PART D

LONGITUDINAL STUDIES: ANALYSIS OF RATES AND SURVIVAL TIMES

In this part of the book we describe methods for the analysis of **longitudinal studies**, that is studies in which subjects are followed over time. These may be subdivided into three main types:

- **cohort studies** in which a group of individuals is followed over time, and the incidence of one or more outcomes is recorded, together with exposure to one or more factors
- **survival studies** in which individuals are followed from the time they experience a particular event such as the diagnosis of disease, and the time to recurrence of the disease or death is recorded
- **intervention studies** in which subjects are randomized to two or more intervention or treatment groups (one of which is often a control group with no active intervention or treatment or with standard care); the occurrence of pre-specified outcomes is recorded

These different types of study are described in more detail in Chapter 34. Our focus is on methods for their analysis, where the *outcome of interest* is *binary*, and where:

1 individuals in the study *are followed over different lengths of time*, and/or
2 we are interested not only in whether or not the outcome occurs, but also the *time at which it occurs*.

Note that for longitudinal studies in which everyone is followed for *exactly* the same length of time, the methods described in Part C can be used if the outcome is defined as the *risk* or *odds* of the event of interest. The exception is studies when most subjects will experience the event of interest by the end of the follow-up. For example, in a trial of a new treatment approach for lung cancer, even if every patient were followed for 10 years, the focus would be on assessing whether the new treatment had extended the survival time, rather than comparing the proportion who survived in each group. This is because lung cancer has a very poor prognosis; the probability of anyone surviving for more than 10 years is close to zero.

In Chapter 22 we explain why variable follow-up times are common and the special issues that arise in their analysis, and we define **rates of disease and mortality** as the appropriate outcome measure. We then introduce the **Poisson distribution** for the *sampling distribution of a rate* and derive a standard error of a rate from it. In Chapter 23 we describe how to compare two rates, and how to control for the effects of *confounding* using **stratification** methods, and in Chapter

24 the use of **Poisson regression** methods. In Chapter 25 we describe the use of standardized rates to enable ready comparison between several groups. This part of the book concludes with the group of methods known as **survival analysis**; Chapter 26 covers the use of life tables, Kaplan–Meier estimates of survival curves and log rank tests, and Chapter 27 describes Cox (proportional hazards) regression for the analysis of survival data. In contrast to the other methods for the analysis of longitudinal studies presented earlier in this part, survival analysis methods do not require the rate(s) to be constant during specified time periods.

We will assume throughout this part of the book that individuals can only experience one occurrence of the outcome of interest. This is not the case where the outcome of interest is a **disease or condition that can recur**. Examples are episodes of diarrhoea, acute respiratory infection, malaria, asthma and myocardial infarction, which individuals may experience more than once during the course of the study. Although we can apply the methods described in this part of the book by defining the outcome as the occurrence of one or more events, and using the time until the *first* occurrence of the event, a more appropriate approach is to use the methods presented in Chapter 31, which describes the analysis of **clustered data**. The methods in Chapter 31 also apply to the analysis of longitudinal studies in which we take **repeated measures of a quantitative outcome variable**, such as blood pressure or lung function, on the same individual.

CHAPTER 22

Longitudinal studies, rates and the Poisson distribution

22.1 INTRODUCTION

In this chapter we introduce the **rate** of event occurrence, as the outcome measure for the analysis of longitudinal studies. We explain why variable follow-up times happen, show how rates are estimated and discuss what they mean and how they relate to the measure of event occurrence described in Part C. We then describe the **Poisson distribution** for the *sampling distribution of a rate*, and use its properties to derive confidence intervals for rates. In the next chapter we introduce two measures used to compare rates in different exposure groups; the **rate ratio** and the **rate difference**.

22.2 CALCULATING PERIODS OF OBSERVATION (FOLLOW-UP TIMES)

In the majority of longitudinal studies, individuals are followed for different lengths of time. Methods that take this into account are the focus of this part of the book. Variable follow-up times occur for a variety of reasons:

- for logistic reasons, individuals may be recruited over a period of time but followed to the same end date
- in an intervention or cohort study, new individuals may be enrolled during the study because they have moved into the study area
- in a survival study, there may be a delay between the diagnosis of the event and recruitment into the study
- some individuals may be lost to follow up, for example because of emigration out of the study area or because they choose to withdraw from the study
- some individuals may die from causes other than the one that is the focus of interest
- in studies where the population of interest is defined by their age, for example women of child bearing age (ie. 15–44 years), individuals may move into or out of the group during the study as they age.

