching, since BMI is on the causal pathway between fat intake and diabetes development. So by matching cases and controls on this factor, it will not be possible to assess the effect of fat intake on the risk of developing the condition. Variables used for matching cannot be studied in the analysis. Individual matching is expensive and time-consuming; alternatively group matching, also called frequency matching, can be used, which is a form of stratified sampling. For instance, the control group could be selected to have the same proportion of women as the case group, and the same distribution of ages stratified into age ranges.

Measurement of dietary exposure

A particular challenge for case-control studies is identifying the past dietary behaviour that will be relevant to the disease process. A disease may have a long pre-clinical phase and so the relevant exposure to diet may have occurred many years before diagnosis. People find it difficult to report past diet accurately and answers to questions on dietary behaviour in the past are strongly influenced by current eating patterns. If cases have changed their diets as a result of the disease process, then this will lead to error. Changes in diet are quite likely in diseases such as cancer or renal problems, which can affect appetite. Ideally, cases should be identified before they become symptomatic, thus reducing the risk of behaviour change as a result of the disease. This is only really possible using screening clinics to identify cases, such as from the breast-screening programme to identify

women with very early-stage breast cancer. Due to the potential for dietary behaviour change occurring in cases, the main method of collecting dietary information in case-control studies would be using food frequency questionnaires, which usually assess intake over the previous 12 months rather than current intake, which may have been affected by the disease.

Nested case-control studies

A nested case-control study can be developed from a cohort study; a subset of non-cases (controls) from the cohort are compared to the incident cases. Controls are selected for each case by matching on factors such as age. Usually, the exposure of interest (diet) is only measured among the cases and the selected controls. This design may be used when the exposure of interest is difficult or expensive to obtain, such as with coding food diaries or when the outcome is rare. By making use of data previously collected from a large cohort study, the time and cost of beginning a new case-control study are avoided. In addition, by only measuring the diet in as many participants as are necessary, the cost and effort of exposure assessment are reduced. Furthermore, since the dietary information was collected prior to disease incidence, the impact of recall bias on the exposure is reduced.

For example, a nested case-control study of dietary fibre intake and colorectal cancer risk was conducted using seven UK cohort studies, which included 579 case patients who developed colorectal cancer and 1996 matched control subjects. Dietary data obtained from four- to seven-day food diaries was used to calculate the odds ratios for colorectal, colon and rectal cancers with the use of conditional logistic regression models that adjusted for relevant covariates. The multivariable-adjusted odds ratio of colorectal cancer for the highest versus the lowest quintile of fibre intake density was 0.66 (95% CI 0.45–0.96), suggesting a protective effect (Dahm *et al.* 2010).

Analysis of case-control data

The main measure of association that is calculated from a case-control study is the odds ratio (OR). This is a measure of association between an exposure and an outcome. The OR evaluates whether the odds of a certain event or outcome are the same for two groups. Specifically, the OR measures the ratio of the odds that an event or result will occur to the odds of the event not happening. The OR represents the odds that an outcome (disease of interest) will occur given a particular exposure (dietary factor of interest), compared to the odds of the outcome occurring in the absence of that exposure. Typically the data consist of counts for each of a set of conditions and outcomes. By creating a 2×2 table (Table 2.1) the OR is a simple statistic to calculate: $[OR = (a \times d)/(b \times c)]$.

Table 2.1 Distribution of exposure in unmatched case-control studies.

	Cases	Control
Exposed to diet factor	a	b
Unexposed to diet factor	c	d

Matched studies use a different approach to calculate the OR, making use of the number of case-control pairs. If controls have been matched to cases, then a special type of logistic regression, conditional logistic regression, is used for the analysis. This means that controls are only compared to cases within the same matched set.

Confounding factors can be taken into account by using a logistic regression model. This will give an estimated OR and associated confidence intervals that are adjusted for the confounders included.

Odds ratios are also used in the analysis of nested case-control studies. Controls can be selected from the cohort to match cases depending on the date of their baseline intake measurement so that follow-up times are comparable; this deals with the varying recruitment dates within a whole cohort.

