Case-Control Studies

I.6

Norman E. Breslow

6.1 Introduction

6.1.1 A Brief History

The case-control study examines the association between disease and potential risk factors by taking separate samples of diseased cases and of controls at risk of developing disease. Information may be collected for both cases and controls on genetic, social, behavioral, environmental or other determinants of disease risk. The basic study design has a long history, extending back at least to Guy's 1843 comparison of the occupations of men with pulmonary consumption to the occupations of men having other diseases (Lilienfeld and Lilienfeld 1979). Beginning in the 1920's, it was used to link cancer to environmental and hormonal exposures. Broders (1920) discovered an association between pipe smoking and lip cancer; Lane-Claypon (1926), who selected matched hospital controls, investigated the relationship between reproductive experience and female breast cancer; and Lombard and Doering (1928) related pipe smoking to oral cancer. The publication in 1950 of three reports on the association between cigarette smoking and lung cancer generated enormous interest in case-control methodology as well as bitter criticism (Levin et al. 1950; Wynder and Graham 1950; Doll and Hill 1950). The landmark study of Doll and Hill (1950, 1952), in particular, inspired future generations of epidemiologists to use this methodology. It remains to this day a model for the design and conduct of case-control studies, with excellent suggestions on how to reduce or eliminate selection, interview and recall bias.

From the mid-1950's to the mid-1970's the number of case-control studies published in selected medical journals increased four- to sevenfold (Cole 1979). Aird et al. (1953) discovered the association between gastric cancer and the ABO blood groups. The impact of hormonal factors on cancers of female organs was brought to light, starting with confirmation of the association between late first pregnancy and breast cancer (MacMahon et al. 1970). Herbst et al. (1971) investigated an unusual outbreak of vaginal adenocarcinoma in young women, finding that mothers of seven of eight cases had exposed their daughters *in utero* to the fertility drug diethylstilbestrol (DES). None of 32 control mothers had a history of estrogen use during pregnancy. Treatment of menopausal women with exogenous estrogens similarly increased the risk of endometrial cancer (Ziel and Finkle 1975; Smith et al. 1975). Powerful joint effects of alcohol and tobacco consumption on esophageal cancer were demonstrated (Tuyns et al. 1977), as was the strong association between liver cancer and hepatitis B carrier status (Prince et al. 1975). These successes encouraged more investigators to adopt the case-control study as the method of choice for the study of rare chronic diseases, particularly cancer. A survey by Correa et al. (1994) identified 223 population-based case-control studies published in the world literature in 1992. Recent discoveries obtained using case-control methodology have included the role of salted fish in the etiology of nasopharyngeal carcinoma in Chinese populations (Armstrong et al. 1983; Yu et al. 1986), the hazards of prone sleeping position for sudden infant death syn-

drome (SIDS) (Fleming et al. 1990) and the relationship between use of intrauterine devices (IUDs) and tubal infertility (Daling et al. 1985).

This plethora of case-control studies, stimulated by their relatively low cost and short duration, also had its drawbacks. Not all investigators were as careful as Doll and Hill in following a protocol for selection of cases and controls, in conducting the study to mitigate against bias and in thoughtfully analysing the collected data. Nor did they have the good fortune to study associations as strong as that between lung cancer and cigarette smoking. The increasing availability of high speed computers made it possible to collect more and more data, and to look for all manner of associations with putative risk factors. Investigators eager for research funding were sometimes too quick to publish their findings and draw media attention to them. The inevitable result was an increasingly negative reaction on the part of the public, and from segments of the scientific community, to the false alarms and contradictory results (Taubes 1995). One goal of this chapter, and of others in this handbook, is to describe basic scientific principles whose application should help to improve public confidence in published findings of epidemiologic studies.

Early Methodologic Developments 6.1.2

> The sophisticated use and understanding of case-control studies is the most outstanding methodologic development of modern epidemiology.
> (Rothman 1986, p. 62)

The initial interpretation of the case-control study was the comparison of exposure histories for a group of diseased cases with those for non-diseased controls. Typical analyses involved two group comparisons of exposure distributions using chi-squared and t-tests. The critics argued that such comparisons provided no information about the quantities of true epidemiologic interest, namely the disease rates. Cornfield (1951) corrected this misconception by demonstrating that the exposure odds ratio for cases vs. controls was equal to the disease odds ratio for exposed vs. non-exposed. With $D = 1$ indicating disease, $D = 0$ disease-free and $X = 1/0$ likewise denoting exposed or non-exposed, he showed using Bayes theorem that

$$\frac{\Pr(D=1|X=1)\Pr(D=0|X=0)}{\Pr(D=0|X=1)\Pr(D=1|X=0)} = \frac{\Pr(X=1|D=1)\Pr(X=0|D=0)}{\Pr(X=0|D=1)\Pr(X=1|D=0)} \qquad (6.1)$$

and noted that the disease odds ratio approximated the *relative risk* $\Pr(D = 1|X = 1)/\Pr(D = 1|X = 0)$ provided the disease was rare. He also pointed out that, if the overall disease risk was known from other data sources, this could be combined with the relative risk to estimate *absolute* disease risks for exposed and non-exposed, respectively.

Disease risk as considered by Cornfield (1951) was *prevalence*, the probability that a member of the population was ill at a given point in time. For studies

of disease etiology, however, it is preferable to work with disease *incidence*, the probability of developing disease during the study period among subjects who are free of disease initially. Otherwise, one confuses the effect of exposure on causation of disease with its effect on the case fatality rate (Neyman 1955). Controls for a study of the cumulative risk of developing disease during a given period would be persons who were free of disease during the entire period. Although it laid the foundation for what was to follow, this conceptualization of the case-control study in terms of cumulative disease risk was awkward, for two reasons. First, as the study interval lengthened the risk of disease increased for both exposed and non-exposed. The relative risk for a common disease could approach one. Even if it did not, it was undesirable to have the basic effect measure so dependent on study duration, which varies between studies. Second, for a study of long duration, ensuring that the controls were disease-free throughout the study period could be problematic in practice. The modern conception of a case-control study involves sampling of controls who are disease-free at random times during the study period (Sect. 6.2.1). Exposure odds ratios are used to estimate ratios of incidence *rates* rather than ratios of risks. No rare disease assumption is needed in this case.

Mantel and Haenszel (1959) clarified the status of the case-control (or retrospective) study in comparison with the cohort (forward or prospective) study in one of the most highly cited papers in the scientific literature (Breslow 1996). They stated emphatically:

> A primary goal is to reach the same conclusions in a retrospective study as would have been obtained from a forward study, if one had been done.
> (Mantel and Haenszel 1959, p. 722)

This insight underlies the modern conception of the case-control study as involving *sampling*, on the basis of outcome, from an ongoing real or imagined cohort study that has been designed to provide the best possible answer to the basic question. Mantel and Haenszel introduced a new test and a simple, highly efficient estimator for the relative risk after stratification on control factors. Their methods required the epidemiologist to carefully examine the tabular data, and thus to identify strata where there was a lack of information or where there were discrepancies between summary and stratum specific relative risks. They remain valuable today as an adjunct to more elaborate model fitting.

By the end of the 1950s, the case-control study was firmly established as the method of choice for the chronic disease epidemiologist, certainly when the budget was limited. The role of statisticians in bringing the study design to this place of scientific respectability was widely acknowledged (Cole 1979; Armenian and Lilienfeld 1994). Further methodological advances were made during the next two decades, particularly in statistical modeling of case-control data. The development of the proportional hazards regression model for life table data (Sheehe 1962; Cox 1972) provided a sound mathematical basis for methods long used by epi-

demiologists, and led to refinements and extensions of those methods (Breslow et al. 1983). The nested case-control study, originally conceived as a method to reduce the computational burden of fitting Cox's model to data from large cohorts (Liddell et al. 1977), was recognized as an efficient epidemiologic design for the collection of expensive explanatory data (Langholz and Goldstein 1996). It now serves as a paradigm for all case-control studies. Many of the methodological developments were described in texts by Breslow and Day (1980) and Schlesselman (1982) that led to further appreciation and use of the case-control study.

Chapter Outline 6.1.3

The remainder of this chapter discusses the modern conceptualization of the case-control study, largely from a statistical perspective. Matching of controls to cases at the design stage is viewed as a technique to be used in carefully limited contexts to increase the statistical efficiency of a highly stratified analysis. The implications of these theoretical developments for the practical selection of cases and controls are explored. Major pitfalls include the unique susceptibility of the case-control study to selection bias and, especially when exposures are assessed by interview, to measurement error. The design of any particular study usually involves tradeoffs between potential biases arising from these sources. Following established principles of sound statistical science, including the use of an appropriate protocol for subject selection and exposure assessment, can help reduce the variability in study results that has contributed to the low esteem accorded risk factor epidemiology in some scientific circles (Breslow 2003).

Conceptual Foundations 6.2

Sampling from a Real or Fictitious Cohort 6.2.1

The Mantel and Haenszel (1959) goal, of reaching the same conclusions from a case-control study as from a cohort study if one had been done, provides the key to understanding of case-control methodology. Rather than start the planning process by thinking about how to conduct a case-control study, it often is helpful to first plan the ideal cohort study that would be conducted to investigate the same hypothesis if unlimited resources were available. Planning would include cohort identification, definition of the times of entry into and exit from the cohort, ascertainment of the disease endpoint, measurement of the exposure histories, consideration of potential confounders and methods of statistical analysis. The corresponding case-control study would then be viewed as the random sampling of subjects from this idealized cohort to achieve, so far as possible, the stated goal.