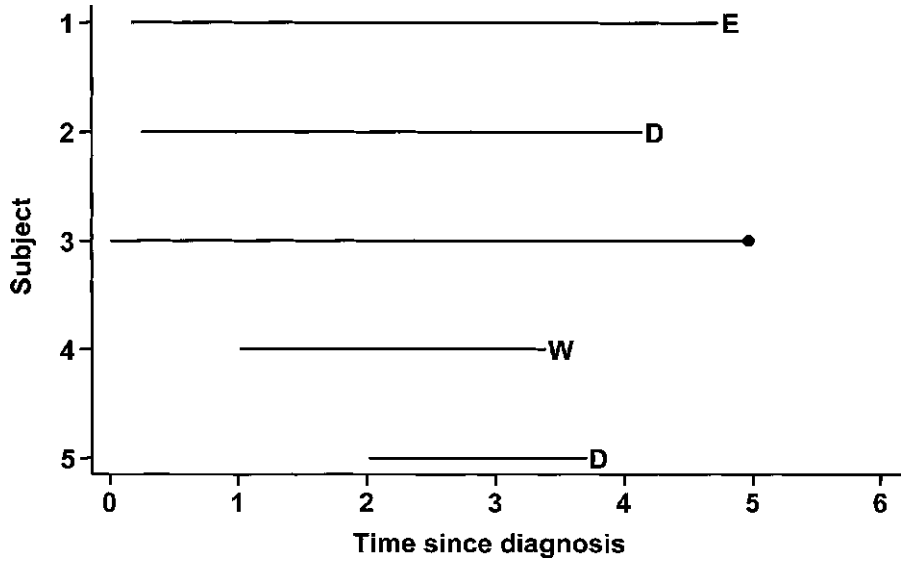

Fig. 22.1 Follow-up histories for 5 subjects in a study of mortality after a diagnosis of prostate cancer (D = died, E = emigrated, W = withdrew, • = reached the end of follow-up without experiencing the disease event).

Figure 22.1 depicts an example from a study of prostate cancer, which shows that subjects were recruited to the study at varying times after diagnosis and exited at different points in time. Only subject 3 was followed for the full 5 years: subjects 2 and 5 died, subject 1 emigrated and subject 4 withdrew from the study. Survival times for subjects who are known to have survived up to a certain point in time, such as subjects 1 and 4, but whose survival status past that point is not known, are said to be **censored**.

An individual's **period of observation** (or **follow-up time**) starts when they join the study and stops when they experience the outcome, are lost to follow-up, or the follow-up period ends, whichever happens first. This is the time during which, were they to experience an event, the event would be recorded in the study. This period is also called the **period at risk**. It is often measured in years, when it is called **person-years-at-risk** or **pyar**.

The occurrence and timings of outcome events, losses to follow-up, and recruitment of new participants are most accurately determined through regular surveillance of the study population. In some countries this may be possible using national databases, for example of deaths or cancer events, by 'flagging' the subjects under surveillance in the study so that the occurrence of events of interest can be routinely detected. In other settings it may be necessary to carry out community-based surveillance. For logistic simplicity, and cost considerations, this is sometimes carried out by conducting just two cross-sectional surveys, one at the beginning and one at the end of the study period, and enquiring about changes in the intervening period. If the exact date of an outcome event, loss to follow-up,

or new recruitment cannot be determined through questioning, it is usually assumed to have occurred half-way through the interval between the surveys.

Using statistical computer packages to calculate periods of follow-up

When analysing longitudinal studies, it is important to choose a statistical computer package that allows easy manipulation of dates. Many packages provide a facility for automatic recoding of dates as the total number of days that have elapsed since the start of the Julian calendar, or from a chosen reference date such as 1/Jan/1960. Thus, for example, 15/Jan/1960 would be coded as 14, 2/Feb/1960 as 32, 1/Jan/1959 as −365 and so on. It is then easy to calculate the time that has elapsed between two dates. If the recoded variables are *startdate* and *exitdate*, and since (taking leap years into account) there are on average 365.25 days in a year, the follow-up time in years is given by:

$$\text{Follow-up time in years} = (exitdate - startdate)/365.25$$

22.3 RATES

The **rate** of occurrence of an outcome event measures the number of new events that occur per person per unit time, and is denoted by the Greek letter λ (lambda). Some examples of rates are:

- In the UK, the *incidence rate* of prostate cancer is 74.3/100 000 men/year. In other words, 74.3 new cases of prostate cancer are detected among every 100 000 men each year
- In the UK, the *mortality rate* from prostate cancer is 32.5/100 000 men/year. In other words 32.5 out of every 100 000 men die from prostate cancer each year
- In the UK, the incidence rate of abortions among teenage girls aged 16–19 years rose from 6.1/1000 girls/year in 1969 to 26.0/1000 girls/year in 1999

The rate is estimated from study data by dividing the total number (d) of events observed by the total (T) of the individual person-years of observation.

$$\text{Rate, } \lambda = \frac{\text{number of events}}{\text{total person-years of observation}} = \frac{d}{T}$$

Note that the sum, T, of the individual person-years is equivalent to the average number of persons under observation multiplied by the length of the study.

The rate is also known as the **incidence rate** (or **incidence density**) of the outcome event, except when the outcome of interest is death, in which case it is called the **mortality rate**. For rare events, the rate is often multiplied by 1000 (or even 10 000

or 100 000) and expressed per 1000 (or 10 000 or 100 000) person-years-at-risk. For a **common disease** such as diarrhoea or asthma, which may occur more than once in the same person, the incidence rate measures the average number of attacks per person per year (at risk). However, the standard methods for the analysis of rates (described in this part of the book) are not valid when individuals may experience multiple episodes of disease. We explain how to deal with this situation in Chapter 31.

Example 22.1

Five hundred children aged less than 5 years living in a community in rural Guatemala were enrolled in a study of acute lower respiratory infections. Fifty-seven were hospitalized for an acute lower respiratory infection, after which they were no longer followed in the study. The study lasted for 2 years, but because of migration, the occurrence of infections, passing the age of 5, and losses to follow-up, the number under surveillance declined with time and the total child-years at risk was $T = 873$ (i.e. an average population size of 436 over the 2 years). The rate of acute lower respiratory infections was therefore estimated to be:

$$\lambda = 57/873 = 0.0653 \text{ per child-year}$$

This can also be expressed per 1000 child-years at risk, as:

$$\lambda = 57/873 \times 1000 = 65.3 \text{ per } 1000 \text{ child-years}$$

Note that the estimated rate will be the same whether the child-years of follow-up arise from following (for example) 1000 children for 1 year, 500 children for 2 years or 250 children for 4 years (and so on).

