16
Case-control studies

In a cohort study, the relationship between exposure and disease incidence is investigated by following the entire cohort and measuring the rate of occurrence of new cases in the different exposure groups. The follow-up allows the investigator to register those subjects who develop the disease during the study period and to identify those who remain free of the disease. In a *case-control* study the subjects who develop the disease (the cases) are registered by some other mechanism than follow-up, and a group of healthy subjects (the controls) is used to represent the subjects who do not develop the disease. In this way the need for follow-up is eliminated. If there is no relationship between exposure and disease incidence the distribution of exposure among the cases should be the same as the distribution among the controls.

Historically the aim of case-control studies was limited to testing for association between exposure and disease. Often little thought went into the selection of control groups, or even of cases to be studied. Frequently, studies were carried out using whatever cases could be traced from medical records at a given centre. In this rather careless climate, case-control studies fell into disrepute. However, it is now understood that properly conducted case-control studies allow *quantitative* estimates of exposure effects and this discovery has clarified the fundamental assumptions of the method. It has also contributed to a clearer understanding of the design of case-control studies issues and to a considerable improvement in the quality of studies.

We shall look first at estimating exposure effects and then consider how best to select controls. In the last section of the chapter there is a brief account of some of the difficulties which arise when case-control studies are based on prevalent rather than incident cases.

16.1 The probability model in the study base

Every case-control study of incidence can be seen within the context of an underlying cohort which supplies the cases on which the case-control study depends. A useful terminology refers to this underlying cohort, observed for the duration of the study, as the study *base*.

To estimate the quantitative relationship between exposure and disease

incidence we need to look more closely at what is happening in the study base. Consider the simple situation where the study base is divided into two groups, unexposed and exposed, and let π_0, π_1 be the probabilities that a member of the unexposed or the exposed group will fail over the period of the study and become a case.

The branches in the probability tree shown in Fig. 16.1 refer to the different possibilities for a randomly chosen member of the study base, and the events are taken in order of occurrence. The first branching of the tree refers to exposure. The subject may have been exposed (E+), or not (E−); we have taken the probability that a subject was exposed as 0.1, for illustration. The next branching refers to failure. The subject may fail (F), or survive (S); these are the probabilities already referred to as π_1 for the exposed group and π_0 for the unexposed group. The final branching refers to whether the subject is selected into the study or not; for illustration we have chosen a probability of 0.97 that a failure is registered and therefore included as a case, and a probability of 0.01 that a surviving subject is selected as one of the sample of controls. Note that the probability that a failure is registered is assumed to be the same for both exposure groups, and the probability that a healthy subject is chosen as a control is assumed to be the same for both exposure groups.

There are 8 possible outcomes for a member of the study base, corresponding to the 8 tips of the tree, but only 4 of these appear in the study. The four outcomes corresponding to the case-control study are: exposed cases, exposed controls, unexposed cases and unexposed controls. The numbers of subjects in these categories are referred to as D_1, H_1, D_0, H_0, respectively, where D refers to cases, H to healthy controls, and the suffices 1 and 0 refer to exposed and unexposed. The probabilities of the four outcomes appearing in the case-control study are calculated by multiplying conditional probabilities along the branches, and are shown to the right of the figure.

The estimation of the disease exposure relationship in the study base from the results of the case-control study may be approached using either a *retrospective* conditional argument or a *prospective* conditional argument. These correspond to two different ways of reorganizing the probability tree.

16.2 The retrospective probability model

In this argument we re-express our model as a model for the conditional probabilities of exposure given that the subject was a case (F) or a control (S). The reordering of the probability tree to reflect this argument is shown in Fig. 16.2. We define the parameter Ω_1 as the odds of having been exposed for a case. From Fig. 16.2, Ω_1 is related to the odds of failure in the study

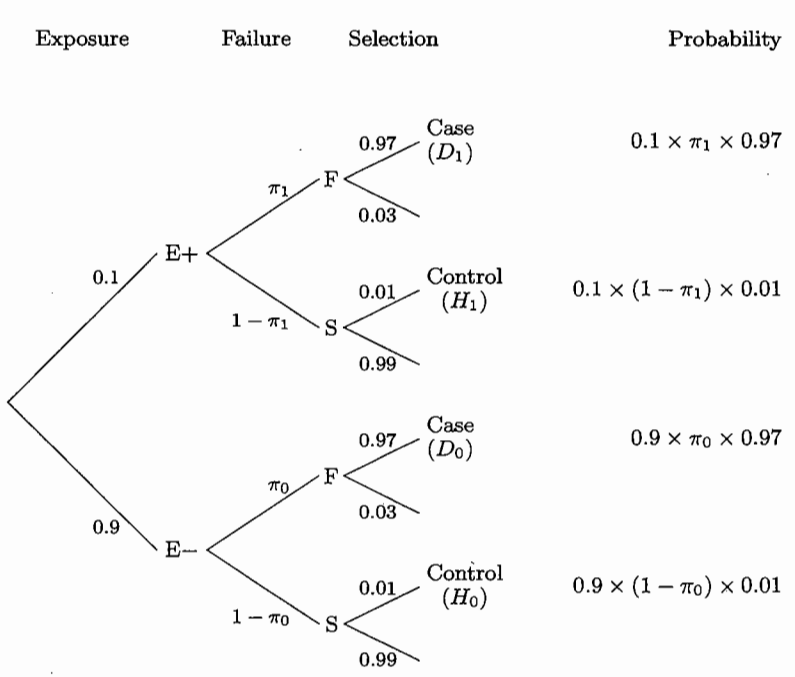


Fig. 16.1. The probability model in the study base.

base by the equations

$$\Omega_1 = \frac{0.1 \times \pi_1 \times 0.97}{0.9 \times \pi_0 \times 0.97} = \frac{0.1}{0.9} \times \frac{\pi_1}{\pi_0}.$$

The value of Ω_1 can be estimated by D_1/D_0, the ratio of exposed to unexposed cases. Similarly, we define Ω_0 as the odds of a having been exposed for a control. From Fig. 16.2,

$$\Omega_0 = \frac{0.1 \times (1-\pi_1) \times 0.01}{0.9 \times (1-\pi_0) \times 0.01} = \frac{0.1}{0.9} \times \frac{1-\pi_1}{1-\pi_0},$$

and the value of Ω_0 can be estimated by H_1/H_0, the ratio of exposed to unexposed controls. Finally the *odds ratio*

$$\frac{\Omega_1}{\Omega_0} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

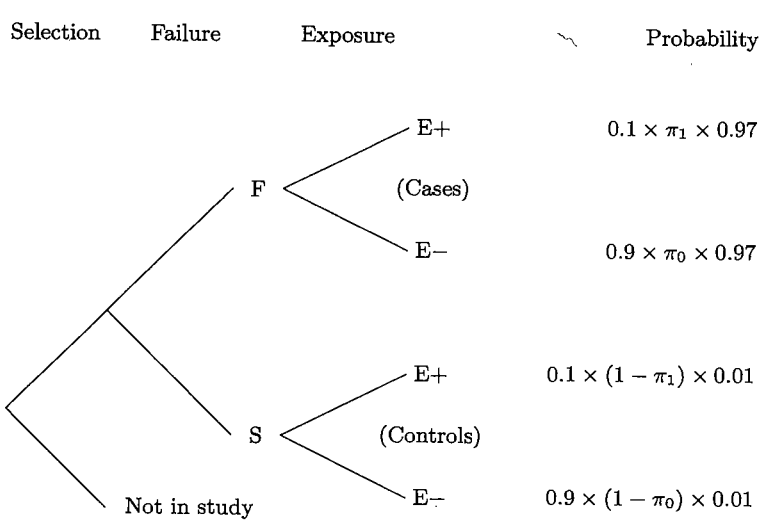


Fig. 16.2. The probability tree for the retrospective argument.

can be estimated by

$$\frac{D_1/D_0}{H_1/H_0}.$$

Thus although it is not possible to estimate π_0 and π_1 separately from a case-control study it *is* possible to estimate the odds ratio.

EXAMPLE: BCG VACCINATION AND LEPROSY

The data in Table 16.1 are from a rather unusual example of a case-control study in which the controls were obtained from a 100% cross-sectional survey of the study base.* The aim of the study was to investigate whether BCG vaccination in early childhood, whose purpose is to protect against tuberculosis, confers any protection against leprosy, which is caused by a closely related bacillus. New cases of leprosy reported during a given period in a defined geographical area were examined for presence or absence of the characteristic scar left by BCG vaccination. During approximately the same period, a 100% survey of the population of this area had been carried out, and this survey included examination for BCG scar. The tabulated data refer only to subjects under 35, because persons over the age of 35 at the time of the study would have been children at a time when vaccination was not widely available.

*From Fine, P.E.M. *et al.* (1986) *The Lancet*, August 30 1986, 499–502.

Table 16.1. BCG scar status in new leprosy cases and in a healthy population survey

BCG scar	Leprosy cases	Population survey
Present	101	46 028
Absent	159	34 594

Table 16.2. A simulated study with 1000 controls

BCG scar	Leprosy cases	Population survey
Present	101	554
Absent	159	446

Exercise 16.1. Estimate the odds of BCG vaccination for leprosy cases and for the controls. Estimate the odds ratio and hence the extent of protection against leprosy afforded by vaccination.