Cohort Definition. In concept the underlying cohort for a case-control study consists of all subjects who, had they experienced the disease endpoint at a specific time, would have been ascertained as a case at that time. When case-control sampling is carried out in the context of an actual cohort study, to select individuals for genotyping or other expensive measurements, for example, the cohort is completely enumerated by and known to the investigator. More typically, however, the underlying cohort is not fully identified and is effectively defined by the method of case ascertainment. When cases are ascertained from a particular hospital, for example, one considers the cohort to consist of all subjects who, had they developed the disease in question, would have been diagnosed in that hospital.

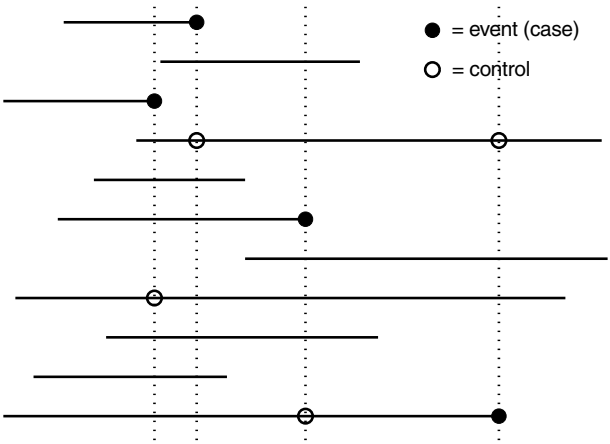

Figure 6.1. Schematic of a (nested) case-control study

Figure 6.1 illustrates the basic idea of case-control sampling. Each of the 11 horizontal lines represents time on study for a member of the cohort. Subjects enter follow-up at the left hand endpoint and exit at the right. They are considered to be *at risk* of becoming a case throughout this period. It is even possible, though not shown here, that a subject could enter the cohort, leave for awhile and then return. Four of the 11 subjects are cases. Their follow-up ends at diagnosis since they are no longer at risk of becoming an incident case thereafter. The vertical dotted lines, plotted at each of the times that a case occurs, intersect the trajectories of those who are at risk at that time, i.e., the trajectories of subjects in the corresponding *risk set*.

Nested Case-Control Sampling. When the cohort study is a real one, so that times of entry and exit are known for all members, the investigator may completely enumerate each risk set. A nested case-control study is then possible in which controls are selected by finite population random sampling, without replacement, from non-cases in the risk set. The usual assumption is that the sampling of controls from each risk set is completely independent of sampling from all other risk sets. Two consequences are that a subject sampled as a control at one point in time

may later become a case, and that the same subject may be sampled more than once as a control. Figure 6.1, which depicts the situation where exactly one control is sampled from each risk set, illustrates each of these possibilities. Robins et al. (1986) describe other sampling schemes, and corresponding methods of analysis, for nested case-control studies.

Density Sampling. These ideas also may be applied, at least in principle, to the more typical situation in which the cohort is not completely enumerated. An essential assumption, which in fact well approximates the design of many studies, is that the cohort is sampled throughout the study period. More specifically, controls are selected at any given time at a rate proportional to the disease incidence rate at that time (Sheehe 1962). Miettinen (1976) termed this *incidence density sampling*. A second assumption is that each subject at risk at a given time has the same probability of being sampled as a control. This implies that, from the standpoint of an individual, the likelihood of being included in the study as a control increases with increasing *time on study*. If the disease incidence rate is constant, someone who is a member of the cohort for twice as long as someone else has twice the chance of being selected as a control. In the statistical literature this is known as *length biased* sampling. One important consequence, under the assumption of constant disease incidence, is that the number of controls sampled is proportional to the *total time at risk*.

Incidence Rate Ratios are Estimable from Odds Ratios 6.2.2

We consider here the simplest situation in which the disease incidence rate is constant and there are two groups of subjects, exposed and non-exposed, that are homogeneous apart from exposure. Confounding is therefore not an issue. Denote by A the total number of incident cases ascertained from the cohort during the study period (t_0, t_1) and suppose that A_0 are determined to be non-exposed whereas $A_1 = A - A_0$ are exposed. Similarly denote by $T = T_0 + T_1$ the total person-time on study, decomposed into its non-exposed (T_0) and exposed (T_1) components. While the numbers of cases A_0 and A_1 are known to the investigator, T_0 and T_1 may not be unless the underlying cohort is a real one. Instead, the case-control study provides information on how many of the total $M = M_0 + M_1$ of controls are non-exposed (M_0) and how many are exposed (M_1). Denoting by M_1/M_0 the observed odds of exposure for controls and likewise by A_1/A_0 the observed odds of exposure for cases, the corresponding exposure odds ratio is $(A_1M_0)/(A_0M_1)$.

Let $\pi\tau$ denote the probability that a subject who contributes τ person-years of followup is sampled as a control. With $T = \sum_{i=1}^{N} \tau_i$ denoting the sum of the times-on-study for N cohort members, i.e., the total time at risk, the expected number of controls is $E(M) = \pi T$. In practice π is often selected by the investigator to yield a fixed number of controls, at least as a target value. Its actual value remains unknown unless information is available about T. Nonetheless, provided π is constant for all subjects, both exposed and non-exposed, $E(M_0) = \pi T_0$ and

$E(M_1) = \pi T_1$. Hence the control ratio M_0/M_1 estimates the corresponding ratio T_0/T_1 of person-time. Since the exposure specific incidence rates are estimated by $\hat{\lambda}_0 = A_0/T_0$ and $\hat{\lambda}_1 = A_1/T_1$, it follows (see Rothman and Greenland 1998, Chap. 10) that the rate ratio may be estimated by the exposure odds ratio:

$$\frac{\hat{\lambda}_1}{\hat{\lambda}_0} = \frac{A_1 T_0}{A_0 T_1} \approx \frac{A_1 M_0}{A_0 M_1} . \tag{6.2}$$

See Sect. 6.3.1 for a numerical example.

6.2.3 Time-dependent Rates and Exposures

Section 6.2.2 assumes that the parameter of interest is the ratio of instantaneous incidence rates, each assumed constant in time, for exposed and non-exposed subjects. A more general conceptualization takes the interest parameter to be the ratio $\psi \equiv \lambda_1(t)/\lambda_0(t)$ of instantaneous rates where the ratio, but not necessarily the underlying rates, is assumed constant in t. Let $N(t)$ denote the total number of subjects at risk at time t in the underlying cohort, of which a proportion $p_1(t)$ are exposed and $p_0(t)$ are non-exposed. These proportions could vary with time either because the exposure status for individual subjects changes, or because the exposure composition of the cohort changes through entries and exits. Note that the expected number of exposed cases is given by $\int N(t)p_1(t)\lambda_1(t)\,\mathrm{d}t$ and similarly for the non-exposed cases. The expected number of controls sampled in the interval $(t, t+\mathrm{d}t)$ is therefore $M(t)\,\mathrm{d}t$ where $M(t) = N(t)[p_0(t)\lambda_0(t) + p_1(t)\lambda_1(t)]$. It follows that the unadjusted exposure odds ratio under density sampling estimates

$$\begin{aligned} \psi^* &= \frac{\int N(t)p_1(t)\lambda_1(t)\,\mathrm{d}t \int M(t)p_0(t)\,\mathrm{d}t}{\int N(t)p_0(t)\lambda_0(t)\,\mathrm{d}t \int M(t)p_1(t)\,\mathrm{d}t} \\ &= \psi \frac{\int N(t)p_1(t)\lambda_0(t)\,\mathrm{d}t \int M(t)p_0(t)\,\mathrm{d}t}{\int N(t)p_0(t)\lambda_0(t)\,\mathrm{d}t \int M(t)p_1(t)\,\mathrm{d}t} \end{aligned} \tag{6.3}$$

(Greenland and Thomas 1982). Thus the exposure odds ratio estimates the incidence rate ratio, i.e., $\psi^* = \psi$, provided either that the exposure proportions are constant in t or else that $\psi = 1$. Otherwise, time t acts as a *confounder* of the exposure–disease association. In this case, a time-matched analysis using standard methods for matched case-control studies (Breslow and Day 1980, Chap. 7) is needed to estimate ψ unbiasedly. The marginal (unmatched) odds ratio usually provides a slightly *conservative* estimate of this parameter.