Understanding rates and their relationship with risks

The rate relates the number of new events to total observation time. This is in contrast to the **risk**, or **cumulative incidence** (see Chapter 15), in which the number of new events is related to the number at risk at the beginning of the observation period; the longer the period of observation the greater the risk will be, since there will be more time for events to occur. Measures of risk therefore contain an implicit but not explicit time element.

Figure 22.2 illustrates the accumulation of new cases of a disease over a 5 year period in a population *initially disease free*, for two somewhat different incidence rates: (a) $\lambda = 0.3$/person/year, and (b) $\lambda = 0.03$/person/year. For ease of understanding, we are illustrating this assuming that the population remains constant over the 5 years, and that there is complete surveillance; that is that there are no losses to follow-up, and no migration either in or out.

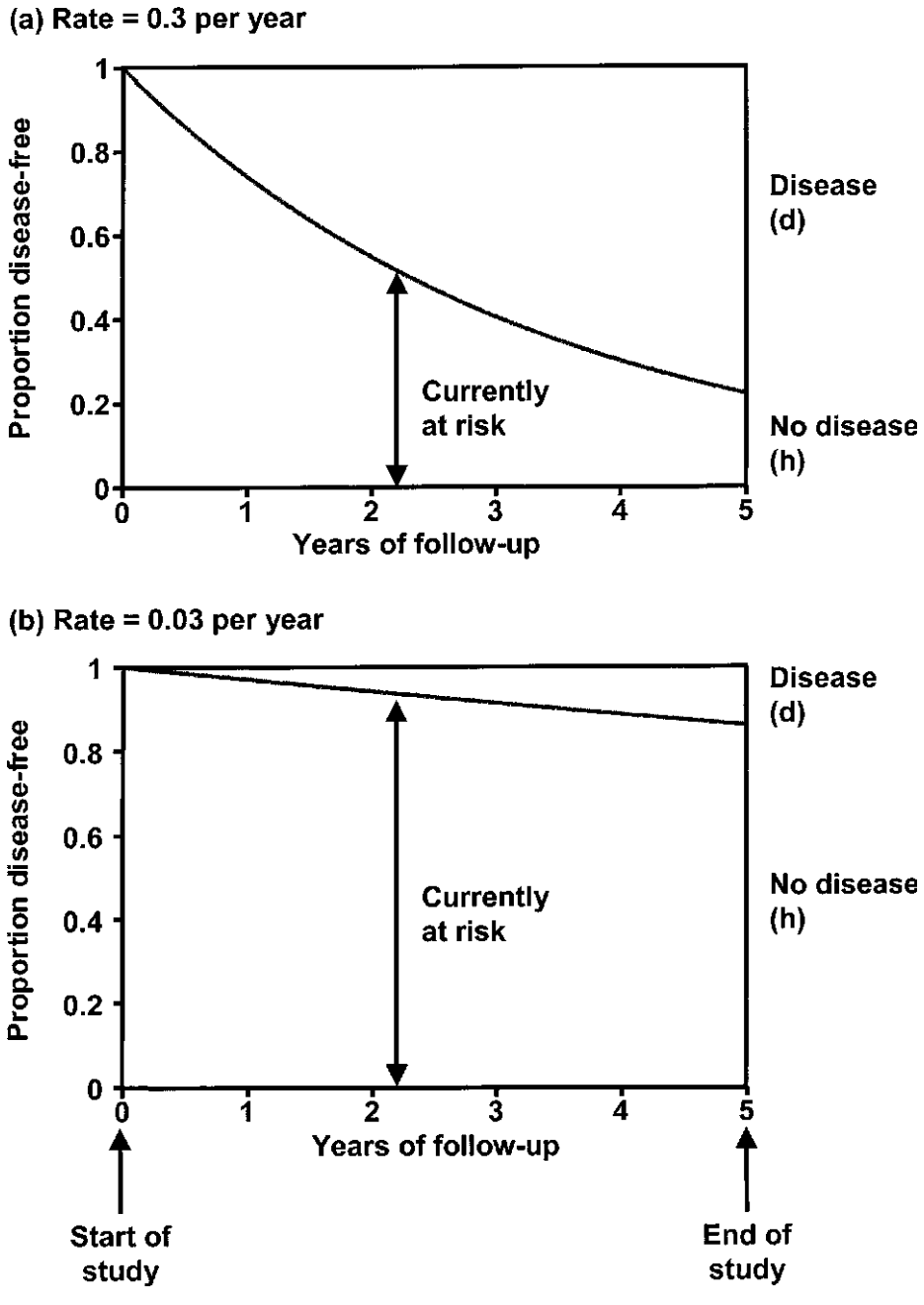

Fig. 22.2 A graphical representation of two follow-up studies which lasted for 5 years. In the top graph (a) the rate of disease is 0.3/person/year, and the disease-free population declines exponentially with time. In the bottom graph (b) the rate is 0.03/person/year, and the decline in the disease-free population is approximately linear over the period of the study.

The disease rate applies to the number of people disease-free at a particular point in time. Understanding the effect of this is a bit like understanding the calculation of compound interest rates. In Figure 22.2(a), the incidence rate is high, and so the proportion of the population remaining disease free is changing rapidly over time. The disease rate is therefore operating on an ever-diminishing proportion of the population as time goes on. This means that the number of new cases per unit time will be steadily decreasing.