This example provides a good illustration of the potential economy of the case-control approach. Here a population survey was available for control but had it not been there would have been no need to carry out such a large-scale exercise. The precision of the odds ratio estimate is dominated by the precision of the odds for BCG scar among the 260 leprosy cases. Perhaps 1000 suitably chosen controls would be enough to estimate the corresponding odds among healthy subjects— there is little gain in precision to be obtained by using 80 000!

Exercise 16.2. Table 16.2 shows the results of a computer-simulated study which picked 1000 controls at random. What is the odds ratio estimate in this study?

16.3 The prospective probability model

In this argument we re-express our model in terms of the conditional probabilities of failure given selection into the study and given exposure status. The re-ordering of the conditional probability tree to reflect this argument is shown in Fig. 16.3. Define the parameter ω_1 as the *odds of being a case* for exposed subjects. By the odds of being a case we mean

$$\frac{\text{Probability of failure given that the subject is in the study}}{\text{Probability of survival given that the subject is in the study}}$$

From Fig. 16.3

$$\omega_1 = \frac{0.1 \times \pi_1 \times 0.97}{0.1 \times (1-\pi_1) \times 0.01} = \frac{0.97}{0.01} \times \frac{\pi_1}{1-\pi_1},$$

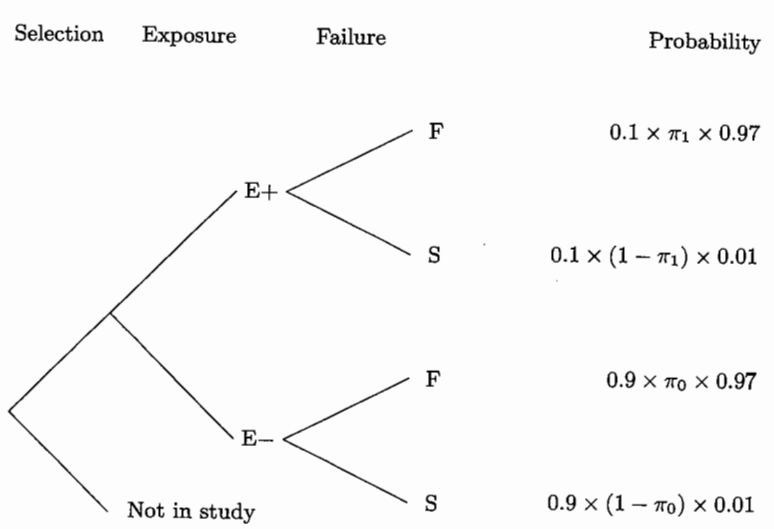


Fig. 16.3. The prospective probability model.

and this can be estimated by the case/control ratio among exposed subjects, D_1/H_1. Similarly the odds of being a case for unexposed subjects is

$$\omega_0 = \frac{0.9 \times \pi_0 \times 0.97}{0.9 \times (1-\pi_0) \times 0.01} = \frac{0.97}{0.01} \times \frac{\pi_0}{1-\pi_0},$$

which can be estimated by the case/control ratio among unexposed subjects, D_0/H_0. Finally, the odds ratio

$$\frac{\omega_1}{\omega_0} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)},$$

can be estimated by

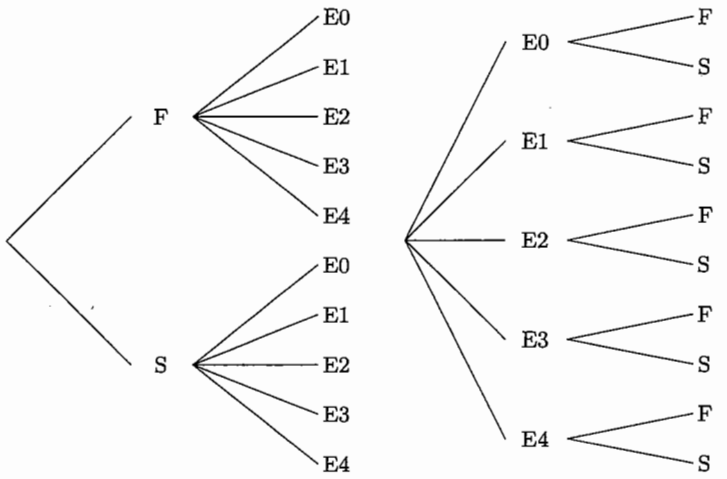
$$\frac{D_1/H_1}{D_0/H_0}.$$

This is the same estimate as that obtained from the retrospective approach since

$$\frac{D_1/D_0}{H_1/H_0} = \frac{D_1/H_1}{D_0/H_0} = \frac{D_1 H_0}{D_0 H_1}.$$

16.4 Many levels of exposure

In the retrospective argument it is the exposure status which is the response (outcome variable); in the prospective argument it is the disease status which is the response. The retrospective argument is more natural,



(a) Retrospective model (b) Prospective model

Fig. 16.4. Five exposure categories.

but the prospective argument leads to the same answers and is more convenient when studying exposures with many levels. This is illustrated by Fig. 16.4, which shows probability trees for both arguments when there are 5 exposure categories. Disease status is indicated by F (for cases) or S (for controls) and the 5 exposure categories are labelled E0 to E4. To construct a likelihood using the retrospective likelihood we must use a probability model for a response with 5 possible outcomes, but the prospective argument only requires the binary probability model. The odds of being a case for subjects in exposure category i is a constant multiple of the corresponding odds of failure in the study base; with the selection probabilities assumed in Fig. 16.1,

$$\omega_i = \frac{\pi_i}{1-\pi_i} \times \frac{0.97}{0.01}.$$

As the complexity of the exposure grouping increases, the retrospective probability model must become ever more complex, while the prospective model remains binary.

As an example of an exposure with more than two levels we shall look at a famous study carried out in the middle of the nineteenth century by William Guy.† This was possibly the first case-control study. The level of physical activity of the occupations of pulmonary tuberculosis outpatients (cases) was compared with that of other outpatients (controls). The data

†From Guy, W.A. (1843) *Journal of the Royal Statistical Society*, **6**, 197–211.

Table 16.3. Physical exertion at work of 1659 outpatients

Level of exertion in occupation	Pulmonary consumption (Cases)	Other diseases (Controls)	Case/ control ratio	Estimated odds ratio
Little	125	385	0.325	1.64
Varied	41	136	0.301	1.52
More	142	630	0.225	1.14
Great	33	167	0.198	1.00

Table 16.4. Alcohol and tobacco use by oral cancer cases and (controls)

Alcohol (oz/day)	Tobacco (cigarette equivalents per day)							
	0		1–19		20–39		40+	
0	10	(38)	11	(26)	13	(36)	9	(8)
0.1 – 0.3	7	(27)	16	(35)	50	(60)	16	(19)
0.4 – 1.5	4	(12)	18	(16)	60	(49)	27	(14)
1.6 +	5	(8)	21	(20)	125	(52)	91	(27)

are shown in Table 16.3. There are four levels of exposure corresponding to different levels of activity and the table shows the ratio of cases to controls. Each of these case-control ratios estimates some constant times the odds of failure conditional on exposure level. Since the constant depends on the probability of registration for cases and selection for controls it will be the same for all exposure levels and the case/control ratios can be compared as though they were the odds of failure.

Looking at the case/control ratios in this way, they suggest that there is a steady increase in the odds of failure (and hence the incidence rate) with decreasing level of physical activity. The table also shows odds ratio estimates with the 'great' activity category taken as reference. By definition, the odds ratio for this reference category is 1. The natural choice of reference category is the one with lowest exposure to adverse factor. In some cases, however, the natural reference category might contain very few cases and controls, leading to poor estimation of all the odds ratios; another reference category should then be chosen.

Exercise 16.3. Table 16.4 shows the distribution of 483 cases of oral cancer by level of alcohol consumption and level of tobacco consumption, together with the corresponding distribution for 447 controls.[‡] Calculate the case/control ratios, and describe the joint action of the two exposures.

[‡]From Rothman, K.J. and Keller, A.Z. (1972) *Journal of Chronic Diseases*, **23**, 711–716.

16.5 Incidence density sampling

We saw in Chapter 1 that, when the probabilities of failure are small, the risk and odds parameters are approximately equal. In these conditions, we showed in Chapter 5 that the risk parameter is also approximately equal to the cumulative rate, λt. It follows that

$$\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \approx \frac{\pi_1}{\pi_0} \approx \frac{\lambda_1}{\lambda_0}$$

for rare events. These ratios are known as the *odds ratio*, the *risk ratio*, and the *rate ratio*, and the condition for these to be approximately the same is usually described as the *rare disease assumption*. Taken together with the arguments developed in this chapter, we see that the odds ratio in a case-control study may be used to estimate the rate ratio in the underlying study base. There are two additional assumptions in this argument:

1. all subjects in the base are observed from the beginning of the study period, that is, there are no *late entries*;
2. all subjects who do not fail from the cause of interest will remain under observation until the end of the study period, that is, there is no *censoring*.

In practice, these assumptions are more likely to be violated than the rare disease assumption.