6.2.4 Cumulative Risk Ratios and Case-Cohort Sampling

While it is generally agreed that case-control studies of chronic disease are best designed using density sampling to estimate the incidence rate ratio, alternative sampling designs may be superior for other purposes. Vaccine efficacy is usually defined as the proportional reduction, over the study period, in the num-

ber of cases among subjects who are vaccinated compared to those who are not. Equivalently, it is 1 minus the ratio of cumulative disease risks for vaccinated vs. non-vaccinated. Suppose the effect of vaccination is to render completely immune a proportion P_I of subjects, while the remainder of those vaccinated have the same disease incidence rates $\lambda_0(t)$ as do non-vaccinated persons (Smith et al. 1984). For simplicity assume that all subjects, both vaccinated and non-vaccinated, are followed from a common starting time t_0 and that there is no loss to follow-up. The cumulative risk of disease by time t_1 for those not vaccinated is $P(t_0, t_1) = 1 - \exp[-\int_{t_0}^{t_1} \lambda_0(t)\,dt]$ and the vaccine efficacy is thus

$$1 - \frac{\text{risk for vaccinated}}{\text{risk for non-vaccinated}} = 1 - \frac{P_I \times 0 + (1 - P_I) \times P(t_0, t_1)}{P(t_0, t_1)} = P_I \,. \qquad (6.4)$$

Here the cumulative risk ratio, not the incidence rate ratio, is independent of study duration $t_1 - t_0$ (Rodrigues and Kirkwood 1990). Suppose now a *subcohort* of M subjects is drawn at random from the combined cohort of vaccinated and non-vaccinated subjects such that each individual has the *same* probability π of inclusion in it, *regardless* of duration of follow-up. If M_0 and M_1 denote the numbers of non-vaccinated and vaccinated in the subcohort, while A_0 and A_1 denote the numbers of disease cases diagnosed by time t_1, then vaccine efficacy is simply estimated as

$$\widehat{P}_I = 1 - \frac{A_1/M_1}{A_0/M_0} \,. \qquad (6.5)$$

More generally, the case-cohort design (Kupper et al. 1975; Miettinen 1982; Prentice 1986) involves random sampling of a subcohort at study entry, without re-

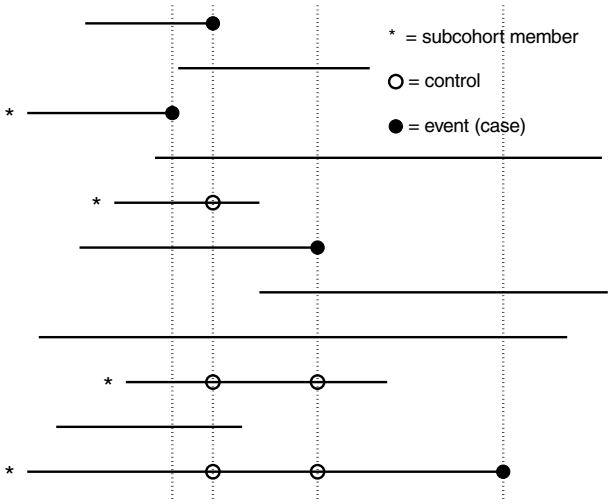

Figure 6.2. Schematic of a case-cohort study

gard to time on study. Figure 6.2 contrasts this design with nested case-control sampling (Fig. 6.1). Incidence rate ratios may be estimated for dynamic (open) cohorts, with staggered entry and loss to follow-up as pictured, just as they are with nested case-control sampling (Prentice 1986; Lin and Ying 1993; Barlow 1994). Subcohort members under observation at the time of disease occurrence serve as the controls for each case in a time-matched analysis. Since the subcohort is a simple random sample from the full cohort, it is suitable for estimation of population genotype or exposure frequencies, whereas the controls from a nested study are not. Furthermore, the same subcohort may be used to provide controls for two or more different types of disease cases. Because of this flexibility, the case-cohort design is increasingly used for sampling from defined cohorts.

6.2.5 Estimation of Absolute Risks

The key feature of case-control sampling in the context of an actual cohort study, where the underlying cohort is completely enumerated and entry and exit times are known for all cohort members, is that the sampling probabilities for cases and controls are known or can be estimated from the available data. The case-control study provides supplementary information on explanatory variables for a randomly selected group of cohort members. Analysis of the combined cohort and case-control data may be approached using standard methods for incomplete data (Little and Rubin 2002). The Horvitz and Thompson (1952) survey sampling approach is often easiest to implement. Here the contribution to estimators or estimating equations from each subject with complete data, i.e., each subject included in the case-control sample, is weighted by (an estimate of) the inverse probability of having been included. Any analysis that could have been carried out were explanatory data available for the entire cohort can also be carried out using the combined data from the cohort and the case-control sample. This principle applies to estimation of absolute as well as relative risks.

Table 6.1. Numbers of lung cancer cases and controls in Greater London among males aged 45–64 years, by average amount smoked in preceding 10 years, with estimated death rates of lung cancer per 1000 persons per year[†]

	Ave. daily number of cigarettes						
	0	1–4	5–14	15–24	25–49	50+	Total
Controls (n_{0j})	38	87	397	279	119	12	$n_{0+} = 932$
Cases (n_{1j})	2	19	197	171	129	21	$n_{1+} = 539$
Rates ($\hat{\lambda}_j$)	0.14	0.59	1.35	1.67	2.95	4.76*	1.57

[†] Reconstructed from data of Doll and Hill (1952), p. 1278
* Doll and Hill give 4.74 for this entry

A demonstration that absolute risks can be estimated from case-control data that are supplemented with information regarding the underlying population was

provided by Doll and Hill (1952). They restricted the analysis to cases and controls drawn from the Greater London area, for which the numbers of persons and the numbers of deaths due to lung cancer were known from government records for each category of age and sex. Table 6.1 shows numbers of male cases n_{1j} and controls n_{0j} aged 45–64 years at the jth of 6 levels of average cigarette consumption during the preceding 10 years ($j = 1, \ldots, 6$). Assuming that the smoking habits of the controls were representative of the habits of the general population in each age-sex category, and likewise that the habits of the cases were reasonably similar to those of persons who died of lung cancer, they were able to estimate the numbers of persons N_j and of deaths D_j at each of the 6 smoking levels. Specifically, knowing that the total male population of Greater London aged 45–64 was $N_+ = 937{,}000$, they estimated the sub-population (in thousands) at the jth smoking level as $\widehat{N}_j = (n_{0j}/932) \times 937$. Similarly, knowing that $D_+ = 1474$ deaths from lung cancer occurred annually in this population, they estimated the numbers of deaths at that level by $\widehat{D}_j = (n_{1j}/539) \times 1474$. Thus the absolute rates per 1000 persons per year at smoking level j were estimated as

$$\hat{\lambda}_j = \frac{\widehat{D}_j}{\widehat{N}_j} = \frac{n_{1j} \cdot n_{0+} \cdot D_+}{n_{0j} \cdot n_{1+} \cdot N_+} .$$

See Table 6.1 and Doll and Hill (1952), Table XII. Neutra and Drolette (1978) formally justified this commonly used procedure while Greenland (1987) provided an extension for matched case-control studies.

Langholz and Borgan (1997) developed more specialized methodology for estimation of absolute risks from nested case-control studies under the Cox (1972) model. The absolute risk of disease over the time period (t_0, t_1) for a subject with explanatory variables x who is disease-free at its start is

$$P(t_0, t_1; x) = \int_{t_0}^{t_1} S(t_0, t; x)\lambda(t; x)\,\mathrm{d}t \ , \tag{6.6}$$

where $S(t_0, t; x)$ denotes the probability that the subject remains on study and free of disease from t_0 to t and $\lambda(t; x)$ is the disease incidence rate. Increments in the baseline cumulative incidence rate function at each time of disease diagnosis, needed to estimate both S and λ, are obtained from the usual formula for the cohort study applied to *reduced* risk sets consisting of the case and sampled control(s). The denominator term, representing the sum of relative risks for all subjects in the risk set, is weighted by n/m where n denotes the size of the risk set and m the number of subjects, including the case, sampled from it. Benichou and Gail (1995) studied similar methodology for unmatched case-control sampling from an actual cohort when all explanatory variables are discrete. Econometricians also have developed methods for incorporation of external information on background rates into the analyses of data collected in "choice-based" sampling designs, the social science analog of case-control studies (Hsieh et al. 1985).

6.3 Matching and Stratification

> While the logical absurdity of attempting to measure an effect for a factor controlled by matching must be obvious, it is surprising how often investigators must be restrained from attempting this. (Mantel and Haenszel 1959, p. 729)

Investigators planning case-control studies used to consider matching of individual controls to cases as a means of making the two groups as comparable as possible, thereby increasing the perceived validity of study results. It is now recognized that such matching, or stratified sampling of controls to make them more like the cases – known as frequency matching, has a much more limited and specific role. This is to improve the efficiency of rate ratio estimators (exposure odds ratios) that are statistically *adjusted* to account for possible confounding effects. Inappropriate matching may have the unintended effect of compromising design efficiency or even of rendering the results completely uninterpretable. Furthermore, since the sampling design must always be considered, matching usually complicates the statistical analysis.

6.3.1 Consequences of Matching

The goal of matching in case-control studies is to balance the numbers of cases and controls within strata that will be used for statistical adjustment purposes. If the factor(s) used for stratification are associated with exposure, the matched control sample will generally have an exposure distribution more like that of the cases than would an unmatched control sample.

Some interesting and important consequences of matching are illustrated by the fictitious data shown in Table 6.2, which is adapted from Table 10-5 of Rothman and Greenland (1998). In the underlying cohort the disease rates for exposed and non-exposed are identical for males and females. Consequently, there is no effect modification nor confounding by sex and the crude (marginal) rate ratio equals the sex-specific ratios. The frequency matching of controls to cases by sex, however, has *induced* apparent confounding in the case-control data. The sex specific rate ratios are correctly estimated by the sex-specific odds ratios, in accordance with Equation (6.2), but they are substantially under-estimated by the crude exposure odds ratio. An analysis that accounts for the matching is essential to correctly estimate the interest parameter.