Problems with case-control studies

The two main areas of concern with case-control studies are dietary measurement error due to recall bias, and choice of controls. Both of these are discussed in the relevant sections earlier in this chapter.

The impact of recall bias on the results of case-control studies can be seen particularly in systematic reviews of the relationship between diet and disease where both case-control and cohort studies have been included. For example, Figure 2.5 shows that results from case-control studies exploring salt intake and risk of stomach cancer have a fivefold increased risk of stomach cancer per additional serving of salty foods per day; this is in comparison with data from cohort studies that show a much more modest and non-statistically significant increased risk. These differences may well be due, at least in part, to the impact of dietary recall bias in the case-control studies.

2.5 Prospective longitudinal studies

In a prospective longitudinal study (also known as a follow-up or cohort study), individuals are followed up over a period of time and disease or health outcomes are identified during the follow-up period. Individuals should be free of the disease being investigated at the start of the study (if not, they should be excluded from the analysis).

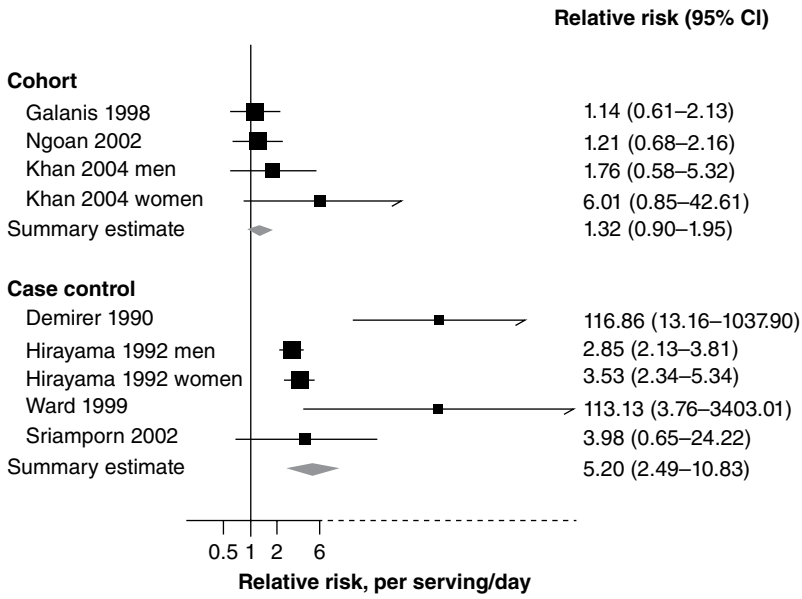

Figure 2.5 Salty/salted foods and stomach cancer risk. Results from a systematic review. This material has been reproduced from the 2007 WCRF/AICR Report *Food, Nutrition, Physical Activity and the Prevention of Cancer: a Global Perspective*.

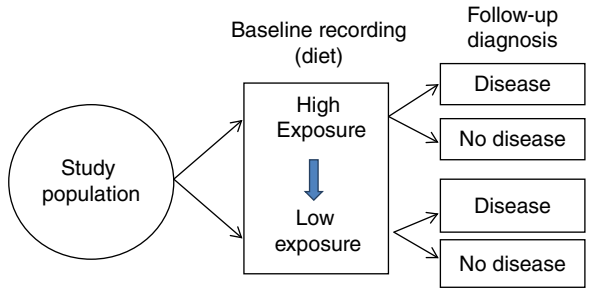

Figure 2.6 Flow chart of a cohort study.

Baseline data on nutritional and lifestyle exposures of interest that may be associated with the disease/health outcome are collected for all individuals (Figure 2.6).