All of these assumptions can be guaranteed by the simple device of selecting a short enough study period. If insufficient cases would be obtained from such a study then the remedy is simple – carry out several *consecutive* short studies. The subjects remaining in the base at the end of one study immediately enter the next study. Each study then provides a separate estimate of the rate ratio, and provided this ratio remains constant over the whole study period, the information can be aggregated using methods very similar to those discussed in Chapter 15.

Taken to the limit, the total time available for the study may be divided into clicks which contain at most one case. Those clicks in which no case occurs are not informative so there is no purpose in drawing controls, but controls are drawn for all clicks in which a case occurs. Thus one or more controls are drawn from the study base immediately after the occurrence of each case. This design is termed *incidence density sampling*.

A study carried out in this way involves matching of controls to cases with respect to time. Methods for stratified case-control studies will be discussed in Chapter 18, but in the special case where the ratio of exposed to unexposed persons in the study base does not vary appreciably over the study period, it is legitimate to ignore the matching by time during the analysis.

One practical problem with this sampling method is that it is possible for the same individual to be included in the study more than once. For example, a control drawn at one point in time may later become a case or may be selected as a control a second time. Is it legitimate to carry out analyses which count the same person more than once? In Chapter 4 we saw that a single subject observed through several consecutive time bands can be treated as a series of different subjects, one for each band. In exactly the same way, in a case-control study it turns out to be correct to allow subjects to be sampled again in later time bands and treated as independent controls.

16.6 Nested case-control studies and case-cohort studies

An important use of incidence density sampling is in *nested* case-control studies, where case-control analysis is used in cohort studies. This is an attractive option whenever the assessment of exposure of any subject is, for some reason or other, expensive. For example, in dietary studies, individual diet may have been assessed by very detailed diary records of food intake, perhaps referring to several periods of time. The coding and transcription of such records for computer analysis is laborious and expensive. Much of this work is avoided in a nested case-control study by coding these records only for cases, as they occur, and for groups of controls drawn for each case. Since there is (usually) little to be gained by drawing more than five controls for each case, there are considerable savings to be made by such a strategy. We shall discuss the design and analysis of nested case-control studies in Chapter 33.

In recent years some authors have suggested that there are sometimes practical advantages in selecting controls by taking a single random sample of the cohort at the *beginning* of the study. This type of study has been termed a *case-cohort* or *case-base* study. If the disease is rare and there is little loss to follow-up, then the analysis may be carried out as usual, after first removing from the control sample any individuals who later became cases. However, if stratification by time becomes necessary the analysis is more difficult.

16.7 Selection bias

One important reason for obtaining wrong answers from case-control studies is incorrect sampling of controls (or cases) from the study base. This is called *selection bias*. It should be clear from this chapter that case-control studies will only yield unbiased estimates of

$$\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

if the selection probabilities for both cases and controls do not vary between exposure groups. Selection bias occurs when this is not true.

A study can only be truly convincing in this respect if its base is closely defined. The type of study with the best defined base is a nested case-control study, in which the study base consists of a documented and closely traced cohort. This method has proved particularly useful in occupational studies, where employment records identify an underlying cohort and pension schemes provide a mechanism for long term follow-up.

In a *geographically* based case-control study the base is defined by residence in a particular geographical area during the period of study. Although all such individuals are not specifically identified, it may nevertheless be possible to carry out a study in such a way that all cases are registered and controls drawn in a manner unrelated to exposure. Such studies require complete registration of disease in the study area, including capture of resident cases diagnosed and treated elsewhere. Control selection may also be difficult, since few countries have accurate and accessible population registers.

Another important base for case-control studies is the patient list of the family doctor. These lists offer good possibilities for representative control selection and for complete registration of cases particularly when, as in the United Kingdom, access to all medical services is channelled through the family practitioner.

For reasons of economy and convenience, a common choice is the *hospital-based* case-control study in which the case series is made up of all new cases presenting at one or more hospitals during the period of the study. Here the study base consists of the *catchment population* comprising all those persons who would have attended these hospitals if they had developed disease during this period. This is ill defined and it is difficult to demonstrate convincingly that the probability of control selection from the study base is independent of exposure. The device of using other patients, attending for unrelated conditions, has two clear difficulties:

1. catchment populations for different specialities in the same hospital do not necessarily coincide, and
2. patients who are sick with other diseases are not necessarily representative of the population of persons free of the disease of interest. In particular, factors associated with increased risk of these diseases may appear to be *protective* against the disease of interest simply because they are over represented in controls.

Against these difficulties must be set the claim that recall bias and other forms of differential exposure misclassification may be reduced when both case and control groups are hospital patients.

Two further points should be made briefly before concluding this section. First, matching is extremely useful in avoiding selection bias although

its use is more frequently advocated on the grounds of efficiency. We shall return to this discussion in Chapter 18. Second, it is important to draw attention to the fact that the best sampling scheme can be invalidated by poor subject compliance. If a substantial number of potential cases and controls refuse to participate there is considerable potential for bias as a result of differential compliance in different exposure groups. All too often case-control studies do not report compliance, and the potential for such bias is hard to assess.

16.8 Prevalent cases

If a case-control study is carried out using prevalent cases it is no longer a study of disease incidence and the odds ratio estimate cannot be interpreted as an estimate of a ratio of incidence rates. However, such studies can be used to study relationships of exposures to the prevalence of disease.

If the cases can be considered a random sample of those with disease in the population, and controls can be considered a random sample of the healthy section of the population, then the odds that a case was exposed divided by the odds that a control was exposed is an estimate of

$$\frac{\text{Prevalence odds in exposed population}}{\text{Prevalence odds in unexposed population}}.$$

When the prevalence in both groups is low this ratio is approximately equal to the prevalence in the exposed population divided by the prevalence in the unexposed population.

The remarks concerning sources of bias in incident case-control studies apply equally here. In particular, recall bias is a serious problem when interviewing prevalent cases who have been sick and in contact with medical professionals for some time. However, the main problems of interpretation are those of interpreting prevalence itself; the odds ratio is affected by factors which influence the *duration* for which a case, once diagnosed, remains in the sampling frame. These include not only factors related to survival, but factors relating to migration which may be complex and difficult to quantify.

Solutions to the exercises

16.1 The estimate of the odds *for* vaccination in leprosy cases is $101/159 = 0.635$ as compared with $46\,028/34\,594 = 1.331$ in the healthy subjects. The odds ratio estimate is $0.635/1.331 = 0.48$.

16.2 The odds ratio is

$$\frac{101/159}{554/446} = 0.51.$$

16.3 The case/control ratios are as follows:

Alcohol	Tobacco (cigs. per day)			
(oz/day)	0	1–19	20–39	40+
0	0.26	0.42	0.36	1.12
0.1–0.3	0.26	0.46	0.83	0.84
0.4–1.5	0.33	1.13	1.22	1.93
1.6 +	0.63	1.05	2.40	3.37

Because the frequencies in the table are small, there is much random variation, but there is an overall tendency for the ratios to increase both from left to right along rows, and from top to bottom down columns. This indicates that *both* variables have an effect on cancer incidence; there is an effect of tobacco when alcohol intake is held constant, and vice versa.

16 Case-control studies

"In a cohort study, the relationship between exposure and disease incidence is investigated by following the entire cohort and measuring the rate of occurrence of new cases in the different exposure groups." [first paragraph]

As per Miettinen 1985, a *cohort* is

> a closed population (from Latin, *cohors*, enclosure); that is, a population whose membership is defined on the basis of some *event*, for ever after; that is, a population with fixed membership.

It is better not to think of different exposure *groups* but of the amount of exposed and unexposed *population-time*. After all, it might be that exposure was intermittent (and effects near-immediate), such as driving while on or off the cell phone (and motor vehicle accident rates within these amounts of experience) or on or off a blood thinner (and rates of bleeding within these). It would of course be different if the concern was accident rates in the 'under the influence of alcohol' and 'not under the influence of alcohol' states, or rates of developing cirrhosis of the liver over the time of follow-up.

The same methods that C&H applied to the bus drivers and 'conductors' *cohort* studied in Chapter 13 to 15 apply also to a *dynamic population*

Again, as per Miettinen 1985, a *dynamic population* is

> an open population; that is, a population whose membership is defined on the basis of some *state*, for the duration of that state; that is, a population with turnover of membership.

An example is the *open population* of McGill biostatistics students, or of Montreal (you are in iit for as long as, and only as long as, you are a student, or a resident of Montreal). Contrast this with the (*closed population*) *cohort* of McGill biostatistics students, which you entered when you were accepted and first registered, but will always be a part of, even if you never graduate, and even after you die (Einstein is still in the cohort of Nobel laureates, which he entered in 1921).

> and a group of healthy subjects (the controls) is used to represent the subjects who do not develop the disease.

This is old-fashioned definition; the newer outlook will be commented on below. As JH says in the Resources website, '*In case-control studies, (ratios of) rates are based on estimated denominators,* and ***the purpose of the 'controls' is to supply these estimated denominators***. That is why it would be a lot clearer, and mathematically simpler, if the cases were referred to as the '*case series*' and the so-called 'controls' as the '*denominator series.*' Also, it would be a lot easier and simpler if (as C&H seem to do in Table 16.2) one took a sample of the 'base' without regard to whether the persons (or person moments) in the denominator series might (slightly) overlap with the persons in the case series. It is clear from the shots on goal example in the Patrick Roy story (see link 'examples of denominator (control) series') that if one were forced to sample from *all* of the shots on goal, one would do so without regard to whether the shot resulted in a goal. All one wishes to do with the denominator series is classify the individual instances into those in the index category of 'exposure' (here, the low shots) vs. the reference category of 'exposure' (here, presumably the higher shots), *so that we have [albeit estimated] denominators to put under the numerators, so that we can compare rates.* The reluctance to modernize our views is addressed in Miettinen's 2012 interview (at about minute 15 in part 2 of the audio version), which can be found in the Web material, using the link '11.05 Woolf55 / Mantel73 / Miettinen76 Miettinen2012Interview'.