6.3.2 Efficiency of Matching

The advantages of a frequency matched sample become evident when one considers extreme situations. In the study of esophageal cancer of Tuyns et al. (1977), for example, 775 controls were sampled at random from electoral rolls for comparison with the 200 cases. Not surprisingly, the lowest age stratum contained only a single

Table 6.2. Distribution of cases and person-years of observation in a fictitious cohort study, and expected distribution of cases and frequency matched controls[†]

	A. Results for underlying cohort study			
	Males		Females	
	Exposed	Non-exposed	Exposed	Non-exposed
Diseased	450	10	50	90
Person-years	90,000	10,000	10,000	90,000
Rate ($\times 10^3$)	5.0	1.0	5.0	1.0
Rate ratio	$\psi_E = 5$		$\psi_E = 5$	
	Crude rate ratio $= \frac{(450+50)/100{,}000}{(10+90)/100{,}000} = 5$			

	B. Expected results for the case-control study			
	Males		Females	
	Exposed	Non-exposed	Exposed	Non-exposed
Cases	450	10	50	90
Controls	414	46	14	126
Odds ratio	$\widehat{\psi}_E = 5.0$		$\widehat{\psi}_E = 5.0$	
	Expected crude odds ratio $\approx \widehat{\psi}_E = \frac{(450+50)\times(46+126)}{(10+90)\times(414+14)} = 2$			

[†] Adapted from Table 10-5 of Rothman and Greenland (1998)

case and 115 controls. Since they contributed very little to the age-stratified odds ratio, the time spent interviewing the 115 youngest controls was largely wasted. When the potential for imbalance is less extreme, however, the advantages of matching are not so clear. Some insight is provided by considering the ratio of asymptotic variances of crude and adjusted (stratified) odds ratio estimators for frequency matched and random samples in the simplest of situations, that involving a binary exposure factor, a binary confounding factor and a rare disease. Assuming equal numbers of cases and controls, and that the exposure rate ratio ψ_E is the same at both levels of the confounder, the variances are determined by five quantities: ψ_E; p_E, the population proportion exposed; p_C, the proportion positive for the confounder; ψ_C, the rate ratio for the confounder; and ψ_{CE}, the odds ratio associating confounder and exposure in the population. Table 6.3, adapted from Breslow (1982), shows ratios of variances and biases for different odds ratio estimators when $p_C = 0.5$ and $p_E = 0.3$. Similar results were given by Thomas and Greenland (1983) and by Smith and Day (1984).

A stratified analysis is not needed to control confounding when $\psi_{CE} = 1$ or $\psi_C = 1$. For as shown in rows 1–5, 9 and 13 of Table 6.3, the bias B_R of the pooled estimator using a randomly selected control sample is then zero. Columns labeled V_M/V_R^* show the increase in variance, i.e., the loss in efficiency, if a matched control sample and stratified analysis were used instead. There is no efficiency loss through matching when $\psi_{CE} = 1$ but increasing loss for estimation of large rate

Table 6.3. Variance ratios and biases, in percent, for estimators of ψ_E in case-control studies with matched and random samples†

		$\psi_E = 1$				$\psi_E = 2$				$\psi_E = 5$				$\psi_E = 10$			
ψ_{CE}	ψ_C	$\frac{V_M}{V_R}$	$\frac{V_M}{V_{R^*}}$	B_R	B_M	$\frac{V_M}{V_R}$	$\frac{V_M}{V_{R^*}}$	B_R	B_M	$\frac{V_M}{V_R}$	$\frac{V_M}{V_{R^*}}$	B_R	B_M	$\frac{V_M}{V_R}$	$\frac{V_M}{V_{R^*}}$	B_R	B_M
1	1	100	100	0	0	100	100	0	0	100	100	0	0	100	100	0	0
	2	97	100	0	0	97	100	0	0	97	100	0	0	97	100	0	0
	5	88	100	0	0	87	100	0	0	87	100	0	0	88	100	0	0
	10	80	100	0	0	79	100	0	0	80	100	0	0	81	100	0	0
2	1	100	103	0	0	100	103	0	−4	100	103	0	−4	101	104	0	−6
	2	95	100	12	0	96	100	12	−4	97	101	12	−4	99	102	12	−5
	5	85	97	24	0	86	97	24	−1	88	98	24	−2	91	99	24	−3
	10	78	96	31	0	78	96	31	−1	81	97	31	−1	85	98	31	−2
5	1	100	113	0	0	99	113	0	−8	101	118	0	−18	105	122	0	−23
	2	93	106	27	0	93	106	27	−7	97	110	27	−14	103	114	27	−18
	5	82	99	58	0	83	99	58	−4	89	102	58	−8	95	105	58	−10
	10	76	95	75	0	77	96	75	−2	82	98	75	−5	88	100	75	−6
10	1	100	126	0	0	98	126	0	−14	100	134	0	−29	107	144	0	−37
	2	91	114	36	0	90	114	36	−11	96	120	36	−23	104	128	36	−28
	5	80	102	82	0	81	102	82	−6	88	107	82	−13	96	111	82	−16
	10	74	97	106	0	76	98	106	−4	82	101	106	−8	90	105	106	−9

† Adapted from Breslow (1982), Table 2 ψ_{CE} = Odds ratio associating exposure and confounder ψ_E = Rate ratio for exposure
ψ_C = Rate ratio for confounder B_R = Bias of pooled estimate of ψ_E, random sample, as percent of ψ_E
B_M = Bias of pooled estimate of ψ_E, matched sample, as percent of ψ_E V_M = Variance of the stratified estimate in the matched sample
V_R = Variance of the stratified estimate in the random sample V_{R^*} = Variance of the pooled estimate in the random sample

ratios when the correlation between confounder and exposure is high. Since the "confounder" is not a risk factor for disease ($\psi_C = 1$), it need not be controlled in the analysis. By needlessly matching on it, the exposure distributions for cases and controls have been made more alike, thus reducing the efficiency of estimation of the exposure effect. The negative biases associated with the crude analysis of the matched data reflect the same phenomenon as the example in Table 6.2. This is a case of *overmatching*.

Stratification *is* needed to control confounding when both $\psi_{CE} > 1$ and $\psi_C > 1$. Then, as shown in rows 6–8, 10–12 and 14–16 of the table, the bias B_R using the unadjusted design and analysis is non-zero and becomes increasingly serious as the effect of the confounder and its correlation with the exposure increase. The efficiency of the matched design to the standard design, using in both cases the correct (stratified) analysis, may be read from columns labeled V_M/V_R. Values *under* 100% indicate greater efficiency, meaning a smaller variance, for the matched design. When the potential confounder increases disease risk but exposure does not, matching is always more efficient and its efficiency increases with the degree of confounding. Even in the most extreme situation ($\psi_{CE} = \psi_C = 10$), however, no more than 26% of efficiency is lost by failure to match. A conclusion is that confounder and disease must be strongly associated for matching to produce major gains. Matching may actually *lose* efficiency when ψ_{CE} and ψ_E are both large.

Overmatching 6.3.3

Overmatching refers to matching on a factor that is not a confounder of the disease-exposure association. There are three possibilities.

Factor Related Only to Exposure. This is the situation just considered in Tables 6.2 and 6.3 (rows 5, 9, 13). Matching is not needed to control confounding and leads to a loss of efficiency.

Factor Related Only to Disease. This has been called "the case of futility" because the matching is effectively at random with respect to exposure (Miettinen 1970). Frequency matching has no effect on efficiency, as the variance ratios $V_M/V_{R^*} = 100$ when $\psi_{CE} = 1$ suggest. Were one to "incorrectly" fail to account for the matching in the analysis, however, there would be efficiency loss relative to the frequency matched analysis; note the percentages below 100 in the column labeled V_M/V_R. Individual pair matching in such circumstances could cause a loss of efficiency because of the need to account for this in the analysis and the consequent reduction in degrees of freedom for estimation of the main effect. With binary exposure measurements, for example, only the discordant case-control pairs would contribute to estimation of the exposure odds ratio, and these would become fewer and fewer as the association between the matching factor and disease increased.

"Confounder" an Intermediate in the Causal Pathway. The most serious type of overmatching occurs when one matches on a factor that is both affected by

exposure and a cause of disease. If the effect of anti-hypertensive medication on the risk of myocardial infarction was being investigated, for example, yet cases and controls were matched on blood pressure measurements taken after treatment commenced, the data would be completely useless for estimation of treatment effect. Ignoring the matching in the analysis would only compound the error by driving the odds ratio even closer towards unity.

6.3.4 When to Match

In view of the drawbacks of overmatching, and the often modest efficiency gains even when statistical adjustment is indicated, one may well ask whether matching is ever justified. The administrative costs of locating matched controls, and the loss of cases from analysis if none can be found, further argue for careful consideration of matched designs. Individual case-control matching is most appealing when needed to control the effects of a confounder that is not easily measured. The paradigm is use of an identical co-twin to control for genotype (Jablon et al. 1967). Otherwise, stratification of the control sample on gender and broad categories of age to achieve rough comparability with the case distribution, provided that this can be accomplished without great cost, is likely all that is advisable. Greater attention to stratification of the control sample may be needed when the primary goal is to evaluate statistical *interaction*, or effect modification, between exposure and a covariate (Smith and Day 1984).

6.4 Selection of Subjects

The two preceding sections outline the basic ideas of sampling of subjects for a case-control study from a theoretical statistical perspective. While the theory is an important guide to practice, implementation is usually imperfect and requires some compromise to minimize the various types of bias to which case-control studies are particularly susceptible (Sect. 6.5). In this section we consider some of the choices available to the investigator for putting the theory into action.

6.4.1 Selection of Cases

Disease Definition

Careful definition of the disease endpoint to conform to the goals of the study is critical to success. Specific cancers are reasonably well defined by primary site and histologic type. Studies of diabetes, rheumatoid arthritis or psychiatric conditions should follow standard criteria for diagnosis established by professional societies. In the typical study of disease etiology, the investigator may choose to enhance efficiency by including only those cases of disease most likely a priori to have been caused by the particular exposure. Thus, instead of "uterine cancer", studies of hormonal risk factors would best be restricted to adenocarcinoma of

the endometrium whereas those investigating sexual practices or viral etiology would focus on squamous cell cancer of the cervix. Of course, in the early stages of an investigation, demonstration that the exposure effect is *specific* to a particular disease subtype can be an important part of the evidence that the association is causal (Hill 1965; Weiss 2002). For case-control studies of the public health impact of exposure, furthermore, a broader definition of disease may be desirable.