In contrast to case-control studies, where systematic differences in reporting of exposures could occur between those with and those without the disease, recall bias associated with the outcome is avoided in prospective studies since exposure data is collected before the outcome data. In general, prospective studies are also less likely to be subject to the selection bias mentioned earlier in this chapter. Additional advantages are that cohort data can be used to study a wide range of disease and health outcomes, and by careful selection of individuals uncommon exposures or specific dietary patterns can be studied (e.g. vegetarian or Mediterranean diet). A well-designed cohort, for which a variety of exposures and confounders have been gathered, can also be used at a later date to test new hypotheses. A wider intake range of exposures can be gathered in cohorts compared to case-control or RCT studies, thus allowing useful dose–response relationships between exposure and outcome to be examined. Unlike cross-sectional studies, the time relationship between exposure and disease can be determined in longitudinal studies, therefore results can elucidate aetiology and may provide some evidence for causality if biases are minimised.

Examples

Using a cohort of 2635 pregnant women recruited between 8 and 12 weeks of pregnancy, caffeine intake during pregnancy was found to be associated with an

increased risk of fetal growth restriction (CARE Study Group 2008). Habitual caffeine intake from all potential sources from 4 weeks before pregnancy and during pregnancy was measured using a validated questionnaire. Details of potential confounders such as smoking, alcohol intake, maternal height and weight and ethnicity were also gathered using this questionnaire and adjusted for in the analysis. The association was found to be stronger in women with faster caffeine clearance compared to slower clearance; the caffeine half-life (the proxy for clearance) was determined by measuring caffeine in saliva (CARE Study Group 2008).

Much larger cohorts, such as the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort, have been established and followed up over many years to investigate associations between dietary intake and a range of cancers and other chronic diseases in the general population. EPIC includes over 500,000 people recruited in 10 European countries: Denmark, France, Germany, Greece, Italy, the Netherlands, Norway, Spain, Sweden and the UK. There are two UK studies in the EPIC group, EPIC-Norfolk and EPIC-Oxford, which recruited over 23 000 and 65 000 people respectively and have used food frequency questionnaires (FFQs) and food diaries to gather dietary intake data (http://epic.iarc.fr/). Blood samples from individuals have also been taken, from which concentrations of nutrients and hormones have been measured and used in analyses to investigate the relationship between diet and chronic diseases. In Epic-Norfolk, plasma ascorbic acid was inversely associated with cancer mortality in men but not women. Pooling the individuals from all or a number of EPIC studies increases the power to detect associations.

Methods

Selection of the study population

First, it is important to define the population from which the study sample (that is, the study population) is to be drawn and to which the results need to be generalised. This should also be undertaken for other types of observational studies. Ideally, a sampling frame needs to be compiled or sourced. This is a list of population members from which the study sample can be selected. If the exposure of interest is common, for instance fruit and vegetable intake, then the study sample can be selected from the general population using electoral registers, school registers or list of patients in general practices. EPIC-Norfolk, for instance, recruited people registered with 35 Norfolk GPs. In order for the findings of the sample to be generalisable to the population from which it is drawn, a large number of individuals need to be randomly selected from the sampling frame. Nevertheless, even for very large cohorts, random selection from subsets of the sampling frame stratified by age, gender, socio-economic status (SES) or other important factors may be necessary to ensure representativeness in these factors, as proposed for the UK Biobank (http://www.ukbiobank.ac.uk/). In this cohort, 500 000 people aged 40–69 from across the UK were recruited from National Health Service (NHS) registers held centrally.

Alternatively, individuals can be randomly selected from geographical clusters within the sampling frame (e.g. schools, GPs). Cluster sampling has advantages of cost and convenience in recruitment, but it is not truly random and similar characteristics of people living within one cluster may affect the results. Such effects may be reduced by increasing the number of clusters and selecting areas to include a range of known influencing characteristics, such as SES. Additionally, it may be necessary to weight the clusters in the analysis in order for the results to be representative of the proportions of individuals in the population rather than the proportion in the sample.

Incompleteness of the sampling frame also needs to be acknowledged, since hard-to-reach individuals such as the homeless or travellers are unlikely to be found on registers. Furthermore, the recruitment of pregnant women during their first trimester from antenatal clinics may miss women who attend their first clinic late in their pregnancy and who are often from lower SES.