16.1 The probability model in the study base

> Every case-control study of incidence can be seen within the context of an underlying cohort which supplies the cases on which the case-control study depends. A useful terminology refers to this underlying cohort, observed for the duration of the study, as the study *base*.

This is a very important point, but JH would widen it, to include not just an underlying *cohort* but also an underlying *dynamic population.* Think of studying the rates of motor vehicle accidents in the 'on the cell-phone' and 'off the cell-phone' driver time. The *base* from which the accidents 'emerge' (in which the accidents occur) is the entire driver experience.

> Consider the simple situation where the study base is divided into two groups, unexposed and exposed, and let π_0, π_1 be the probabilities that a member of the unexposed or the exposed group will fail over the period of the study and become a case.

Again, JH would not limit it to unexposed or the exposed *groups* but would also consider 'exposed person-time' and 'unexposed person-time' in the study base. He would also like to widen the terminology so that instead of 'fail,'

one speaks of an 'event.' It could be a desirable one: Boston fans wouldn't see scoring a goal on Montreal's Patrick Roy as a failure. The probabilities π_0 and π_1 might be of the order of 1 goal in 15 or so shots, on average.

This use of probabilities, or proportions, π_0 and π_1, brings up another point, to do with events. It is more common to think of *probabilities per unit time*, (as C&H do in Chapter 5) i.e., rates, than probabilities per se. C&H move to rates in section 16.5.

> Figure 16.1 The probability model in the study base

C&H deliberately chose the sampling probability applied to cases to be less than 1 (they chose 0.97), to make the point that it does not affect the rate ratio per se, but rather the precision with which it is estimated. Naturally, if 'cases' are common, and the budget is limited, one might not study all of them.

C&H chose the sampling probability to be applied to the base (or to the 'non-cases') to be 0.01. Again, the sampling fraction does not affect the estimated rate ratio per se, but rather its precision. Naturally, if 'non-cases' are common, and the budget is limited, one might not study all of them; indeed, as we will see when examining the variance (Ch 17), sampling can save a lot of resources while still getting close to the precision that comes from studying the entire material available.

The important point is that the sampling of the F's is carried ***without regard to E***; likewise for the sampling of the S's.

We will also see that *even if one does not know the two sampling fractions* (here 0.97 and 0.01), they cancel out in the ultimate calculations of the rate ratio estimate (or just affect the intercept in the logistic regression).

16.2 The retrospective probability model

> In this argument we re-express our model as a model for the conditional probabilities of exposure given that the subject was a case (F) or a control (S). ... We define the parameter Ω_1 as the odds of having been exposed for a case.

It would be good to get away from this older thinking, as it perpetuates the *temptation to compare cases with controls*, instead of *comparing incidence rates* in the exposed and unexposed persons or person-time. Moreover, it would get us away from the awkward parameter called the 'odds ratio,' and help us ***focus on the parameter of interest***, the ***rate*** ratio.

> Finally the odds ratio can be estimated by
>
> $$\frac{D_1/D_0}{H_1/H_0}.$$
>
> Thus although it is not possible to estimate π_1 and π_0 separately from a case-control study it is possible to estimate the odds ratio.

Even though we should move on from comparing exposure rates (and instead compare rates), the format of this *rate*-ratio estimator is of note, since it does point to the *statistical models* we will invoke in order to study its sampling distribution. It is good to think of the $D_1 : D_0$ as the ***binomial*** split of the entire case series, of size $D = D_1 + D_0$. And it is good to think of the $H_1 : H_0$ as the ***binomial*** split of the entire denominator ('control') series, of size $H = H_1 + H_0$. Better still, if the H series was sampled independently of (without worrying about any possible overlap with) the case series, then the sampling distribution of this rate-ratio estimator becomes even more tractable: we can either model the 2 random variables as 2 independent binomials, or condition on all 4 margins of the 2×2 table, and treat the one free entry as a non-central hypergeometric distribution modulated by the rate-ratio parameter θ. And if we cannot achieve a Gaussian-looking likelihood, we can use the likelihood from the exact non-central hypergeometric distribution.

EXAMPLE: BCG VACCINATION AND LEPROSY

This is an excellent teaching example (JH had put the Fine et al. 1986 Lancet paper (Protective efficacy of BCG against leprosy in Northern Malawi) on the Resources website. It has the same '***known denominators***' structure as John Snow's study of cholera deaths in the two South London water companies in 1854 – see 'examples of denominator (control) series'. Snow relied on a 'Government Return' (Survey) from the year before to provide the 'denominators' (see Resources website for more details). Both examples let us simulate various 'what if these demominators were not available?' scenarios and sampling approaches.

> Here a population survey was available for control but **had it not been** there would have been no need to carry out such a large-scale exercise. The precision of the odds ratio estimate [JH: a *rate ratio estimate*!] is dominated by the precision of the odds for BCG scar among the 260 leprosy cases. Perhaps 1000 suitably chosen controls [JH: a *denominator series* of 1000!] would be enough to estimate the corresponding odds [JH: prevalence of a BCH scar] among healthy[1]

[1] It is not quire clear from the wording whether C&H suggest sampling from *just* the

subjects - there is little gain in precision to be obtained by using 80 000!

We (and C&H) investigate this in Chapter 17.

SYNTHETIC RETROSPECTIVE STUDY

Mantel used 'case-control' sampling in 1973 (see full text on website under Mantel73). His reason for sampling was to cut down on computer time when running the logistic regression.

> THE SYNTHETIC RETROSPECTIVE STUDY - SAMPLING FROM A PROSPECTIVE STUDY
>
> My present purpose is to propose, discuss, and validate use of retrospective approach procedures in a prospective study situation. A particular prospective-study situation which I encountered gave rise to only 165 cases of a particular condition in a cohort of about 4,000 individuals.
>
> Preliminary analyses were undertaken using a limited number of the variables on which data had been collected. But even with simple maximum likelihood analyses of the form used involving only one or two of the study variables, the computer time required was somewhat prolonged. This could then preclude making analyses as comprehensive and as extensive as we should have liked.
>
> As it turned out, the key cause for prolonged computer time was the large number of observations involved. Computation was simple and rapid once the necessary totals were obtained for all 4,000 individuals. But time was consumed for entering all the information and for computing at each iterative stage certain quantities appropriate for each individual. The number of iterative cycles for convergence could be reduced by a device for obtaining suitable entering approximations (see below), but even this would not resolve our problem.
>
> A possible remedy envisaged was to convert the study, in principle, to a retrospective one. Suppose we included in the analysis a random proportion, π_1 of our cases and another random proportion, π_2, of the negatives. If we chose π_1 as 1 and π_2 as 0.15, we would have all the cases and 3.5 negatives per case. By the reasoning that $n_1 n_2/(n_1 + n_2)$ measures the relative information in a comparison of two averages based on sample sizes of n_1 and n_2 respectively, we might expect by analogy, which would of course not be exact in the present case, that this approach would result in only a moderate loss of information (***The practicing statistician is generally aware of this kind of thing. There is little to be gained by letting the size of the control group,*** n_2***, become arbitrarily large if the size of the experimental group,*** n_1***, must remain fixed.***) But the reduction in computer time would permit much more effective analyses. Ostensibly we would be meeting the additional conditions assumed for validity of the retrospective study approach; that is the retained individuals would be a random sample of the cases and disease-free individuals arising in the prospective study.

(80,622 - 260 = 80,362), or from *all* 80,622. Sampling from *all* 80,622, without worrying about a possible small overlap with cases, makes it easier to calculate the sampling variation, avoids having to talk about odds ratios, and allows one to explain the rate ratio estimate to grade 6 students. In their answer to exercise 17.1 (p173), it appears that they sample from *all* 80,622.

16.3 The prospective probability model

Be careful here as to the meaning of (lower case) omega:

$$\omega_1 = \frac{Prob[F|E+]}{1 - Prob[F|E+]},$$

and so it has the opposite directionality from the upper case OMEGA in section 16.2:

$$\Omega_1 = \frac{Prob[E+|F]}{1 - Prob[E+|F]}.$$

The final result in section 16.3 is that since (***algebraically***)

$$\frac{\widehat{\omega_1}}{\omega_0} = \frac{D_1/H_1}{D_0/H_0} = \frac{D_1\ H_0}{D_0\ H_1} = \frac{D_1/D_0}{H_1/H_0} = \frac{\widehat{\Omega_1}}{\Omega_0},$$

this, together with the stipulations in section 16.5, is taken as justification for the statement (p161) that the *crossproduct* ratio $(D_1\ H_0)/(D_0\ H_1)$ in a case-control study may be used to estimate the rate ratio in the underlying study base.