As mentioned in Sect. 6.1.2, studies of disease etiology are best restricted to incident cases. This may not always be possible, however. Congenital anomalies are generally ascertained as those that are prevalent at birth, and consideration of possible exposure effects on fetal loss forms an important part of the interpretation. Cohort studies may be preferable for estimating the true effects of exposure on reproductive outcomes (Weinberg and Wilcox 1998).

Sources of Cases

Population Registries. Population based disease registries, particularly of cancer and birth defects, are often considered the ideal source of cases. This is because the population at risk, whose identification is needed for control selection, is well defined by geographic or administrative boundaries. Practical limitations on their use include the speed with which cases can be identified and interviewed, to avoid selection bias from exclusion of those who may have died, and the feasibility of random sampling of controls.

Health Maintenance Organizations. Large health maintenance organizations (HMOs) are advantageous as a source of cases, for several reasons. The source population is enumerated and demographic data, as well as some exposure and covariate data, may already be available for everyone. This permits judicious selection of cases and controls using nested case-control, case-cohort or stratified two-phase sampling designs (Sect. 6.6). Relatively objective and inexpensive exposure assessments may be possible using routine medical or pharmacy records, some of which may already exist in electronic form. Similarly, cases are usually easily ascertained from reports of diagnoses within the organization. Of course, some assurance is needed that members of the HMO are unlikely to go elsewhere for diagnosis and treatment.

Hospitals and Clinics. Historically, many case-control studies have been conducted using either a single or a small group of hospitals or clinics. This facilitates timely access to cases and increases the likelihood of their cooperation, thus limiting selection bias. On the other hand, definition of the source population from which the cases arose may be problematic, not to mention the practicality of obtaining random samples of controls from it.

Exclusion Criteria

In principle, any exclusion criteria may be used for cases so long as they are equally applied to the controls, and vice-versa, since they serve simply to restrict the source

population. Thus subjects may be excluded who reside in areas difficult to reach or who are not native speakers of the language of interview. Practical applications of this rule can be more subtle, however. Wacholder (1995) argues, for example, that exclusion of cancer cases who lacked a histologic diagnosis could inadvertently tend to exclude those from smaller, rural hospitals who were more likely to have exposures related to agriculture.

Exposure Opportunity. Case-control studies are most informative when there is a substantial degree of exposure variability, so that the exposure is neither rare nor ubiquitous (Chase and Klauber 1965). Subjects known a priori to have no opportunity for exposure could be excluded on grounds of efficiency if the exposure was rare, since they would contribute little additional information. Thus, for example, women who were past reproductive age when oral contraceptives became popular should be excluded from a study of OC use and breast cancer (Wacholder et al. 1992a). On the other hand, since they provide valid information on the non-exposed, there is no logical basis for insisting that subjects without the opportunity for exposure should be routinely excluded from cohort and case-control studies (Schlesselman and Stadel 1987; Poole 1987).

6.4.2 Selection of Controls

Principles of Control Selection

Wacholder et al. (1992a) described three basic principles of control selection. The first two correspond roughly to considerations already developed regarding conceptual foundations and the use of matching. The third stems from the desire to minimize the effects of measurement error to which case-control studies are particularly susceptible.

The "study-base" Principle. This is the principle that controls be randomly selected from disease-free members of the underlying cohort, also known as the source population (Kelsey et al. 1996) or study-base (Miettinen 1985), at the times that cases are being ascertained (Sect. 6.2). When controls are in fact selected later, it sometimes mandates the random selection of a reference date for each control so that the distributions of the case diagnosis dates and control reference dates are comparable. Only exposures occurring prior to the reference/diagnosis date would be taken into account. This principle also implies that whatever exclusion criteria have been applied to the cases must also be applied equally to the controls.

The Deconfounding Principle. This principle underlies the stratified sampling of controls to render possible, or improve the efficiency of, an adjusted analysis designed to control confounding (Sect. 6.3).

The Comparable Accuracy Principle. This principle, controversial even in the authors' view, suggests that controls be selected so that the errors of measurement

of their exposures and covariates are comparable to the measurement errors of the cases. The suggestion that dead controls be selected for dead cases, for example, is sometimes made on the basis of the comparable accuracy principle (Gordis 1982). Unfortunately, there is no guarantee that adherence to the principle will eliminate or even reduce bias (Greenland and Robins 1985). Unless the measurement error can be completely controlled, for example by obtaining error free measurements for a *validation* subsample of cases and controls and appropriately incorporating these data in the analysis, it can seriously compromise study validity even if case and control data are equally error prone (Sect. 6.5.3).

Sources of Controls

The appropriate source population for sampling of controls is determined by the study-base principal. When cases arise from an enumerated source population such as an HMO, controls may be sampled from this cohort using a nested case-control or case-cohort design (Figs. 6.1 and 6.2). One principal advantage of conducting epidemiologic studies in the Nordic countries is their maintenance of national disease and population registers which may be exploited for case and control selection, respectively (see Chap. I.4 of this handbook). Standard survey sampling methods are often used to select controls for "population based" studies in countries that do not maintain population registers. The most difficult and controversial problems of control selection arise with hospital based studies.

Survey Sampling. Methods for scientific sampling of populations have been developed by census bureaus and other government agencies throughout the world. The particular method most advantageous for any given epidemiologic study will likely depend on the local administrative infrastructure. Survey sampling often proceeds in stages, where one first samples a large administrative unit, then a smaller one and finally arrives at an individual household or subject. Such multi-stage "cluster" sampling introduces modest correlations in the responses of individuals sampled from the same primary sampling unit, more marked ones for individuals sampled from the same lower level cluster. Although often ignored by epidemiologists, usually at the cost of some underestimation of the variability in estimated relative risks, these correlations should be accounted for in a rigorous statistical analysis (Graubard et al. 1989). Fortunately, simple methods to accomodate cluster sampling are now routinely incorporated in the standard statistical packages.

Random Digit Dialing. In view of the high costs of census bureau techniques in the United States, methods of survey sampling through the telephone exchanges have been developed (Waksberg 1978; Harlow and Davis 1988). Random digit dialing (RDD) has become increasingly popular for control selection in populations that have high rates of telephone access. Some implementations start with the telephone exchange of each case for sampling of controls that are thereby matched on somewhat ill-defined neighborhood factors (Robison and Daigle 1984). RDD

methods may be costly for ascertainment of controls from minority populations, requiring dozens of calls to locate a suitable household (Wacholder et al. 1992b). They are particularly susceptible to bias because of higher selection probabilities for households that have more than one phone line or more than one eligible control and because of high rates of nonresponse (Sect. 6.5.1). The latter problem is likely to become increasingly serious in view of the persistent use of answering machines to screen out unwanted calls. The popularity of cell phones, moreover, eventually may make it infeasible to use RDD to draw a random control sample from a source population defined by geographic or administrative boundaries.

Neighborhood and Friend Controls. Matched controls may also be selected from neighbors or friends of each case. For the former method, a census is taken of all households in the immediate geographic area of the case and these are approached in a random order until a suitable control is found. Care must be taken to ensure that the control was resident at the same time the case was diagnosed. Even with these precautions, neighborhood sampling may yield biased controls for hospital based studies since it will not be guaranteed that the control would have been ascertained as a case if ill, thus violating the study-base principle (Wacholder et al. 1992b). Neighborhood controls are also susceptible to overmatching due to their similarity to the cases on factors associated with exposure that are not risk factors for disease (Sect. 6.3.3). These same difficulties confront the use of friend controls, whereby a random selection is taken from among a census of friends provided by each case. There may be further selection on factors related to popularity since the friend selected as control may well not have listed the case as a friend had the friend become ill (Robins and Pike 1990). The primary advantage of friend controls would be a low level of nonresponse.

Hospital Based Controls. Many studies that ascertain cases through hospitals also select controls from these same hospitals, which is of obvious logistical convenience. Such controls are likely to have the same high response levels as the cases. The fact that they may be interviewed in a hospital setting, as the cases are, is an advantage from the perspective of the comparable accuracy principle (Mantel and Haenszel 1959). The major difficulties stem from the fact that the hypothetical study-base, the *catchment* of persons who would report to the particular hospital if they developed the disease under study, may be different from the catchment population for other diseases. Furthermore, many of the disease categories from which controls could be selected may themselves be associated with the exposure. A large part of the planning of hospital based case-control studies is devoted to the choice of disease categories thought to be independent of exposure and to have a similar catchment. The hope is that controls with such diseases will effectively constitute a random sample, vis-à-vis exposure, from the study-base. Since the independence of exposure and disease diagnosis is rarely known with great certainty, a standard recommendation is to select controls having a variety of diagnoses so that the failure of any one of them to meet the criterion does not compromise the study (Wacholder et al. 1992b; Rothman and Greenland 1998, p. 101). If it is found

later that a certain diagnosis is associated with exposure, those controls can be excluded.

How Many Controls per Case? How Many Control Groups? 6.4.3

Case-Control Ratios. For a fixed number of study subjects, statistical power for testing the null hypothesis is optimized by having equal numbers of cases and controls. When the disease is extremely rare or acquisition of cases particularly expensive, however, it may be important and cost-effective to increase the numbers of controls. In order to have the same statistical power (to reject the null hypothesis of no exposure effect against local alternatives) as a design with equal numbers of cases and controls, a design with M controls per case would need only $(M + 1)/2M$ as many cases. When $M = 2$, for example, this would imply the use of 3/4 as many cases, but twice as many controls, to achieve the same power as a design with equal numbers. For a fixed number of cases, the relative efficiency of a design with M controls per case relative to one that uses an unlimited number of controls is therefore only $M/(M + 1)$. Since 80% of maximum efficiency can thus be obtained with $M = 4$, it is often inadvisable to seek a higher ratio. Exceptions occur when sampling and data collection for controls is substantially cheaper than for cases or if accurate estimation of large rate ratios, rather than a test of the null hypothesis, is the primary statistical objective (Breslow 1982; Breslow et al. 1983).