If the exposure is not particularly common, for instance a vegetarian diet, then the study population can be selected based on their exposure to ensure that sufficient individuals with this exposure are included. The UK Women's Cohort Study (UKWCS) was established to compare vegetarians, fish eaters and red meat eaters. All eligible women who replied to an initial World Cancer Research Fund (WCRF) survey and stated that they were vegetarian or that they did not eat red meat were selected to take part in the cohort (Cade *et al.* 2004). However, only a proportion of the red meat–eating majority was invited to participate and a method of selection that was likely to avoid bias was used: for each vegetarian the next non-vegetarian in the list aged within 10 years was selected. Alternatively, vegetarians could be targeted more directly, as undertaken by the EPIC-Oxford study, which mailed members of the Vegetarian Society of the UK. Similarly, Seventh-Day Adventists, who usually follow a vegetarian diet, were recruited directly in studies established in California, USA, the Netherlands and other countries. Additionally, the UK Biobank may contain sufficient vegetarians to power analyses. Individuals participating completed an online 24-hour diet recall questionnaire that asked whether they routinely followed a vegetarian diet.

Power calculations utilising statistical packages are needed to determine the size of cohort required based on estimates such as the overall risk of the outcome in the

population, the ratio of unexposed to exposed individuals in the population and the follow-up time.

Smaller nested case-control studies can be established within large cohorts. Nevertheless, care is needed to avoid bias in the selection of the controls and cases, as outlined earlier.

Measurement of dietary exposures

Usually diet is measured only at baseline (the start of the cohort study). An assumption in longitudinal studies is that eating habits remain relatively stable before and after baseline data collection. However, this may not always be the case due to changes in dietary fashion and advice. To overcome this, some cohorts have undertaken additional wave collections of dietary exposures at a number of follow-up time points, although this adds to the study resource requirements, as well as leading to losses to follow-up and complexities of analysis. Although dietary intake between assessment points would be unknown, an average may be used for analysis, or respondents may be categorised for instance as 'always', 'sometimes' or 'never' within specified intake ranges at assessment. Similarly, the effect of intermittent supplement use (at least one but not all assessment points) and more consistent use (at all assessment points) could be compared in the analyses to never reporting use. Alternatively, questions relating to supplement use may be worded to obtain information on length of use as well as type, dose and numbers taken.

Although the disease outcome is collected prospectively in longitudinal studies, there is an element of retrospective recall of exposure data with the use of some instruments, such as FFQs and diet histories. FFQs usually obtain estimated average intake relating to the previous 12 months. More current and detailed, but short-term, dietary intake may be gathered by 24-hour recalls or by diary over a period of four to seven days. However, transferring information from paper-based diaries to electronic format requires substantially more time and resources in large-scale cohort studies. Resource requirements can be reduced by creating a much smaller nested case-control study within a large cohort or a number of cohort studies such as the UK Dietary Cohort Consortium, which was used to explore the relationship between breast cancer risk and vitamin C intake from both diet and supplements (Hutchinson *et al.* 2012).

Analysis of cohort data

Cohort studies allow us to measure disease incidence, since we have a healthy population who are followed up over time and the rate of new disease development (incidence) can be calculated. If the follow-up times for all the individuals in the cohort are similar, then relative risk ratios can be estimated. For instance, in birth cohorts the relative risk of an outcome in the offspring at a specified age in relation to specified intake during pregnancy can be calculated. The relative risk is the cumulative incidence in the exposed group compared to the cumulative incidence in the unexposed group. However, since it is important to adjust for potential confounders in all risk analyses of cohort data, multiple regression analysis is most often undertaken. If the outcome is continuous, then multiple linear regression can be used, but if the outcome is dichotomous, then logistic regression should be carried out.

In the CARE study, multiple linear regression analysis was used to estimate the reduction in birth weight with higher caffeine intake after adjustment for various factors. However, logistic regression was used to estimate the odds of giving birth to a baby with fetal growth restriction (birth weight < 10th centile after accounting for maternal factors) depending on caffeine intake during pregnancy. In nutritional epidemiology, intake is usually split into quartiles or quintiles for reporting estimates with confidence intervals; in this study, intake of < 100 mg/day was compared to intake groups of 100–199, 200–299 and ≥ 300 mg/day. Dose–response relationships, which can provide some evidence of causality, were found in testing for trends using intake as a continuous exposure and, as commonly done, p values for these were reported. Additionally, the risk of fetal growth restriction was plotted against increasing caffeine intake using fractional polynominal regression, a more advanced statistical technique. This showed a linear dose–response relationship with no threshold effects.