Note that JH prefers to use the term empirical '***crossproduct***' ratio rather than empirical 'odds' ratio so as to keep our ***focus on the real estimand (the rate ratio)***, and so as not to introduce an parameter (odds ratio) with *no scientific meaning*.

> In the retrospective argument it is the exposure status which is the response (outcome variable); in the prospective argument it is the disease status which is the response. The retrospective argument is more natural, ..

Treating the exposure status as the random variable (actually there are 2 Bernoulli series: the Bernoulli for each case in the 'case series', and the Bernoulli for each 'exposure probe' in the 'denominator series') is merely a *statistical* – and *artificial* – step in the point and interval estimation of the Rate Ratio.

16.4 Many levels of exposure

> ... we shall look at a famous study carried out in the middle of the nineteenth century by William Guy (Guy, W.A. Journal of the Royal Statistical Society, 6, 197-211.) This was ***possibly the first case-control study***.

Case-control 'thinking', and the retrospective vantage, is fundamental to assessing causality. It is what we do when we get a headache, or a rash, or when our computer keeps crashing. Many of the religious and other proscriptions agains eating certain foods (e.g., pork, shellfish, the blowfish that featured in the Homer Simpson episode, ...) probably rest on evidence (long since forgotten) gained from case-control thinking. Indeed, JH invites scholars to search biblical sources to see if we can uncover some of them.

The study by Guy predates the 20th-century ones described in the '***Origins and early development of the case-control study*: part 1, Early evolution' article (on the Resources website).

The whimsical one by Thornton Wilder (see below) is a fun read.

THORNTON WILDER'S ORIGINAL DESIGN OF A CASE-CONTROL STUDY

To my knowledge, results of the first case-control study (of reproductive factors and breast cancer) were reported by Lane- Claypon in 1926 (1). The original copyright date of The Bridge of San Luis Rey (2) by Thornton Wilder is 1927. Since Lane-Claypon's study and Wilder's book were conceived at about the same time, and on different sides of the Atlantic, it is safe to assume that the work of neither person influenced 3 that of the other. Therefore, does not the following passage from Wilder's book, although fictional, constitute an independent, original description of the case-control method of epidemiologic enquiry?

> It was by dint of hearing a great many such sneers at faith that Brother Juniper became convinced that the world's time had come for proof, tabulated proof, of the conviction that was so bright and exciting within him. ***When the pestilence visited his dear village of Puerto and carried off a large number of peasants he secretly drew up a diagram of the characteristics of fifteen victims and fifteen survivors, the statistics of their value*** "sub specie aeternitatis". Each soul was rated upon a basis of ten as regards its goodness, its diligence in religious observation, and its importance to its family group. Here is a ***fragment*** of this ambitious chart:

	Goodness	Piety	Usefulness
Alfonso G.	4	4	10
Nina	2	5	10
Manuel B.	10	10	0
Alfonso V.	-8	-10	10
Vera N.	0	10	10

> The thing was more difficult than he had foreseen. Almost every soul in a difficult frontier community turned out to be indispensable economically, and the third column was all but useless. The examiner was driven to the use of minus terms when he confronted the personal character of Alfonso V., who was not, like Vera N., merely bad he was a propagandist for badness and not merely avoided church but led others to avoid it. Vera N. was indeed bad, but she was a model worshipper and the mainstay of a full hut. ***From all this saddening data Brother Juniper contrived an index for each peasant. He added up the total for the victims and compared it with***

> ***the total for survivors to discover that the dead were five times more worth saving.***
>
> From The Bridge of San Luis Rey by Thornton Wilder. Copyright 1927 by Albert Boni, Inc. Copyright renewed 1955 by Thornton Wilder. Reprinted by permission of Harper & Row, Publishers, Inc.

This quote is taken out of context, and is not meant to convey or imply anything about Wilder's fine and moving book. It is presented as a missed contribution to epidemiology from an unlikely source.

References

1. Lane-Claypon JE. A further report on cancer of the breast. Reports on Public Health and Medical Subjects 32. London: Her Majesty's Stationery Office, 1926. 2. Wilder T. The Bridge of San Luis Rey. New York: Washington Square Press, 1939 and 1968.

David B. Thomas, Program in Epidemiology,Fred Hutchinson Cancer Research Center, Seattle, WA 98104

Letter to Editor, American Journal of Epidemiology, Vol 124(2) pp 342-343, 1986

16.5 Incidence density sampling

> All of these assumptions can be guaranteed by the simple device of selecting a short enough study period. If insufficient cases would be obtained from such a study then the remedy is simple - carry out ***several consecutive short studies***. The subjects remaining in the base at the end of one study immediately enter the next study. Each study then provides a separate estimate of the rate ratio, and provided this ratio remains constant over the whole study period, the information can be aggregated using methods very similar to those discussed in Chapter 15,

This is similar in spirit to what Mantel did (see above). He tells us

> The actual number of individuals was substantially less than 4,000. An initial cohort of about 1,350 men was studied to evaluate the short-term prognostic value of various factors in coronary heart disease. **Men remaining free of disease for two years could be re-entered into the analysis for the next two years using their new X_i values.**

C&H go on to say....

> Taken to the limit, the total time available for the study may be divided into ***clicks which contain at most one case***. Those clicks in which no case occurs are not informative so there is no purpose in drawing controls, but ***controls are drawn for all clicks in which a case occurs***. Thus one or more controls are drawn from the study base immediately after the occurrence of each case. *This design is termed incidence density sampling.* A study carried out in this way involves matching of controls to cases with respect to time. Methods for stratified case-control studies will be discussed in Chapter 18, but in the special case where the ratio of exposed to unexposed persons in the study base does not vary appreciably over the study period, it is legitimate to ignore the matching by time during the analysis.

The reason '**this design is termed incidence density sampling**' probably goes back to Miettinen's 1976 paper, where, in getting rid of the 'rare disease assumption' used in page 161, he also introduced the term 'incidence density':

> 3.1. The parameters. ***Incidence density*** ("force of morbidity" or "force of mortality") – perhaps the most fundamental measure of the occurrence of illness – is ***the number of new cases divided by the population- time (person-years of observation) in which they occur***.

Miettinen (at minute 20:50 in part 2 of the audio of the 2012 interview) tells us that he not adopted the term 'incidence density *sampling*' at all.

> The whole point is, when your study base is population-time, a dynamic population-time, then you sample that for your denominator series. There is nothing special about the *sampling* itself. If there is anything special, it is the nature of the study base, namely a dynamic population. So that is what you are sampling. It is not that you have a peculiar way of sampling; you sample what you must sample. It is not a sampling option.

[JH: note cases '*emerging*'].