Multiple Control Groups. Early case-control investigations, including the classic study of Doll and Hill (1952), often utilized two or more control groups. Indeed, multiple control groups were recommended by Dorn (1959) to improve the case-control study so that it would "provide a more valid basis for generalization". As explained by Hill (1971, pp 47–48) "If a whole series of control groups, e.g., of patients with different diseases, gives much the same answer and only the one affected group differs, the evidence is clearly much stronger than if the affected group differs from merely one other group." Similar informal arguments have been put forward in favor of multiple control groups as a means of addressing the possible biases that may be associated with the use of any one of them (Ibrahim and Spitzer 1979). Working from a more formal perspective, Rosenbaum (1987) concluded that a second or third control group was useful *only* if supplemental information was available on whether such use addressed a specific bias. If controls sampled from separate sources have different exposure histories, even after statistical adjustment for potential confounders, this indeed suggests that similar adjustment of the case-control comparison may be inadequate to control confounding. However, failure to detect a difference among control groups may give a false sense of security unless they were deliberately selected to differ with respect to unmeasured potential confounders. Implementation of this last criterion would clearly require some guess as to what those unmeasured confounders might be.

Recent reviews of case-control methods have tended to shy away from the use of multiple control groups (Rothman and Greenland 1998, p. 106; Wacholder et al. 1992b). They argue that there is usually a single "best" control group, and that since the discovery of an adjusted exposure difference with other control groups will force these to be discarded, the effort involved will have been wasted. However, there may not be a "best" control group, or its identification may be controversial. Discovery of a difference between control groups should generally encourage the investigator to seriously suspect that confounding may have compromised study results.

6.5 Pitfalls

Case-control studies are susceptible to the same biases and problems of interpretation that afflict all observational epidemiological studies. These include confounding, selection or sampling bias, measurement error and missing data. Selection bias can be considered an extreme version of bias due to missing data where the entire observational record is missing for subjects who are in the source population but fail to be included in the study. Each of these topics is considered in detail in other chapters of this handbook. Many methods described there for dealing with such issues apply to case-control studies as well as to cohort studies. Attention is confined here to a few of the potential problems to which case-control studies are particularly susceptible.

6.5.1 Selection Bias

As elaborated at length in Sect. 6.2.1, the cases and controls in a case-control study are best viewed as resulting from outcome dependent sampling from an underlying, often idealized cohort study. The goal is to estimate the degreee of association of disease risk with exposure that would have been found had complete records been available for the entire cohort. The sampling of controls and sometimes even of cases may be stratified, for example by sex and broad categories of age, but otherwise is supposed to be random within the subpopulations of diseased and non-diseased subjects. Selection bias arises when the sampling is in fact not random. It poses a major threat to the validity of case-control studies.

The effect of sampling bias is easy to demonstrate quantitatively for an exposure variable with two levels. For simplicity, we consider the effect on the odds ratio associating exposure with the cumulative risk of disease during a defined study period. The first 2×2 subtable displayed in Table 6.4 contains the population frequencies of subjects who are exposed and become diseased during the study period (P_{11}), who are not exposed and become diseased (P_{01}) and likewise the frequences of being exposed or non-exposed and remaining disease-free (P_{10} and P_{00}, respectively). The target parameter of interest is the odds ratio ψ based on these population frequencies. As shown in the next two subtables, the odds ratio ψ^*

Table 6.4. Effect of selection bias on odds ratio measures of association

	Population frequencies		Sampling fractions		Expected sample frequencies	
	Case	Cont	Case	Cont	Case	Control
Exposed	P_{11}	P_{10}	f_{11}	f_{10}	$f_{11} \times P_{11}$	$f_{10} \times P_{10}$
Non-exposed	P_{01}	P_{00}	f_{01}	f_{00}	$f_{01} \times P_{01}$	$f_{00} \times P_{00}$
Odds ratios	$\psi = \frac{P_{11} \times P_{00}}{P_{10} \times P_{01}}$		$\psi_f = \frac{f_{11} \times f_{00}}{f_{10} \times f_{01}}$		$\psi^* = \psi_f \times \psi$	

expected from the case-control sample equals the product of the true odds ratio, ψ, times the cross products ratio of the sampling frequencies, denoted ψ_f. Hence $\psi = \psi^*$, i.e., there is no bias, provided that $\psi_f = 1$. This will occur when the sampling fractions for cases and controls are all the same, depend *only* on the disease outcome, i.e., $f_{10} = f_{00}$ and $f_{11} = f_{01}$, or depend only on exposure, i.e., $f_{01} = f_{00}$ and $f_{11} = f_{10}$. Often the sampling fractions for cases are both near 1 whereas those for the controls are much smaller. The fact that this does not matter, provided that the sampling fractions for exposed cases and non-exposed cases are the same and similarly for controls, is another way of understanding why case-control studies provide estimates of the relative risk (disease odds ratio). Bias does occur when the sampling fractions depend *jointly* on exposure and disease, usually because exposed controls are more or less likely to be sampled than non-exposed controls. In a study that ascertained all the cases, but sampled exposed persons as controls with twice the frequency as non-exposed persons, the estimated relative risk (odds ratio) would be twice the correct value. This is known as Berkson bias (Berkson 1946).

Some of the factors that contribute to selection bias are as follows.

Patient Dies Before Interview. When cases are ascertained through a population based disease registry, a significant interval of time may elapse between initial diagnosis and notification to the registry. Some patients whose disease course is rapidly fatal may therefore not be interviewed in person, but are either excluded from the study or represented by a proxy interview subject to increased measurement error. This selection factor may affect both cases *and* controls in hospital based studies. It constitutes a major problem in reproductive epidemiology (see Chap. III.5 of this handbook).

Physician Refuses Consent. Committees charged with protection of human research subjects may require that permission for participation be given by the patient's physician. This could affect control participation in hospital based studies or case participation in general.

Subject Refuses Participation. The most common reason for selection bias in case-control studies is refusal of the subject to participate, either actively by refus-

ing to sign a consent form or passively by failure to return a questionnaire or turn up at the appointed hour for a laboratory examination. Cases with disease are often highly motivated to participate, whereas controls selected from the population are not. Unfortunately, control participation rates often depend on some correlate of exposure. Refusal rates for telephone surveys, for example, are higher for people who are older, have fewer social relationships, are less well educated and have lower income (O'Neil 1979).

Subjects Ascertained Through Their Household. Selection bias can occur when controls are ascertained by first contacting households to determine whether a control lives there who is suitable for matching to the case, and only a single control is selected from each household. In studies of childhood disease, where controls are matched on age within two years of the case, a child with a sibling in the same age range is less likely to be selected than one who has no such siblings (Greenberg 1990).

Random Digit Dialing. Some other problems of selection bias are associated with the use of RDD for control ascertainment besides the fact that this method identifies households rather than individuals. Households without telephones stand no chance of selection, for example, whereas those with multiple telephones will be over-represented. The absence of a telephone may particularly affect minority populations.

6.5.2 Adjustments for Selection Bias in Study Design and Analysis

The most important consideration regarding selection bias is to avoid it so far as possible. At the design phase of the study, the exclusion criteria for both cases and controls may be chosen to maximize the probability of their ascertainment and participation. If RDD is used for control selection, this means taking the obvious step of excluding cases from households that lack telephones. Demographic, geographic and linguistic factors may enter into the exclusion criteria for the same reason.

If selection bias cannot be avoided, as much data as possible should be gathered on *potential* case and control subjects to allow *prediction* of which of them go on to participate and which refuse. When sampling from the general population, it may be possible to use a recent survey of the same population for this purpose, provided of course that the survey itself had nearly complete response. If cases and controls are drawn from an enumerated population such as an HMO, data may already exist in medical or other records that can be used for this purpose.

At the time of analysis, one may attempt to adjust for selection bias in the same way that one adjusts for missing data. This is to use sampling weights for each participating subject, i.e., those with "complete data", equal to the inverse predicted probability that the subject would have been selected given the data

collected for this purpose at the design stage. This is only useful, of course, if there is substantial variability in the predicted probabilities. Alternatively, or additionally, one may statistically adjust the analysis for factors that are thought to be associated with selection but for which data are only available for participating subjects. Such adjustment would consist of stratification of the analysis on factor levels, or inclusion of the factor in a regression model for disease given exposure, just as one adjusts for confounders (Breslow and Day 1980, Sect. 3.8). However, if there is a substantial degree of nonresponse, it is quite unlikely that any adjustment will mitigate the serious biases that can result. There is simply no way to deal with it if selection fractions within factor levels used for adjustment purposes depend jointly on disease and exposure.

Measurement Error

6.5.3

A second major limitation of case-control studies is their susceptibility to measurement error. Cases and controls are often ascertained long after the relevant exposures have occurred. In spite of Dorn's (1959) admonition to use *objective* measures of exposure, most case-control studies of environmental risk factors continue today to measure exposure by interview or questionnaire. The potential for misclassification of exposure levels in such research is enormous. First, subjects may have only a vague memory of past exposures. Second, those who are diseased at the time of interview may recall these past events in a different way than those who are healthy controls. This may be in part because the early stages of their disease led to changes in behavior that made recollection of past practices more difficult. Interviewers may solicit and record answers differently if they have knowledge of the diagnosis or of the patient's status as case or control.