If follow-up times differ substantially between individuals in the cohort, then the total person-time at risk is needed to calculate hazard ratios (rate ratios) in time-to-event analysis (also called survival analysis, even when the event is a disease incidence and not a death). This method is useful when recruitment takes place over a number of years. Person-years at risk is calculated for each individual as the time from the measurement of dietary intake at their baseline date until disease incidence or the censor date or the individual was lost to follow-up. Cox regression (also known as proportional hazards regression) is one method of time-to-event analysis. Individuals known to have the outcome at the set-up of the cohort should be excluded from all risk analyses. This is one of the main analysis methods used in the EPIC and UKWCS cohorts mentioned earlier.

One of the biggest issues with the reliability of the results in cohort studies is confounding, which is explained in Section 2.6. The selection of confounders used for adjustment in an analysis of cohorts and other observational studies may appear to be more of an art than a science, since it often requires subjective decision-making. Univariate analyses should be undertaken to

determine associations between the potential confounders and the outcome and then between the potential confounders and the dietary exposure. Variables that are significantly associated with both should be considered for adjustment. Variables that do not meet this criterion in the study but where there is strong prior evidence of confounding from previous studies should also be considered for inclusion for adjustment. Visual methods, for instance creating diagrams called directed acyclic graphs (DAGs), can help clarify the direction of the effects of variables to help identify which may or may not be potential confounders (Greenland, Pearl and Robins 1999). Variables that appear to be on the causal pathway between the dietary exposure and the outcome should not be included, since controlling for these mediators would attenuate associations between exposure and outcome. Finally, over-adjustment can occur if too many confounders are included, particularly if they are collinear; that is, they are strongly correlated. This can lead to associations being missed where they may really exist.

Another way of controlling for a confounder in cohort studies, other than through multiple regression, is by restricting the analysis to those individuals who are not affected by the confounder. For instance, the majority of studies exploring associations between nutrient intake and disease have only measured intake from diet, and have not included the nutrient intake from supplements, which are commonly consumed in the Western world. Furthermore, some studies that have gathered supplement data have not obtained the strength dosage of supplements. To avoid supplement intake being a confounder, all supplement users could be excluded from an analysis, if this basic information is provided. These results could then be compared to results prior to exclusions.

Problems with longitudinal studies

Longitudinal studies are very time-consuming and expensive. Since only a small percentage of individuals may develop the outcome by the end of the follow-up period, very large sample sizes as well as long follow-up periods are needed to detect a significant result if an association exists. Sample size calculations are recommended prior to the creation of the cohort and also prior to analyses. Additionally, as seen in later chapters, the methods for collecting and electronically capturing exposures are very time-consuming for the individual and the research team. Nevertheless, new technologies such as the 'My Meal Mate' (MMM) mobile smartphone application and the online 24-hour recalls 'ASA24' and 'myfood24' are being developed to improve accuracy and reduce data-capture resource requirements. Since participants may find these methods less burdensome, they may be willing to provide extra days of consumption to classify their intake more appropriately compared to traditional methods.

In particular, cohorts are an inefficient or impractical method of studying relatively rare outcomes such as pancreatic cancer, since a large number of individuals would be needed to find a significant association, if one existed. Nevertheless, this may be partially overcome in cohort studies by undertaking a meta-analysis of results from a number of studies, or, better still, by pooling individual data for analysis from a number of cohort studies.