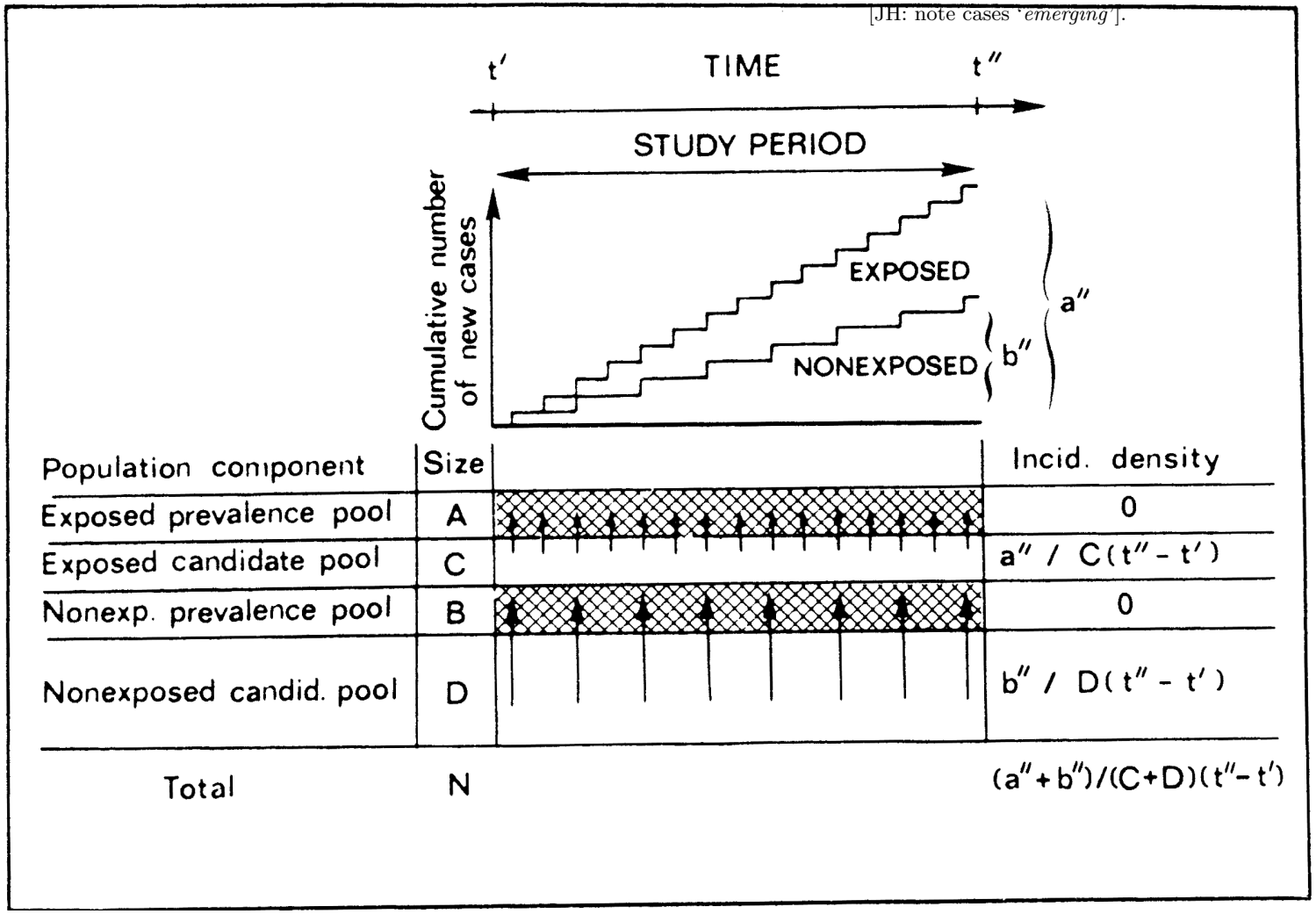


Figure 1. Static population (e.g. a particular age group) over the time-span (t' to t'') of a case-referent study based on incident cases. The sizes of the different component populations remain static, but there is turnover of membership in each compartment. The arrows indicate occurrences of new cases, i.e., transitions from the candidate pools to the prevalence pools. Note that the incidence-densities are zero in each of the prevalence pools, and that the incident cases are referred to the follow-up experiences in the candidate pools only. Also note that **the incidence density ratio is** $(\mathbf{a}''/\mathbf{b}'')/(\mathbf{C}/\mathbf{D})$, with $\mathbf{a}''/\mathbf{b}''$ and $\mathbf{C}/\mathbf{D}$is estimable from the (incident) cases and referents, respectively, regardless of the levels of incidence or prevalence.

TABLE 1

Case-referent data by Cole et al. (4) and Cole (7) relating the incidence of bladder cancer to cigarette-smoking in men of various ages. The study involved complete ascertainment of newly-diagnosed cases in a population (eastern Massachusetts) of known size by age. Interviews were confined to a sample of cases as well as of non-cases. See section 6.

Age (years)	No. of new cases within 18 months	Size of source population in 10^3	Overall incidence density* in $(10^5 \text{ years})^{-1}$		No. of study subjects				Incidence density ratio	Smoking-related fraction of cases		Exposure-specific incidence density in $(10^5 \text{ years})^{-1}$	
					Cases		Referents						
			Study	Connecticut region (8)	Sm.+†	Sm.−	Sm.+	Sm.−		Etiologic	Preventive	Sm. +	Sm. −
50–54	35	77.4	30	(29)	24	1	22	4	4.36	.74		34	8
55–59	52	68.4	51	(48)	35	2	35	4	2.00	.47		54	27
60–64	52	61.5	56	(65)	31	5	38	3	.49		.48	52	110
65–69	86	47.4	120	(130)	46	7	42	15	2.35	.50		140	60
70–74	105	38.0	180	(170)	60	13	51	28	2.53	.50		230	90
75–79	76	23.2	220	(200)	39	14	32	20	1.74	.31		260	150

30-year risk at age 50 years given survival from other illness:

Smokers: $\hat{R}_{50,\,80} \doteq [(34 + 54 + 52 + 140 + 230 + 260)/10^5 \text{ years}]\,(5 \text{ years}) = 0.0385 = 3.9$ per cent.

Nonsmokers: $\hat{R}_{50,\,80} \doteq [(8 + 27 + 110 + 60 + 90 + 150)/10^5 \text{ years}]\,(5 \text{ years}) = 0.0223 = 2.2$ per cent.

Risk ratio estimate: $\hat{RR}_{50,\,80} = 3.85/2.23 = 1.7$

Risk difference estimate: $\hat{RD}_{50,\,80} = (3.85 - 2.23)$ per cent $= 1.6$ per cent.

* Two-digit accuracy.

† Sm+ and Sm−: smoker and nonsmoker, respectively.

[Text from Miettinen's 1976 article]: Cole et al. identified all newly-diagnosed cases of bladder cancer in a (static) population (eastern Massachusetts) of known size over an **18-month period**, drew a **reference series** [or '**denominator series**'] from the **source-population of the cases**, and inquired (inter alia) into the subjects' histories with respect to cigarette smoking. Data are presented in table 1.

Worked e.g. by JH: In the **50-54** age stratum, the total population-time (PT) is $PT = 77.4 \times 1.5 = 116.1$K person-years. The overall incidence density is $35/116.1K = 30$ per 10^5 years. The estimated Sm.+ : Sm- split of the PT is 22:4, so that $\widehat{PT_{SM+}} = \frac{22}{26} \times PT$, and $\widehat{PT_{SM-}} = \frac{4}{26} \times PT$, and so the estimated incidence density ratio (Rate ratio) is

$$\frac{24/PT_{SM+}}{1/PT_{SM-}} = \frac{24/\{\frac{22}{26} \times PT\}}{1/\{\frac{22}{26} \times PT\}} = \frac{24/1}{22/4} = \frac{24 \times 4}{22 \times 1} = \frac{a \times d}{b \times d} = \mathbf{4.36}.$$

16.6 Nested case-control studies

> An important use of incidence density sampling is in nested case-control studies, where case-control analysis is used in cohort studies. This is an attractive option whenever the assessment of exposure of any subject is, for some reason or other, expensive. For example, in dietary studies, individual diet may have been assessed by very detailed diary records of food intake, perhaps referring to several periods of time. The coding and transcription of such records for computer analysis is laborious and expensive. Much of this work is avoided in a nested case-control study by coding these records only for cases, as they occur, and for groups of controls drawn for each case. Since there is (usually) little to be gained by drawing more than five controls for each case, there are considerable savings to be made by such a strategy. We shall discuss the design and analysis of nested case-control studies in Chapter 33.

A local (McGill) example, one of the earliest of this kind, is the case-control study of asbestos and lung cancer, carried out within the Quebec cohort of over 10,000 miners and millers of asbestos. The expense involved obtaining detailed work and smoking histories. The 'nested case-control' approach is described by F. D. K. Liddell; J. C. McDonald; D. C. Thomas; and Stella V. Cunliffe in the article "Methods of Cohort Analysis: Appraisal by Application to Asbestos Mining' in the Journal of the Royal Statistical Society. Series A (General), Vol. 140, No.4 (1977), 469-491. See Resources on Website for the full text of that 1977 JRSS paper.

The following is an excerpt from their report:

> The approach was evaluated by considering five controls for each of the 215 lung cancer deaths. The selection of controls was strictly at random from among men born in the same year as the case and known to have survived at least into the year following that in which the case died. That there were only 1,290 men in this study made it possible to re-examine all smoking history questionnaires which failed validity checks or otherwise aroused doubts. As a result, codes were changed for 122 men (nearly 10 per cent of 1,290), although altering the classification for only 39 men (3 per cent). However, the opportunity was taken to reclassify those who had given up smoking, according to the report, into those who had been ex-smokers for at least seven years when the case died, and recent smokers.

FURTHER COMMENTS by JH

Theoretical basis for "odds ratio" as estimator of Rate Ratio, together with statistical model for the estimator

The old-fashioned and very loose justification for using the empirical odds ratio, or[2], as an estimator of the theoretical rate ratio goes back to Cornfield in the 1950s. Unfortunately it still is the one given in many 'modern' texts, despite the much more general justification provided by Miettinen in 1976.

The old justification rested on algebraic arguments using *persons*, not *population time*. The outcome *proportions* involved refer to *cumulative* incidence.

Clayton and Hills also start with proportions (risks), and thus are forced to use the now-very-old-fashioned "rare disease assumption". Later they argue that if we slice time very finely, we do not need this assumption, but it would be nice if they started with this modern way of viewing it.[3]

One way to be modern, and emphasize that we are generally in the 'rate' (rather than than 'risk') business is to use the empirical odds ratio as an estimate of the rate (i.e., intensity) ratio of interest, and to *call it a rate ratio estimate*; in other words, even if we derive the estimate from a crossproduct that *looks like* an odds ratio, or from the output of a logistic regression where the estimate is labelled 'odds ratio' (or the coefficient is labelled 'log odds ratio'), it is estimating a rate ratio. The estimand (the parameter to be

[2]Sometimes JH will use the lower case *or* to denote the observed or empirical odds ratio, and upper case OR to denote the true (but unobservable) odds ratio.