In other words, although the disease rate is constant over time, the cumulative incidence and risk do not increase at a constant pace; their increase slows down over time. This is reflected by a steadily decreasing gradient of the graph showing how the disease-free population is diminishing over time (or equivalently how the number who have experienced the disease, that is the cumulative incidence, is accumulating). It can be shown mathematically that when the rate is constant over time, this graph is described by an *exponential* function, and that:

Proportion disease free at time $t = e^{-\lambda t}$

Risk up to time $t = 1 - e^{-\lambda t}$

Average time to contracting the disease $= 1/\lambda$

In Figure 22.2(b), the incidence rate is low and so the proportion of the population remaining disease-free decreases slowly over time. It remains sufficiently close to one over the 5 years that the exponential curve is approximately linear, corresponding to a constant increase of new cases (and therefore of risk) over time. In fact when the value of λ *is very small*, the risk is approximately equal to the rate multiplied by the time:

When λ is very small, risk up to time $t \approx \lambda t$, so that

$$\lambda \approx \frac{\text{risk}}{t}$$

Table 22.1 shows the values of the risks (up to 1, 2 and 5 years) that result from these two very different rates. This confirms what we can see visually in Figure 22.2. For the high rate ($\lambda = 0.3$/person/year), the number of new cases per unit time is steadily decreasing; the increase is always less than the rate because the size of the 'at risk' population is decreasing rapidly. Thus at 1 year, the cumulative risk is a bit less than the rate (0.26 compared to 0.3), at 2 years it is considerably less than twice the rate (0.45 compared to 0.6), and so on. In contrast, for the low rate ($\lambda = 0.03$/person/year), the number of new cases is increasing steadily, and the risk increases by approximately 0.03/year.

Table 22.1 Risks of disease up to 1, 2 and 5 years corresponding to rates of $\lambda = 0.3$/person/year, and $\lambda = 0.03$/person/year.

	Risk of disease		
Rate of disease	Over 1 year	Over 2 years	Over 5 years
0.3/person/year	$1 - e^{-0.3} = 0.26$	$1 - e^{-0.3\times2} = 0.45$	$1 - e^{-0.3\times5} = 0.78$
0.03/person/year	$1 - e^{-0.03} = 0.03$	$1 - e^{-0.03\times2} = 0.06$	$1 - e^{-0.03\times5} = 0.14$

We have demonstrated that when λ is very small, the risk up to time t approximately equals λt. This is equivalent to the rate, λ, being approximately equal to the value of the risk per unit time (risk/t). We will now show that the value of risk/t also gets close to the rate as the length of the time interval gets very small. This is true whatever the size of the rate, and is the basis of the **formal definition of a rate**, as the value of risk/t when t is very small.

$$\lambda = \frac{\text{risk}}{t}, \text{ when } t \text{ is very small}$$

Table 22.2 illustrates this for the fairly high rate of $\lambda = 0.3$/person/year. Over 5 years, the risk per year equals 0.1554, just over half the value of the rate. If the length of time is decreased to 1 year, the risk per year is considerably higher at 0.2592, but still somewhat less than the rate of 0.3 per year. As the length of time decreases further, the risk per year increases; by one month it is very close to the rate, and by one day almost equal to it.

Table 22.2 Risk of disease, and risk/t, for different lengths of time interval t, when the rate, $\lambda = 0.3$/person/year.

	Length of time interval, t						
	5 years	1 year	1 month (30 days)	1 week	1 day	1 hour	1 minute
t (years)	5	1	0.08219	0.01918	0.002740	0.0001142	0.000001900
risk $= 1 - e^{-0.3t}$	0.7769	0.2592	0.02436	0.005737	0.0008216	0.00003420	0.0000005710
risk/t	0.1554	0.2592	0.2963	0.2992	0.2999	0.3000	0.3000

22.4 THE POISSON DISTRIBUTION

We have already met the normal distribution for means and the binomial distribution for proportions. We now introduce the **Poisson distribution**, named after the French mathematician, which is appropriate for describing the *number* of occurrences of an event during a period of time, provided that these events

occur independently of each other and at random. An example would be the number of congenital malformations of a given type occurring in a particular district each year, provided that that there are no epidemics or specific environmental hazards and that the population is constant from year to year (also see Example 22.2).

The Poisson distribution is also appropriate for the *number* of particles found in a unit of space, such as the number of malaria parasites seen in a microscope field of a blood slide, provided that the particles are distributed *randomly* and *independently* over the total space. The two properties of randomness and independence must both be fulfilled for the Poisson distribution to hold. For example, the number of *Schistosoma mansoni* eggs in a stool slide will not be Poisson, since the eggs tend to cluster in clumps rather than to be distributed independently.

After introducing the Poisson distribution in general for the number of events, we will explain its application to the analysis of rates.

Definition of the Poisson distribution

The Poisson distribution describes the sampling distribution of the number of occurrences, d, of an event during a period of time (or region of space). It depends upon just one parameter, which is the mean number of occurrences, μ, in periods of the same length (or in equal regions of space).

$$\text{Probability } (d \text{ occurrences}) = \frac{e^{-\mu}\,\mu^{d}}{d!}$$

Note that, by definition, both 0! and μ^0 equal 1. The probability of zero occurrences is therefore $e^{-\mu}$ (e is the mathematical constant 2.71828...).

$$\text{Mean number of occurrences} = \mu$$
$$\text{s.e. of number of occurrences} = \sqrt{\mu}$$

The standard error for the number of occurrences equals the square root of the mean, which is estimated by the square root of the observed number of events, $\sqrt{d}$.