Austin et al. (1994) reviewed published reports of nine case-control studies of diet and cancer in which an attempt had been made to assess the accuracy of recall of dietary histories separately for cases and controls. According to their authors, three studies provided "weak" and four "moderate" evidence for recall bias. However, these results themselves were likely subject to measurement error and may have been understated in consequence.

Measurement error, whether or not it is differential between cases and controls, can compromise conclusions by seriously biasing the relative risk estimates from case-control studies that use dietary self reports or similarly error-prone measurements. Prentice (1996) developed a mathematical model for measurement error that allowed for correlation of the error with the true exposure level and for systematic underreporting of exposure for persons with high exposure levels. He fitted the model to replicate measures of dietary fat intake, some taken using a four day food record and others using a food-frequency questionnaire, for control subjects enrolled in the Women's Health Trial (Henderson et al. 1990). Employing results from international geographic correlation studies to generate the "true model", in which subjects at the 90th percentile of the distribution of dietary fat intake had 3 or 4 times the risk of disease as those at the 10th percentile, he showed that measurement error could plausibly reduce the

relative risks to 1.1. The obvious conclusion from these calculations was that "dietary self-report instruments may be inadequate for analytic epidemiologic studies of dietary fat and disease risk because of measurement error biases" (Prentice 1996).

A substantial and concerted effort has been made by statisticians to develop methods of data analysis that correct for the bias in relative risk estimates caused by measurement error (see Chap. II.5 of this handbook and the text by Carroll et al. 1995). Some require the availability of "gold standard", i.e., error-free, measurements on a fairly large number of subjects in the validation subsample. Others assume that statistically independent true replicate measurements are available. Unfortunately, data collected in case-control studies rarely meet these stringent requirements, at least not in their entirety. It therefore behooves us to recall Bradford Hill's (1953, p. 995) sage advice:

> One must go and seek more facts, paying less attention to techniques of handling the data and far more to the development and perfection of the methods of obtaining them.

6.6 Conclusions

The case-control study played a major, successful role during the second half of the twentieth century in identifying risk factors for chronic disease. It has also proven helpful for evaluation of the efficacy of vaccination (Comstock 1994) and screening (Weiss 1994) programs. The twenty-first century will witness its continued use as a cost-effective study design, with increasing application in genetic epidemiology (Khoury and Beaty 1994) and particularly in the study of gene-environment interactions (Andrieu and Goldstein 1998). Statisticians and epidemiologists will continue to develop more efficient study designs and methods of data analysis that take full advantage of all available data. When a case-control study is conducted in an HMO, for example, some data will likely be available on either the exposure or the control variables for all subjects in the underlying cohort. *Two-phase* sampling designs, whereby *biased* samples of cases and controls are selected using the data available for all subjects, then offer the potential for much greater efficiency than the standard case-control design (White 1982; Breslow and Cain 1988; Langholz and Borgan 1995; Breslow and Chatterjee 1999). Chapter I.7 of this handbook discusses these and other evolving study designs and analyses.

The advantages of case-control methodology in terms of speed and cost may have also contributed, ironically, to a diminished stature for epidemiology and biostatistics in the eyes both of the scientific community and of the general public (Breslow 2003). Part of the problem is an inherent aversion to the "black box" approach of risk factor epidemiology that associates cause and effect without the need for any understanding of pathogenetic mechanisms. Epidemiologic findings are most convincing when supported by relevant laboratory research. Another part

of the problem is the saturation of the news media with conflicting reports based on case-control and other studies that are too small, poorly designed, improperly analyzed or overly interpreted. Taubes (1995) began his controversial and influential article on the limitations of epidemiology with the observation: "The news about health risks comes thick and fast these days, and it seems almost constitutionally contradictory." The epidemiologists he interviewed for this article cited the ability of confounding, selection bias and measurement error to overwhelm smaller exposure effects. One even suggested that no single study, no matter how well conducted, should be viewed as "persuasive" unless the lower limit of the 95% confidence interval for the rate ratio exceeded 3 or 4. Very few published studies, even when reported by the press as "suggestive" of an association, meet this stringent criterion.

Medical science and public health would be well served by fewer, larger case-control studies designed to test specific hypotheses that are carefully articulated in advance. Studies that can barely "detect" a relative risk of 2 may not provide convincing evidence of a dose-response gradient and are unlikely to enable one to determine whether an elevated relative risk in a particular disease subgroup, even one specified in advance, is evidence for the *specificity* of association that can be useful in causal interpretation (Weiss 2002). (There are of course exceptions, as when a unique exposure contributes to an outbreak of an extremely rare disease. Recall the DES-adenocarcinoma of the vagina story mentioned in the Introduction.) Investigators are also well advised to develop a strict protocol for selection of cases and controls and for collection and *analysis* of the data. Doll and Hill (1952) utilized such a protocol. They also had the advantage of working during the punch card era that discouraged "data dredging" and the inclusion of all but the most important variables in the analysis. A reasonable strategy might be to perform a maximum of three carefully planned analyses of the association between the primary exposure and disease: one without adjustment; one adjusted for a short list of confounders known a priori to be associated with disease; and the third adjusted for a specified list of known and suspected confounders. In case of conflict, the major interpretation would be based on the second analysis though the results of all three would be reported. Flexibility would be needed in application, of course, especially to accommodate changes in the study protocol after the study had commenced. Finally, investigators would be well advised to exercise greater caution in advertising their findings to the press before confirmation was forthcoming from other sources. By following basic principles of good statistical and scientific practice, the case-control study can gain credibility within the research community and enhance its standing as a basis for public health action.

Acknowledgements. I am indebted to Sander Greenland, Noel Weiss, the Editors and an anonymous referee for helpful comments on an earlier draft. This work was supported in part by grant R01-CA40644 from the US Public Health Service.

References

Aird L, Bentall HH, Roberts JAF (1953) A relationship between cancer of stomach and the ABO blood groups. British Medical Journal 1:799–801

Andrieu N, Goldstein AM (1998) Epidemiologic and genetic approaches in the study of gene-environment interaction: An overview of available methods. Epidemiologic Reviews 20:137–147

Armenian HK, Lilienfeld DE (1994) Overview and historical perspective. Epidemiologic Reviews 16:1–5

Armstrong RW, Armstrong MJ, Yu MC, Henderson BE (1983) Salted fish and inhalants as risk-factors for nasopharyngeal carcinoma in Malaysian Chinese. Cancer Research 43:2967–2970

Austin H, Hill HA, Flanders WD, Greenberg RS (1994) Limitations in the application of case-control methodology. Epidemiologic Reviews 16:65–76

Barlow WE (1994) Robust variance estimation for the case-cohort design. Biometrics 50:1064–1072

Benichou J, Gail MH (1995) Methods of inference for estimates of absolute risk derived from population-based case-control studies. Biometrics 51:182–194

Berkson J (1946) Limitations of the application of fourfold table analysis to hospital data. Biometrics Bulletin 2:47–53

Breslow N (1982) Design and analysis of case-control studies. Annual Review of Public Health 3:29–54

Breslow NE (1996) Statistics in epidemiology: The case-control study. Journal of the American Statistical Association 91:14–28

Breslow NE (2003) Are statistical contributions to medicine undervalued? Biometrics 59:1–8

Breslow NE, Cain KC (1988) Logistic regression for two-stage case-control data. Biometrika 75:11–20

Breslow NE, Chatterjee N (1999) Design and analysis of two-phase studies with binary outcomes applied to Wilms tumor prognosis. Applied Statistics 48:457–468

Breslow NE, Day NE (1980) Statistical methods in cancer research I: The analysis of case-control studies. International Agency for Research on Cancer, Lyon

Breslow NE, Lubin JH, Marek P, Langholz B (1983) Multiplicative models and cohort analysis. Journal of the American Statistical Association 78:1–12

Broders AC (1920) Squamous-cell epithelioma of the lip. A study of five hundred and thirty-seven cases. Journal of the American Medical Association 74:656–664

Carroll RJ, Ruppert D, Stefanski LA (1995) Measurement error in nonlinear models. Chapman and Hall, London

Chase G, Klauber MR (1965) A graph of sample sizes for retrospective studies. American Journal of Public Health 55:1993–1996

Cole P (1979) The evolving case-control study. Journal of Chronic Disease 32:15–27

Comstock GW (1994) Evaluating vaccination effectiveness and vaccine efficacy by means of case-control studies. Epidemiologic Reviews 16:77–89

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

Correa A, Stewart WF, Yeh HC, Santos-Burgoa C (1994) Exposure measurement in case-control studies: Reported methods and recommendations. Epidemiologic Reviews 16:18–32

Cox DR (1972) Regression models and life-tables (with discussion). Journal of the Royal Statistical Society (Series B) 34:187–220

Daling JR, Weiss NS, Metch BJ, Chow WH, Soderstrom RM, Stadel BV (1985) Primary tubal infertility in relation to the use of an intrauterine-device. New England Journal of Medicine 312:937–941

Doll R, Hill AB (1950) Smoking and carcinoma of the lung. Preliminary report. British Medical Journal 2:739–748

Doll R, Hill AB (1952) A study of the aetiology of carcinoma of the lung. British Medical Journal 2:1271–1286

Dorn HF (1959) Some problems arising in prospective and retrospective studies of the etiology of disease. New England Journal of Medicine 261:571–579