[3]There are a few instances where time *per se* isn't involved, e.g., success rates (with 'rate' as a proportion) of low versus high shots on ice hockey goalie Patrick Roy, acceptance rates of males versus females to medical school, etc. In these instances, it is still helpful to think of the 'successes' (or events) as numerators (classified as low/high, or male/female), and *samples* of all shots or applications, also classified in same way, as estimates of the relative sizes of the respective denominators. Note that if we sample from all candidates (shots or applicants) without excluding those that happen to form the numerator series, we *can directly estimate the ratio of the two proportions without ever involving anything that looks like an odds ratio.* To see this, imagine that candidates in the index and reference categories of the determinant in question would be expected to produce π_1 and π_0 'success' proportions respectively respectively, so that the estimand, the ratio of the two proportions, is π_1/π_0. Suppose we get to observe the two numerators y_1 and y_0, but not the full sizes N_1 and N_0 – the two denominators – of the two candidate pools. We can estimate the relative sizes by sampling say n from the $N_1 + N_0$ and classifying them into n_1 and n_0 respectively. If we know the sampling fraction f, then we can estimate the two 'success' proportions as $\hat{\pi}_i = y_i/[n_i \times (1/f)]$ and estimate their difference or their ratio. However, in the ratio, the sampling fraction cancels out, leaving us with the estimator $\widehat{\pi_1 \div \pi_0} = (y_1/n_1) \div (y_0/n_0)$ that does not require knowledge of the sampling fraction. In the statistical model in this context, y_1 and y_0 are two binomials with unknown denominators, N_1 and N_0, while the n_1/n split is a hypergeometric random variable (that could be approximated by a binomial if the sampling fraction is small. Theoretically, or via simulation, one can show that $(y_1/n_1) \div (y_0/n_0)$ is median-unbiased for $\pi_1 \div \pi_0$.

estimated) is $\theta = \lambda_1/\lambda_0$.

The truly modern way is to think of the cases as arising in population-time, and to think of the population time involved as an infinite number of person-moments - think of a person-moment as a person at a particular moment. Say that a proportion π_E of these are "exposed" person moments, and the remaining proportion π_0 are "non-exposed" person-moments. Suppose further that the (theoretical) event rates in the exposed and unexposed amounts of population-time are

$$\lambda_E = \frac{E[no.events]}{PT_E} \; ; \; \lambda_0 = \frac{E[no.events]}{PT_0},$$

with (theoretical) Rate Ratio $\theta = \lambda_E/\lambda_0$.

Denominator Series *[overall size d; d_0, d_1 in 'exposure' categories 0, 1]*

Suppose we take a finite random sample, of size d,[4] of the infinite number of person moments in the base that generated the cases, and classify them into d_E "exposed" person moments and $d_0 = d - d_E$ "non-exposed" person-moments. We will refer to this sample of d as the *denominator* series. What is the statistical model for $d_E \mid d$? Clearly, it is

$$d_E \sim Binomial(d, \pi_E).$$

Numerator (Case) Series *[overall size c; c_0, c_1 in 'exposure' categories 0, 1]*

Denote by c the observed number of events; we classify them into c_E events in "exposed" population-time and $c_0 = c - c_E$ in the "non-exposed" population-time. We will refer to this sample of c as the *case* series.

What is the statistical model for $c_E \mid c$? We can think of c_E as the realization of a Poisson r.v. with mean (expectation) $\mu_E = (PT_E \times \pi_E) \times \lambda_E$. Likewise, think of for c_0 as the realization of a Poisson r.v. with mean (expectation) $\mu_0 = (PT_0 \times \pi_0) \times \lambda_0$.

Now, it is a statistical theorem (Casella and Berger, p194, exercise 4.15) that

$$c_E \mid c \sim Binomial(c, \mu_E/[\mu_E + \mu_0]).$$

Thus we can identify the distribution of the 4 random variables involved in the OR estimator

$$\hat{OR} = or = c_E/d_E \; \div \; c_0/d_0 \; = \; c_E/c_0 \; \div \; d_E/d_0 = (c_E \times d_0) \; \div \; (c_0 \times d_E).$$

[4] In this section, JH has borrowed from Miettinen the notation of 'c_1' and 'c_0' for the numbers of exposed and unexposed cases, and 'd_1' and 'd_0' for the numbers of exposed and unexposed denominators, instead of C&H's D_1 and D_0, and Y_1 and Y_0. Both of these sets of notations are more memorable and instructive than the A, B, C, and D in Mantel and Haenszel's 1959 paper, and the a, b, c, d also widely used elsewhere is statistics.

The $c_E : c_0$ split is governed by **one binomial**, *involving* θ and other parameters, while the $d_E : d_0$ split is governed by **a separate binomial**, involving the same other parameters, but *not* involving θ.

If one replaces μ_E and μ_0 by their constituents, one can show that the odds that an unexposed person-moment in the series of $c + d$ represents a "case" is $c : d$, whereas the corresponding odds for an exposed person moment is $(\theta \times c) \; : \; d$.

In other words, in the dataset of $c + d$,

$$logit[Prob[case|0] = \log(c/d) = \beta_0 \; ;$$

$$logit[Prob[case|E] = \log(c/d) + \log\theta = \beta_0 + \beta_E E,$$

where E is an indicator variable.

So, one can estimate $\log\theta = \log OR$ by an unconditional logistic regression of the Y's ($c + d$ observations in all, with $Y = 1$ if in case series; $= 0$ if in denominator series) on the corresponding set of $c + d$ indicators of exposure (1 if exposed, 0 if not). If need be (if e.g., there are matched sets), one can use conditional logistic regression.

C & H preamble

"If there is no relationship between exposure and disease incidence the distribution of exposure among the cases should be the same as the distribution among the controls."

We often use this reasoning informally, and supply a 'denominator series' from our own experience, as when we are confronted with the statement that 90% of all driving accidents occur within 10 Kilometres of home, or that in most pancreas cancer cases in North America the patient gives a history of coffee-drinking, or that most goals in ice hockey are on low shots, or that those in professional sports tended to be born in certain months of the year. We do not formally call this experience the 'control series'; rather it is more aptly referred to as a 'denominator' series.

Many of todays epidemiologists still regard a case-control study as a comparison of the exposure frequency (or exposure odds) in the case group with that in the control group. It is time to take the modern view. It says that 'in better families, one never compares cases vs. controls. Even though the exposure data were collected after the fact, 'epidemiologists are students of (event) rates' and so we compare (the event rate in) the exposed experience with (that in) the unexposed experience, even if it means having to estimate the (relative) magnitudes of the denominators involved in these event rates.

Supplementary Exercise 16.1

You are asked in conduct a study on the possible role of MMR vaccination on the etiology of autism (cf. Danish Study). Suppose that 315 cases of autism were diagnosed in the dynamic population,[5] 1991-1998, followed to December 1999 ("the base"). The base can be split into 36 age-year 'cells': each of the 28 full-year cells (see diagram in notes for CH15. Each square contains 67,000 infant-years of experience, and each of the other 8 contains half this amount.

Unfortunately, there is no electronic vaccination registry that documents this base. The vaccination records are maintained by regional heath authorities in card files, one per family. They can determine if a child was/was not vaccinated before a specified age/date. The Agency can extract this information and deliver you a .csv file. However, because this is a tedious process, it costs \$10/probe (each 'probe' returns a 'yes/no' if at the child-moment in question, the child had/had not already been vaccinated. Furthermore, because of confidentiality concerns, and because different abstractors work independently on the case series and the base series, in the file delivered to you, you cannot check any overlap in the person-moments in the base-series portion of the file with those in the case-series portion, or with probes into other cells.

For this exercise, focus on just 1 square, the 1993 birth cohort, in the children time between their 3rd and 4th birthdays. This child time consisted of some Y_1=60,143 vaccinated child-years, and Y_0=6,857 unvaccinated child-years.

Count the numbers of cases in these vaccinated and unvaccinated child-years.

First, without using the Y_1 and Y_0, simulate a denominator series by sticking several (say 10) pins at random into this square and counting what proportion of them land on the vaccinated and unvaccinated person moments. [That's one of the ways the ancient mathematicians estimated the value of π – they drew a circle inside a square, and counted the proportion of random shots that fell inside the circle.]

Second, using the Y_1 and Y_0, simulate a larger denominator series (say 100) using the binomial random number generator in R.

Supplementary Exercise 16.2

The article by Woolf (1955) does not have an abstract. As a way to summarize the contents and messages of the article, write a structured abstract, of 200 words or fewer, paraphrasing the article. Devote a good portion of your abstract to the Background/Rationale section, summarizing the first two paragraphs of his paper.[6]

[5] See (available on course site) the expository IJE articles on this topic by Vandenbroucke.

[6] As many of todays epidemiologists still do, does he regard a case-control study as a comparison of the exposure frequency (or exposure odds) in the case group with that in the control group? Or, does he take the modern view? It holds that in better families, one never compares cases vs. controls. Even though the exposure data were collected after the fact, epidemiologists are students of (event) rates and so we compare (the event rate in) the exposed experience with (that in) the unexposed experience, even if it means having to estimate the (relative) magnitudes of the denominators involved in these event rates.

Supplementary Exercise 16.3

In an investigation of the role of blood group in the etiology of ovarian cancer, would a denominator-series consisting of males be acceptable? Must those in the denominator-series be at non-zero risk of ovarian cancer? Explain.

Supplementary Exercise 16.4

McDonnell Douglas was a major American aerospace manufacturing corporation and defense contractor formed by the merger of McDonnell Aircraft and the Douglas Aircraft Company in 1967. Between then and its own merger with Boeing thirty years later, it produced a number of well-known commercial and military aircraft such as the DC-10 airliner and F-15 Eagle air-superiority fighter.

One summer during that timepsan, there was greater than usual number of airline crashes, several of them involving planes manufactured by McDonnell Douglas. Photographs of the crashed planes were prominent in news stories and cover pages of magazines such as TIME and Newsweek.