Due to long follow-up periods, substantial numbers of individuals may drop out during this time (attrition). If the outcomes for these individuals cannot be determined and these losses to follow-up are related to both the exposure and the outcome, then this differential loss to follow-up produces a form of selection bias. Multiple methods of contact and surveillance may reduce losses to follow-up. Since health workers and civil servants are usually more easily traced than the general population, cohorts have been established using these study populations, for instance the Whitehall Study (Marmot and Brunner 2005) and the Nurse's Health Study (Zhang *et al.* 1999). In countries such as the UK and USA, losses to follow-up for some outcomes can now be mainly overcome, and therefore the quality of cohort studies increased, by obtaining outcome data from national and regional registries, for example for cancer and heart disease outcomes, rather than by direct contact with the individual. Consent to obtain this information must be obtained from individuals at the start of the study. Although selection bias by researchers is less of a problem in cohorts than in case-control studies, self-selection – that is, the type of person who volunteers for participation in cohorts – can create selection bias.

Despite efforts to control for confounders in analyses, residual confounding may remain and may account for some of the significant though unknowingly spurious results that are published. Nevertheless, on the whole, cohort results usually provide better support for aetiological suggestions than other observational studies, since they have fewer methodological limitations.

2.6 Confounding

In observational studies, confounders must be taken into account because they can influence the estimated size, direction and/or significance of association between the dietary factor of interest and the disease outcome. As mentioned in the introduction to this chapter, potential confounders are variables that may be associated with both the exposure of interest (diet) and the outcome variable (disease), but are not on the causal pathway

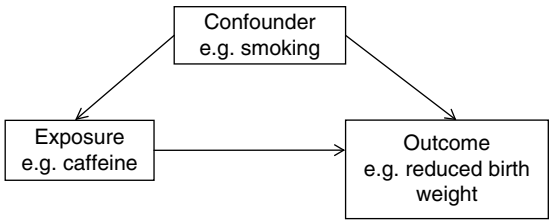

Figure 2.7 An example of confounding.

between exposure and outcome. A confounder can provide an alternative explanation for an apparent association between a dietary exposure and a disease/health outcome. For instance, a positive association between increased coffee consumption and reduced birth weight could be explained by higher smoking levels, because it is well established that smoking during pregnancy is associated with reduced birth weight, and smoking has also been associated with increased coffee consumption (this relationship is illustrated in Figure 2.7). In the Caffeine and Reproductive Health (CARE) study, smoking status was adjusted for as a confounder in multiple regression analyses to produce results showing associations between maternal coffee drinking and fetal birth weight independent of smoking status (CARE Study Group 2008).

Confounding variables need to be identified, measured and controlled for in analyses. Depending on the study design, confounders can be dealt with during the design of the study by matching or by restricting study members; and through data analysis by stratification (e.g. age standardisation), by restriction or by adjustment in regression models.

Even after controlling for measured confounders, some residual confounding may remain. However, in general, confounding variables can often be measured more accurately than dietary variables, and therefore any significant associations with health outcomes found in unadjusted analyses often disappear or are greatly attenuated after adjustment for confounders. This may be due to genuine confounding; alternatively, true dietary associations may be masked by confounders because these dietary exposures are measured less accurately.

Although confounding is less likely in RCTs because covariates are randomly distributed between interventions, nutritional interventions for RCTs may not be feasible or may present their own problems. Individuals may not be willing to be randomised to a diet for the number of years that would be necessary to cause substantial alterations in disease risk. Furthermore, unlike drug interventions where the accessibility of drugs is restricted and amounts provided are known, nutritional interventions can never be completely controlled. For instance, vitamin supplements as an active intervention can be easily obtained by study individuals, meaning that supplement intake can also happen in the control group; furthermore, additional supplement intake can easily occur in the intervention group. Due to the limitations of RCTs in nutrition research, results of observational studies remain useful for exploring the role of nutritional exposures in the causation of disease, despite confounding and other weaknesses.

2.7 Conclusion

Population-based observational studies use epidemiological techniques to explore associations between diet and health outcomes. Ecological and cross-sectional studies are useful first steps in generating hypotheses; however, they cannot be used to test aetiological theories due to methodological limitations. Case-control studies are useful for studying diet and disease relationships, but they present problems in the choice of appropriate controls and recall of past diet. Cohort studies allow more rigorous testing of diet–disease relationships than other approaches. All of these methods are potentially influenced by confounding factors, variables that may be associated with both the exposure of interest (diet) and the outcome (disease).