Fleming PJ, Gilbert R, Azaz Y, Berry PJ, Rudd PT, Stewart A, Hall E (1990) Interaction between bedding and sleeping position in the sudden-infant-death-syndrome – a population based case-control study. British Medical Journal 301:85–89

Gordis L (1982) Should dead cases be matched to dead controls? American Journal of Epidemiology 115:1–5

Graubard BI, Fears TR, Gail MH (1989) Effects of cluster sampling on epidemiologic analysis in population-based case-control studies (Corr: V47 p. 779–780). Biometrics 45:1053–1071

Greenberg ER (1990) Random digit dialing for control selection – A review and a caution on its use in studies of childhood-cancer. American Journal of Epidemiology 131:1–5

Greenland S (1987) Estimation of exposure-specific rates from sparse case-control data. Journal of Chronic Disease 40:1087–1094

Greenland S, Robins JM (1985) Confounding and misclassification. American Journal of Epidemiology 122:495–506

Greenland S, Thomas DC (1982) On the need for the rare disease assumption in case-control studies. American Journal of Epidemiology 116:547–553

Harlow BL, Davis S (1988) Two one-step methods for household screening and interviewing using random digit dialing. American Journal of Epidemiology 127:857–863

Henderson MM, Kushi LH, Thompson DJ, Gorbach SL, Clifford CK, Thompson RS (1990) Feasibility of a randomized trial of a low-fat diet for the prevention of breast-cancer – Dietary compliance in the womens health trial vanguard study. Preventive Medicine 19:115–133

Herbst AL, Ulfelder H, Poskanzer DC (1971) Adenocarcinoma of the vagina. New England Journal of Medicine 284:878–881

Hill AB (1953) Observation and experiment. New England Journal of Medicine 248:995–1001

Hill AB (1965) The environment and disease: association or causation? Proceedings of the Royal Statistical Society of Medicine 58:295–300

Hill AB (1971) Principles of medical statistics. Oxford University Press, New York

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47:663–685

Hsieh DA, Manski CF, McFadden D (1985) Estimation of response probabilities from augmented retrospective observations. Journal of the American Statistical Association 80:651–662

Ibrahim MA, Spitzer WO (1979) The case-control study: The problem and the prospect. Journal of Chronic Disease 32:139–144

Jablon S, Neel JV, Gershowitz H, Atkinson GF (1967) The NAS-NRC twin panel: Methods of construction of the panel, zygosity diagnosis, and proposed use. American Journal of Human Genetics 19:133–161

Kelsey JL, Whittemore AS, Evans AS, Thompson WD (1996) Methods in observational epidemiology, 2nd edn. Oxford University Press, New York

Khoury MJ, Beaty TH (1994) Applications of the case-control method in genetic epidemiology. Epidemiologic Reviews 16:134–150

Kupper LL, McMichael AJ, Spirtas R (1975) A hybrid epidemiologic study design useful in estimating relative risk. Journal of the American Statistical Association 70:524–528

Lane-Claypon JE (1926) A further report on cancer of the breast. Her Majesty's Stationery Office, London

Langholz B, Borgan O (1995) Counter-matching: A stratified nested case-control sampling method. Biometrika 82:69–79

Langholz B, Borgan O (1997) Estimation of absolute risk from nested case-control data. Biometrics 53:767–774

Langholz B, Goldstein L (1996) Risk set sampling in epidemiologic cohort studies. Statistical Science 11:35–53

Levin ML, Goldstein H, Gerhardt PR (1950) Cancer and tobacco smoking. A preliminary report. Journal of the American Medical Association 143:336–338

Liddell FDK, McDonald JC, Thomas DC (1977) Methods of cohort analysis: Appraisal by application to asbestos mining. Journal of the Royal Statistical Society (Series A) 140:469–491

Lilienfeld AM, Lilienfeld DE (1979) A century of case-control studies: Progress? Journal of Chronic Disease 32:5–13

Lin DY, Ying Z (1993) Cox regression with incomplete covariate measurements. Journal of the American Statistical Association 88:1341–1349

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York

Lombard HL, Doering CR (1928) Cancer studies in Massachusetts. 2. Habits, characteristics and environment of individuals with and without cancer. New England Journal of Medicine 198:481–487

MacMahon B, Cole P, Lin TM, Lowe CR, Mirra AP, Ravnihar B, Salber EJ, Valaoras VG, Yuasa S (1970) Age at first birth and breast cancer risk. Bulletin of the World Health Organization 43:209–221

Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute 22:719–748

Miettinen O (1982) Design options in epidemiologic research: An update. Scandinavian Journal of Work Environment and Health 8:7–14

Miettinen OS (1970) Matching and design efficiency in retrospective studies. American Journal of Epidemiology 91:111–118

Miettinen O (1976) Estimability and estimation in case-referent studies. American Journal of Epidemiology 103:226–235

Miettinen OS (1985) Theoretical epidemiology: Principles of occurrence research in medicine. Wiley, New York

Neutra RR, Drolette ME (1978) Estimating exposure-specific disease rates from case-control studies using Bayes theorem. American Journal of Epidemiology 108:214–222

Neyman J (1955) Statistics – servant of all sciences. Science 122:401–406

O'Neil MJ (1979) Estimating the nonresponse bias due to refusals in telephone surveys. Public Opinion Quarterly 43:218–232

Poole C (1987) Critical appraisal of the exposure-potential restriction rule. American Journal of Epidemiology 125:179–183

Prentice RL (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73:1–11

Prentice RL (1996) Measurement error and results from analytic epidemiology: dietary fat and breast cancer. Journal of the National Cancer Institute 88:1738–1747

Prince AM, Szmuness W, Michon J, Demaille J, Diebolt G, Linhard J, Quenum C, Sankale M (1975) A case-control study of the association between primary liver cancer and hepatitis B infection in Senegal. International Journal of Cancer 16:376–383

Robins J, Pike M (1990) The validity of case-control studies with nonrandom selection of controls. Epidemiology 1:273–284

Robins JM, Gail MH, Lubin JH (1986) More on 'biased selection of controls for case-control analyses of cohort studies'. Biometrics 42:293–29

Robison LL, Daigle A (1984) Control selection using random digit dialing for cases of childhood cancer. American Journal of Epidemiology 120:164–165

Rodrigues L, Kirkwood BR (1990) Case-control designs in the study of common diseases: Updates on the demise of the rare disease assumption and the choice of sampling scheme for controls. International Journal of Epidemiology 19:205–213

Rosenbaum PR (1987) The role of a second control group in an observational study (with discussion). Statistical Science 2:292–316
Rothman KJ (1986) Modern epidemiology. Little, Brown, Boston
Rothman KJ, Greenland S (1998) Modern epidemiology, 2nd edn. Lippincott-Raven, Philadelphia
Schlesselman JJ (1982) Case-control studies. Oxford University Press, New York
Schlesselman JJ, Stadel BV (1987) Exposure opportunity in epidemiologic studies. American Journal of Epidemiology 125:174–178
Sheehe PR (1962) Dynamic risk analysis in retrospective matched pair studies of disease. Biometrics 18:323–341
Smith DC, Prentice R, Thompson DJ, Herrmann W (1975) Association of exogenous estrogen and endometrial carcinoma. New England Journal of Medicine 293:1164–1167
Smith PG, Day NE (1984) The design of case-control studies: The influence of confounding and interaction effects. International Journal of Epidemiology 13:356–365
Smith PG, Rodrigues LC, Fine PEM (1984) Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. International Journal of Epidemiology 13:87–93
Taubes G (1995) Epidemiology faces its limits. Science 269:164–169
Thomas DC, Greenland S (1983) The relative efficiencies of matched and independent sample designs for case-control studies. Journal of Chronic Disease 36:685–697
Tuyns AJ, Péquignot G, Jensen OM (1977) Le cancer de l'oesophage en Ille-et-Vilaine en fonction des niveaux de consommation d'alcool et de tabac. Bulletin du Cancer 64:45–60
Wacholder S (1995) Design issues in case-control studies. Statistical Methods in Medical Research 4:293–309
Wacholder S, McLaughlin JK, Silverman DT, Mandel JS (1992) Selection of controls in case-control studies I. Principles. American Journal of Epidemiology 135:1019–1028
Wacholder S, Silverman DT, McLaughlin JK, Mandel JS (1992) Selection of controls in case-control studies II. Types of controls. American Journal of Epidemiology 135:1029–1041
Waksberg J (1978) Sampling methods for random digit dialing. Journal of the American Statistical Association 73:40–46
Weinberg CR, Wilcox AJ (1998) Reproductive epidemiology. In: Rothman KJ, Greenland S (eds) Modern epidemiology, 2nd edn., Chap. 29, pp 585–608. Lippincott-Raven, Philadeplphia
Weiss NS (1994) Application of the case-control method in the evaluation of screening. Epidemiologic Reviews 16:102–108
Weiss NS (2002) Can the 'specificity' of an association be rehabilitated as a basis for supporting a causal hypothesis? Epidemiology 13:6–8
White JE (1982) A two stage design for the study of the relationship between a rare exposure and a rare disease. American Journal of Epidemiology 115:119–128

Wynder EL, Graham EA (1950) Tobacco smoking as a possible etiologic factor in bronchogenic carcinoma. A study of six hundred and eighty-four proved cases. Journal of the American Medical Association 143:329–336

Yu MC, Ho JHC, Lai SH, Henderson BE (1986) Cantonese-style salted fish as a cause of nasopharyngeal carcinoma - Report of a case-control study in Hong-Kong. Cancer Research 46:956–961

Ziel HK, Finkle WD (1975) Increased risk of endometrial carcinoma among users of conjugated estrogens. New England Journal of Medicine 293:1167–1170