This unwelcome publicity prompted a letter from the CEO and Chairman of the company to one of the magazines, with the following text (paraphrased from memory, since JH is unable to locate the letter itself):

> Recent articles in your magazine have included stories and photographs in which the McDonnell Douglas name was very visible and prominent, suggesting that our products are not as safe as those of our competitors. But this impression is a false one, just as it is false to conclude, from the fact that 90% of car accidents occur within 10 miles of home, that accident rates in the driving involving driving-time further from one's home are lower. It is not fair that your journalists concentrate only on the commercial planes that have crashed onto the ground. If they were to position themselves high up in the sky, and also photograph commercial planes they find up there, they would find that many of these also carry the McDonnell Douglas name. The planes we make are just as safe as those of our competitors.

His point about focusing on the numerators, and ignoring the denominators, is well taken, but at the time, his suggestion to assemble a 'denominator series'

by sampling the 'base' (airplanes in flight) was not all that feasible. However, thanks to this website `http://www.flightradar24.com`, it is possible today. And so we will use it to demonstrate the statistical efficiency of the 'case-control' (or 'case/base') approach in comparing the accident rate in flights made with the different major aircraft manufacturers.

Imagine, in this simplified and *entirely fictional* example (loosely modelled after the New York story 'Pedestrians Fatally Injured by Motor Vehicles' in page 98 in Friedman's Primer of Epidemiology, but – in the interests of feasibility for this class– not matched on time of day), that there had been 7 airline crashes involving planes from 2 of the leading manufacturers A and B and that they occurred at the following (Latitude, Longitude) locations

Instance:	1	2	3	4	5	6	7
Latitude:	50	45	40	35	30	25	20
Longitude:	-170	-120	-70	20	70	120	170
Manufacturer:	A	B	B	B	A	B	B
Denominator Sample	5	5	5	5	5	5	5
Manufacturer:							
A (no.)	..	..	..	..	..	..	..
B (no.)	..	..	..	..	..	..	..

Use the website[7] to obtain a denominator sample of 5 flights at each of these locations, and split each denominator by manufacturer, as per the table above. Calculate a (crude) Rate Ratio estimate $\frac{5/2}{??/??}$ for the B:A contrast.

Supplementary Exercise 16.5

In his book 'Outliers', the Canadian-born author Malcolm Gladwell related the observation made in the early 1980s by Roger and Paula Barnsley, that a disproportionate number of elite Canadian hockey players are born in the first few months of the calendar year.

You are asked to use the online database `http://www.hockeydb.com` to determine if this pattern continues to hold.

Focus on the NHL and limit your attention to Canadian-born players.

Month of Birth:	J	F	M	A	M	J	J	A	S	O	N	D
Name starts with..												
B:	..	..	..	..	..	..	..	..	..	..	..	..
C:	..	..	..	..	..	..	..	..	..	..	..	..
H:	..	..	..	..	..	..	..	..	..	..	..	..
L:	..	..	..	..	..	..	..	..	..	..	..	..
M:	..	..	..	..	..	..	..	..	..	..	..	..
P:	..	..	..	..	..	..	..	..	..	..	..	..
S:	..	..	..	..	..	..	..	..	..	..	..	..

Rather than count by hand, you will probably find it easier to download the file, save it as a .csv file, and process it using the R code given in the Resources for case-control studies.

Comment on the pattern, and how it might have arisen.[8] How might you could test if the pattern is monotonically decreasing?

The annual numbers of live births, by month, in Canada provinces and territories, for the years 1991 to 2007 can also be found on the website. The non-uniformity over months was probably a little more extreme in earlier decades.

Supplementary Exercise 16.6

Refer to the article "Road Trauma in Teenage Male Youth with Childhood Disruptive Behavior Disorders: A Population Based Analysis" by Redelmeier et al. in PLoS Medicine, November 2010, Volume 7, Issue 11, e1000369.

The following is an excerpt from the Abstract:

> A history of disruptive behavior disorders was significantly more frequent among trauma patients than controls (767 of 3,421 versus 664 of 3,812), equal to a one-third increase in the relative risk of road trauma (odds ratio = 1.37, 95% confidence interval 1.22–1.54, p,0.001). The risk was evident over a range of settings and after adjustment for measured confounders (odds ratio 1.38, 95% confidence interval 1.21–1.56, p¡0.001).
>
> The risk explained about one-in-20 crashes, was apparent years before the event, extended to those who died, and persisted among those involved as pedestrians.

In the Methods, the authors state:

[7] To see the planes currently flying near location 50 North, 170West, i.e., (50,-170), use the URL `http://www.flightradar24.com/50,-170/6` The '/6' indicates the zoom level. Increase it to 7 or 8 to see a smaller area, or decrease it to 5 to see a larger area.

[8] After you give up, you can look at this video. Hanley J. How-your-birth-month-affects-your-future – Interview Global TV, Montreal, August 5, 2013. The link is under Reprints/Talks on JH's website. Don't be discouraged: JH didn't figure it out either – he got the explanation from others.

> We excluded teenage girls from both groups to avoid Simpsons paradox (a spurious association created by loading on a null-null position) since this group has much lower rates of crash involvement.

In the Discussion, the authors state:

> A third limitation that causes our study to underestimate the association of disruptive behavior disorders with road trauma is that the data excluded girls [74]. To address this issue we retrieved the original databases, replicated our methods in girls rather than boys, and conducted a post hoc analysis. As anticipated, the results yielded a smaller sample (n = 4,156) and about the same estimated risk (odds ratio 1.31, 95% confidence interval 1.07–1.61, chi- square = 6.8, p = 0.010). Hence, the association of disruptive behavioral disorders with road trauma extended to both teenage boys and girls. Of course, many issues remain for future research including medication level at time of injury, amount of driving, extent of brain trauma, and sequelae among those not hospitalized [75,76].

Questions:

i. Reproduce the (crude) odds ratio and CI reported in the abstract.

ii. Figure out how the authors came up with the statement that "the risk [factor] explained about one-in-20 crashes".

Hint: (i) in what fraction of crashes was the factor present? [note that the factor cannot explain cases among those in in which the factor was absent].

(ii) Even when it was present, the factor wasn't responsible for all of these crashes. Most of them would have happened even in the absence of the factor. Use the reported rate ratio (take the 'adjusted' 1.38 rather than the crude 1.37). Work out, using say 138 accidents in males in whom the factor *was* present, what fraction of them would have occurred because of unrelated background factors and what fraction would be 'excess' cases due to the factor itself (or ask yourself: for every crash among those with the factor that was because of background (unrelated) causes, how many would there because of the factor? The (relative) rates in those with and without the factor are the key here). This latter fraction is called the 'etiologic fraction among the exposed'.

Then multiply this etiologic fraction (of cases among the exposed that are due to the exposure) by the fraction in (i) to get the overall fraction of all cases that is due to the factor.

This overall fraction (a fraction of a fraction) is called the overall or population etiologic fraction (sometimes called the population 'attributable' faction). The two formulae for it are not well understood by epidemiologists. See the article 'A heuristic approach to the formulae for the population attributable fraction' under the 'r e p r i n t s' tab on JH's homepage.

iii. Come up with a reasonably realistic example of a behaviour/trait and the occurrence of some health event, where, in a 'case-control' study, if one simply added the 4 frequencies in the 2×2 table for boys and the corresponding ones for girls, and used the combined frequencies from this overall 2×2 table to produce one odds ratio, one could produce a very different odds ratio than the (say) common odds ratio in each of the gender-specific tables.

Do you think Redelmeier need have been concerned with Simpson's paradox in his context?

iv. The reported odds ratio for girls (1.31) is accompanied by both a confidence interval and a (null) chi-quare statistic and p-value. The CI was probably arrived at using Woolf's formula. Compute a test-based confidence interval instead and comment on how close it comes to the reported interval.

v. From the reported odds ratio of 1.31, and assuming that appendicitis is just about as common in boys and girls, that the 4,156 is the total number of trauma and appendicitis admissions in girls, and stating any other assumptions you are forced to make, try to reconstruct what the 2×2 table must have looked like for girls. (the reported CI should be of considerable help!)

vi. We briefly discussed in class how one could merge(combine) the odds ratios for boys and girls to get a single point estimate and associated CI. [notice that *combining the odds ratios* to get 1 new odds ratio can yield a very different result from *combining the raw frequencies* into one 2×2 table, and making one odds ratio from this one table]. Formally merge the results in 3 ways:

   (a) Using the antilog of the weighted average of the logs of the gender-specific odds ratios (also known as Woolf's method) As part of this exercise, prove that the linear combination Woolf uses is the linear combination with the minimum variance.

   (b) Using the Mantel-Haenszel summary odds ratio, and accompanying your point estimate by a *test-based* CI. [Whereas the M-H point

estimate and test-statistic formulae date from 1959, we had to wait much longer for a specialized variance formula to accompany the point estimate.]

(c) Using a likelihood-based approach to estimation of $\theta = \log OR$, in which you represent each of the two items of data as normal-based log likelihoods centered on $\hat{\theta}_M$ and $\hat{\theta}_M$,then add the two log-likelihoods. Hint: since each log-likelihood is a quadratic form in θ, and since their sum is again a a quadratic form in θ, this amounts to working out where the new log-likelihood is centered, and what its curvature is. Show that its centre has the same form as the one used by Woolf.