Example 22.2

A district health authority which plans to close the smaller of two maternity units is assessing the extra demand this will place on the remaining unit. One factor being considered is the risk that on any given day the demand for admissions will

exceed the unit's capacity. At present the larger unit averages 4.2 admissions per day and can cope with a maximum of 10 admissions per day. This results in the unit's capacity being exceeded only on about one day per year. After the closure of the smaller unit the average number of admissions is expected to increase to 6.1 per day. The Poisson distribution can be used to estimate the proportion of days on which the unit's capacity is then likely to be exceeded. For this we need to determine the probability of getting 11 or more admissions on any given day. This is most easily calculated by working out the probabilities of 0, 1, 2 . . . or 10 admissions and subtracting the total of these from 1, as shown in Table 22.3. For example:

$$\text{Probability (three admissions)} = \frac{e^{-6.1}\,6.1^3}{3!}$$

The calculation shows that the probability of 11 or more admissions in a day is 0.0470. The unit's capacity is therefore likely to be exceeded 4.7% of the time, or on about 17 days per year.

Table 22.3 The probabilities of the number of admissions made during a day in a maternity unit, based on a Poisson distribution with a mean of 6.1 admissions per day.

No. of admissions	Probability
0	0.0022
1	0.0137
2	0.0417
3	0.0848
4	0.1294
5	0.1579
6	0.1605
7	0.1399
8	0.1066
9	0.0723
10	0.0440
Total (0 – 10)	0.9530
11+ (by subtraction, 1 – 0.9530)	0.0470

Shape of the Poisson distribution

Figure 22.3 shows the shape of the Poisson distribution for various values of its mean, μ. The distribution is very skewed for small means, when there is a sizeable probability that zero events will be observed. It is symmetrical for large means and is adequately approximated by the normal distribution for values of $\mu = 10$ or more.

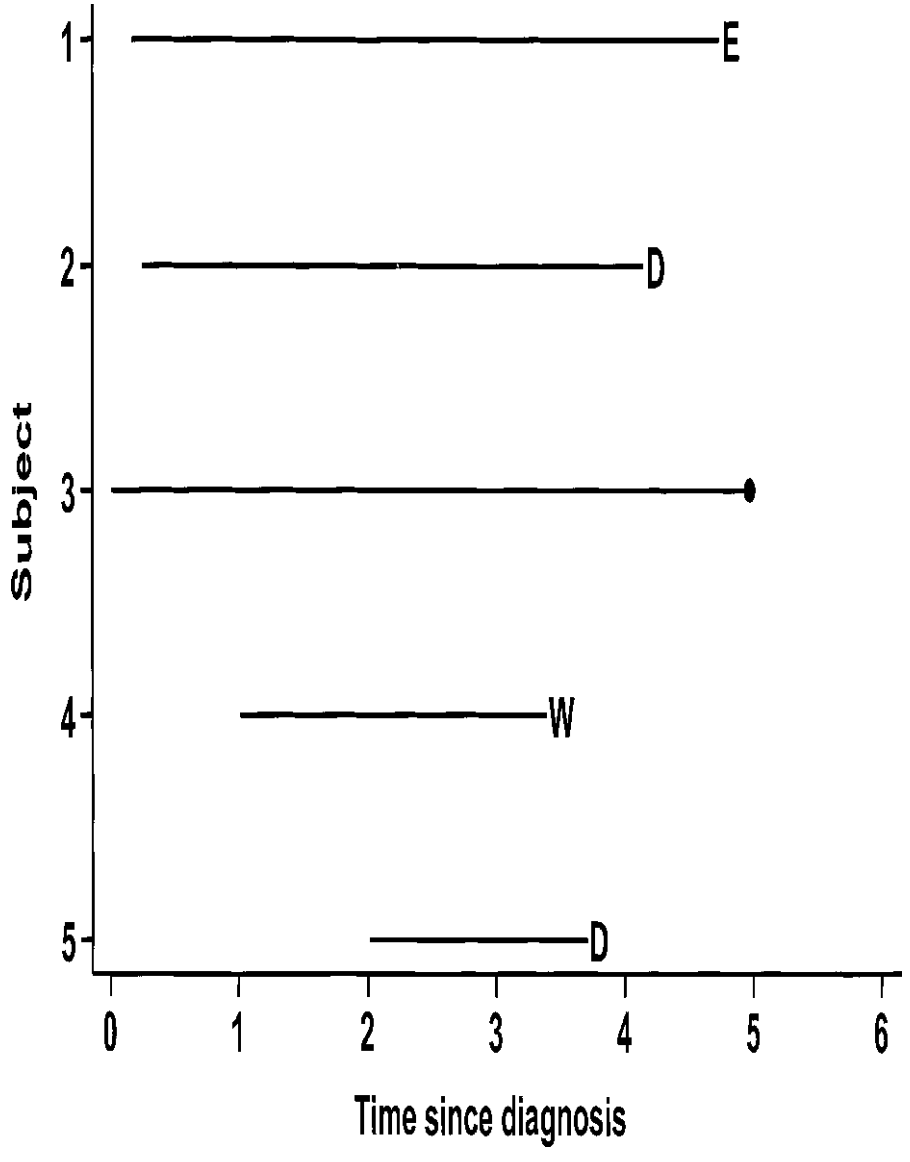

Fig. 22.3 Poisson distribution for various values of μ. The horizontal scale in each diagram shows values of the number of events, d.