References and further reading

Bowen, L., Ebrahim, S., De Stavola, B. *et al.* (2011) Dietary intake and rural-urban migration in India: a cross-sectional study. *PLoS One*, **6** (6), e14822.

Cade, J.E., Burley, V.J., Greenwood, D.C.; The UKWCS Steering Group (2004) The UK Women's Cohort Study: Comparison of vegetarians, fish-eaters and meat-eaters. *Public Health Nutrition*, **7** (7), 871–878.

CARE Study Group (2008) Maternal caffeine intake during pregnancy and risk of fetal growth restriction: A large prospective observational study. *British Medical Journal*, **337**, a2332.

Corsi, D.J., Neuman, M., Finlay, J.E. and Subramanian, S. (2012) Demographic and health surveys: A profile. *International Journal of Epidemiology*, **41** (6), 1602–1613.

Dahm, C.C., Keogh, R.H., Spencer, E.A. *et al.* (2010) Dietary fiber and colorectal cancer risk: A nested case-control study using food diaries. *Journal of the National Cancer Institute*, **102** (9), 614–626.

Duffey, K.J. and Popkin, B.M. (2013) Causes of increased energy intake among children in the U.S., 1977–2010. *American Journal of Preventive Medicine*, **44** (2), e1–8.

Foley, W., Ward, P., Carter, P. *et al.* (2010) An ecological analysis of factors associated with food insecurity in South Australia, 2002–7. *Public Health Nutrition*, **13** (2), 215–221.

Greenland, S., Pearl, J. and Robins, J.M. (1999) Causal diagrams for epidemiologic research. *Epidemiology*, **10** (1), 37–48.

Hutchinson, J., Burley, V.J., Greenwood, D.C. *et al.* (2011) High-dose vitamin C supplement use is associated with self-reported histories of breast cancer and other illnesses in the UK Women's Cohort Study. *Public Health Nutrition*, **14** (5), 768–777.

Hutchinson, J., Lentjes, M., Greenwood, D. *et al.* (2012) Vitamin C intake from diary recordings and risk of breast cancer in the UK Dietary Cohort Consortium. *European Journal of Clinical Nutrition*, **66** (5), 561–568.

Keys, A., Menotti, A., Karvonen, M.J. *et al.* (1986) The diet and 15-year death rate in the Seven Countries Study. *American Journal of Epidemiology*, **124** (6), 903–915.

Lim, S.S., Vos, T., Flaxman, A.D. *et al.* (2012) A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, **380** (9859): 2224–2260.

Marmot, M. and Brunner, E. (2005) Cohort profile: The Whitehall II Study. *International Journal of Epidemiology*, **34** (2), 251–256.

Ng, T.P., Aung, K.C., Feng, L. *et al.* (2012) Homocysteine, folate, vitamin B-12, and physical function in older adults: Cross-sectional findings from the Singapore Longitudinal Ageing Study. *American Journal of Clinical Nutrition*, **96** (6), 1362–1368.

Pickett, K.E., Kelly, S., Brunner, E. *et al.* (2005) Wider income gaps, wider waistbands? An ecological study of obesity and income inequality. *Journal of Epidemiology and Community Health*, **59** (8), 670–674.

Pomerleau, J., Lock, K. and McKee, M. (2003) Discrepancies between ecological and individual data on fruit and vegetable consumption in fifteen countries. *British Journal of Nutrition*, **89** (6), 827–834.

Stephen, A.M., Mak, T.N., Fitt, E. *et al.* (2013) Innovations in national nutrition surveys. *Proceedings of the Nutrition Society*, **72** (1), 77–88.

Yusuf, S., Hawken, S., Ounpuu, S. *et al.* (2004) Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study. *Lancet*, **364** (9438), 937–952.

Zhang, S., Hunter, D.J., Forman, M.R. *et al.* (1999) Dietary carotenoids and vitamins A, C, and E and risk of breast cancer. *Journal of the National Cancer Institute*, **91** (6), 547–556.