Use of the Poisson distribution

The Poisson distribution (and its normal approximation) can be used whenever it is reasonable to assume that the outcome events are occurring independently of each other and randomly in time. This assumption is, of course, less likely to be

true for infectious than for non-communicable diseases but, provided there is no strong evidence of disease clustering, the use is still justified. Specific techniques exist to detect disease clustering in time and/or space (see Elliott *et al*., 2000), such as the possible clustering of cases of leukaemia or variant Creutzfeldt–Jakob disease in a particular area. Such clusters violate what might otherwise be a Poisson distribution.

22.5 STANDARD ERROR OF A RATE

We now discuss the use of the Poisson distribution for the analysis of rates. Recall that:

$$\text{Rate, } \lambda = \frac{\text{number of events}}{\text{total person-years of observation}} = \frac{d}{T}$$

Although the value of the total person-years of observation (T) is affected by the number of events, and the time at which they occur (since an individual's period of observation only contributes until they experience an event, as then they are no longer at risk), it can be shown that we do not need to explicitly consider this variation in T. We can therefore calculate the standard error of a rate as follows:

$$\text{s.e. (rate)} = \frac{\text{s.e. (number of events)}}{T} = \frac{\sqrt{d}}{T} = \sqrt{\frac{\lambda}{T}}$$

The right hand version of the formula (derived by replacing $\sqrt{d}$ with $\sqrt{(\lambda T)}$) makes it clear that the standard error of the rate will be smaller the larger the total person-years of observation, as λ will be the same, on average, whatever the value of this.

Example 22.1 (continued)

We showed earlier that in the 2-year morbidity study in rural Guatemala the rate of acute lower respiratory infections, expressed per 1000 child-years at risk, was estimated to be 65.3 per 1000 child-years. The standard error of the rate is:

$$\text{s.e.} = \frac{\sqrt{d}}{T} \times 1000 = \frac{\sqrt{57}}{873} \times 1000 = 8.6$$

22.6 CONFIDENCE INTERVAL FOR A RATE

A confidence interval for a rate can be derived from its standard error, in the usual way. However, it is preferable to work on the log scale and to derive a confidence interval for the log rate, and then to antilog this to give a confidence interval for a rate, since this takes account of the constraint that the rate must be greater than or equal to zero. We now show how to do this.

The formula for the standard error of the log rate is derived using the **delta method** (see Box 16.1 on p. 157), and is:

$$\text{s.e.}(\log \text{ rate}) = \frac{1}{\sqrt{d}}$$

Thus, perhaps surprisingly, the standard error of the log rate depends only on the number of events, and not on the length of follow-up time. In the same way as shown in Chapter 16, the steps of calculating the confidence interval on the log scale and then converting it to give a confidence interval for the rate can be combined into the following formulae:

$$95\%\ \text{CI (rate)} = \text{rate/EF to rate} \times \text{EF}$$

$$\text{Error factor (EF)} = \exp(1.96/\sqrt{d})$$

Example 22.1 (continued)

For the Guatemala morbidity study there were 57 lower respiratory infections in 873 child-years at risk. The log rate per 1000 child-years at risk, is $\log(\lambda) = \log(1000 \times 57/873) = \log(65.3) = 4.179$. The standard error of this log rate is:

$$\text{s.e.}(\log \text{ rate}) = 1/\sqrt{d} = 1/\sqrt{57} = 0.132$$

1 The 95% confidence interval for the log rate is therefore:

$$95\%\ \text{CI} = 4.179 - (1.96 \times 0.132) \text{ to } 4.179 + (1.96 \times 0.132) = 3.919 \text{ to } 4.438$$

The 95% confidence interval for the rate is:

$$95\%\ \text{CI} = \exp(3.919) \text{ to } \exp(4.438) = 50.36 \text{ to } 84.65 \text{ infections per 1000 child-years}$$

2 Alternatively, we may calculate the 95% CI using the 95% error factor (EF) for the rate:

$$\text{EF} = \exp(1.96/\sqrt{d}) = \exp(1.96/\sqrt{57}) = 1.296$$

The 95% confidence interval for the rate is:

$$\begin{aligned} 95\%\ \text{CI} &= \frac{\lambda}{\text{EF}} \text{ to } \lambda \times \text{EF} = 65.3/1.296 \text{ to } 65.3 \times 1.296 \\ &= 50.36 \text{ to } 84.65 \text{ infections per 1000 child-years} \end{aligned}$$

CHAPTER 23

Comparing rates

23.1 INTRODUCTION

In this chapter we describe the two measures used to compare rates in different exposure groups: the **rate difference** and the **rate ratio**. We then show how to use Mantel–Haenszel methods to estimate rate ratios controlling for confounding factors. In Part C we emphasized the similarity between Mantel–Haenszel methods, which use stratification to estimate odds ratios for the effect of exposure controlled for the effects of confounding variables, and logistic regression models. Mantel–Haenszel methods for rate ratios are closely related to the corresponding regression model for rates, Poisson regression, which is introduced in Chapter 24.

23.2 COMPARING TWO RATES

We now see how the rates of disease in two exposure groups may be compared, using two different measures: the **rate difference** and the **rate ratio**.

Rate differences

Example 23.1

The children in the Guatemala morbidity study analysed in Example 22.1 were subdivided according to the quality of their housing conditions. The data are shown in Table 23.1, together with the notation we will use. We will consider children living in poor housing conditions to be the exposed group and, as in Part C, denote exposed and unexposed groups by the subscripts 1 and 0 respectively. The **rate difference** comparing poor with good housing is $93.0 - 46.3 = 46.7$ infections per 1000 child-years.

Table 23.1 Incidence of lower respiratory infection among children aged less than 5 years, according to their housing conditions.

Housing condition	Number of acute lower respiratory infections	Child-years at risk	Rate/1000 child-years
Poor (exposed)	$d_1 = 33$	$T_1 = 355$	$\lambda_1 = 93.0$
Good (unexposed)	$d_0 = 24$	$T_0 = 518$	$\lambda_0 = 46.3$
Total	$d = 57$	$T = 873$	$\lambda = 65.3$

The **standard error of a rate difference** is:

$$\text{s.e. (rate difference)} = \sqrt{\left(\frac{d_1}{T_1^2} + \frac{d_0}{T_0^2}\right)}$$

This can be used in the usual way to derive a 95% confidence interval. In this example,

$$\begin{aligned}\text{s.e.} &= \sqrt{\left(\frac{d_1}{T_1^2} + \frac{d_0}{T_0^2}\right)} = \sqrt{\left(\frac{33}{355^2} + \frac{24}{518^2}\right)} \times 1000 \\ &= 18.7 \text{ infections per 1000 child-years}\end{aligned}$$

and the 95% confidence interval is:

$$\begin{aligned}&46.7 - 1.96 \times 18.7 \text{ to } 46.7 + 1.96 \times 18.7 \\ &= 10.0 \text{ to } 83.4 \text{ infections per 1000 child-years}\end{aligned}$$

With 95% confidence, the rate of lower respiratory infections among children living in poor housing exceeds the rate among children living in good housing by between 10.0 and 83.4 infections per 1000 child-years.

Rate ratios

As explained in more detail in the next chapter, the analysis of rates is usually done using **rate ratios** rather than rate differences. The rate ratio is defined as:

$$\text{Rate ratio} = \frac{\text{rate in exposed}}{\text{rate in unexposed}} = \frac{\lambda_1}{\lambda_0} = \frac{d_1/T_1}{d_0/T_0} = \frac{d_1 \times T_0}{d_0 \times T_1}$$

As for risk ratios and odds ratios, we use the **standard error of the log rate ratio** to derive confidence intervals, and tests of the null hypothesis of no difference

between the rates in the two groups. This (again derived using the **delta method**) is given by:

$$\text{s.e. of log(rate ratio)} = \sqrt{(1/d_1 + 1/d_0)}$$

The **95% confidence interval for the rate ratio** is:

$$95\%\ \text{CI} = \text{rate ratio/EF to rate ratio} \times \text{EF, where}$$
$$\text{EF} = \exp[1.96 \times \text{s.e. of log(rate ratio)}]$$

z-test for the rate ratio

A ***z*-test** (Wald test, see Chapter 28) of the null hypothesis that the rates in the two groups are equal is given by:

$$z = \frac{\text{log(rate ratio)}}{\text{s.e. of log(rate ratio)}}$$

Example 23.1 (continued)
The rate ratio comparing children living in poor housing with those living in good housing is:

$$\text{rate ratio} = \frac{33/355}{24/518} = 2.01$$

The standard error of the log(rate ratio) is $\sqrt{(1/33 + 1/24)} = 0.268$, and the 95% error factor is:

$$95\%\ \text{EF} = \exp(1.96 \times 0.268) = 1.69$$

A 95% confidence interval for the rate ratio is thus:

$$95\%\ \text{CI} = 2.01/1.69 \text{ to } 2.01 \times 1.69 = 1.19 \text{ to } 3.39$$

With 95% confidence, the rate of acute lower respiratory infections among children living in poor housing is between 1.19 and 3.39 times the rate among children living in good housing. The z statistic is $\log(2.01)/0.268 = 2.60$; the corresponding P value is 0.009. There is therefore good evidence against the null hypothesis that infection rates are the same among children living in good and poor quality housing.

Relationship between rate ratio, risk ratio and odds ratio

From Chapter 16, we know that for a rare event the risk ratio is approximately equal to the odds ratio. And in the last chapter we saw that for a rare event, risk up to time t approximately equals λt. It therefore follows that for a rare event the risk ratio and rate ratio are also approximately equal:

$$\text{Risk ratio} \approx \frac{\lambda_1 t}{\lambda_0 t} = \frac{\lambda_1}{\lambda_0} = \text{Rate ratio} \approx \text{Odds ratio}$$

However when the event is not rare the three measures will all be different. These different measures of the association between exposure and outcome event, and of the impact of exposure, are discussed in more detail in Chapter 37.

23.3 MANTEL–HAENSZEL METHODS FOR RATE RATIOS

Recall from Chapter 18 that a **confounding** variable is one that is related both to the outcome variable and to the exposure of interest (see Figure 18.1), and that is not a part of the causal pathway between them. Ignoring the effects of confounding variables may lead to bias in our estimate of the exposure–outcome association. We saw that we may allow for confounding in the analysis via **stratification**: restricting estimation of the exposure–outcome association to individuals with the same value of the confounder. We then used **Mantel–Haenszel** methods to combine the stratum-specific estimates, leading to an estimate of the *summary odds ratio*, controlled for the confounding.

We now present Mantel–Haenszel methods for rate ratios. Table 23.2 shows the notation we will use for the number of events and person-years in each group, in stratum i. The notation is exactly the same as that in Table 23.1, but with the subscript i added, to refer to the stratum i.

Table 23.2 Notation for the table for stratum i.

	Number of events	Person-years at risk
Group 1 (Exposed)	d_{1i}	T_{1i}
Group 0 (Unexposed)	d_{0i}	T_{0i}
Total	$d_i = d_{0i} + d_{1i}$	$T_i = T_{0i} + T_{1i}$

The data consist of c such tables, where c is the number of different values the confounding variable can take. The estimate of the rate ratio for stratum i is

$$RR_i = \frac{d_{1i}/T_{1i}}{d_{0i}/T_{0i}} = \frac{d_{1i} \times T_{0i}}{d_{0i} \times T_{1i}}$$

Mantel–Haenszel estimate of the rate ratio controlled for confounding

As for the odds ratio, the **Mantel–Haenszel estimate of the rate ratio** is a *weighted average* (see Section 18.3) of the rate ratios in each stratum. The weight for each rate ratio is:

$$w_i = \frac{d_{0i} \times T_{1i}}{T_i}$$

Since the numerator of the weight is the same as the denominator of the rate ratio in stratum i, $w_i \times RR_i = (d_{1i} \times T_{0i})/T_i$. These weights therefore lead to the following formula for the **Mantel–Haenszel estimate of the rate ratio**:

$$\mathrm{RR_{MH}} = \frac{\Sigma(w_i \times RR_i)}{\Sigma w_i} = \frac{\sum \frac{d_{1i} \times T_{0i}}{T_i}}{\sum \frac{d_{0i} \times T_{1i}}{T_i}}$$

Following the notation of Clayton and Hills (1993), this can alternatively be written as:

$$\mathrm{RR_{MH}} = Q/R, \text{ where}$$
$$Q = \sum \frac{d_{1i} \times T_{0i}}{T_i} \text{ and } R = \sum \frac{d_{0i} \times T_{1i}}{T_i}$$

Example 23.2

Data on incidence of acute lower respiratory infections from a study in Guatemala were presented in Example 23.1 and Table 23.1. The rate ratio comparing children living in poor with good housing conditions is 2.01 (95% CI 1.19 to 3.39). Table 23.3 shows the same information, stratified additionally by the type of cooking stove used in the household. ■

Table 23.3 Association between incidence of acute lower respiratory infection and housing conditions, stratified by type of cooking stove.

(a) Wood burning stove (stratum 1)

Housing condition	Number of infections	Child-years at risk	Rate/1000 child-years
Poor (exposed)	$d_{11} = 28$	$T_{11} = 251$	$\lambda_{11} = 111.6$
Good (unexposed)	$d_{01} = 5$	$T_{01} = 52$	$\lambda_{01} = 96.2$
Total	$d_1 = 33$	$T_1 = 303$	$\lambda_1 = 108.9$
Rate ratio = 1.16 (95% CI 0.45 to 3.00), $P = 0.76$			

(b) Kerosene or gas stove (stratum 2)

Housing condition	Number of infections	Child-years at risk	Rate/1000 child-years
Poor (exposed)	$d_{12} = 5$	$T_{12} = 104$	$\lambda_{12} = 48.1$
Good (unexposed)	$d_{02} = 19$	$T_{02} = 466$	$\lambda_{02} = 40.8$
Overall	$d_2 = 24$	$T_2 = 570$	$\lambda_2 = 42.1$
Rate ratio = 1.18 (95% CI 0.44 to 3.16), $P = 0.74$			

Table 23.4 Person-years of observation according to housing conditions and type of cooking stove.

	Type of stove	
Housing condition	Wood burning stove	Gas or kerosene stove
Poor (exposed)	$T_{11} = 251$	$T_{21} = 104$
Good (unexposed)	$T_{10} = 52$	$T_{20} = 466$

Examination of the association between quality of housing and infection rates in the two strata defined by type of cooking stove shows that there is little evidence of an association in either stratum. Type of cooking stove is a strong confounder of the relationship between housing quality and infection rates, because most poor quality houses have wood burning stoves while most good quality houses have kerosene or gas stoves. This can be seen by tabulating the person-years of observation according to housing condition and cooking stove, as shown in Table 23.4.

Table 23.5 shows the calculations needed to derive the Mantel–Haenszel rate ratio combining the stratified data, presented in Table 23.3, on the association between housing conditions (the exposure variable) and the incidence of acute lower respiratory infection (the outcome), controlling for type of stove.

The Mantel–Haenszel estimate of the rate ratio equals:

$$\mathrm{RR}_{\mathrm{MH}} = Q/R = 8.89/7.61 = 1.17$$

Table 23.5 Calculations required to derive the Mantel–Haenszel summary rate ratio, with associated confidence interval and P value.

Stratum i	RR_i	$w_i = \frac{d_{0i} \times T_{1i}}{T_i}$	$w_i \times RR_i$	V_i	d_{1i}	E_{1i}
Wood stove ($i = 1$)	1.16	4.14	4.81	4.69	28	27.34
Kerosene/gas ($i = 2$)	1.18	3.47	4.09	3.58	5	4.38
Total		$R = 7.61$	$Q = 8.89$	$V = 8.27$	$O = 33$	$E = 31.72$

After controlling for the confounding effect of type of stove, the rate of infection is only slightly (17%) greater among children living in poor housing conditions compared to children living in good housing conditions.

Standard error and confidence interval for the Mantel–Haenszel RR

As is usual for ratio measures, the **95% confidence interval for RR_{MH}** is derived using the standard error of $\log(RR_{MH})$, denoted by s.e._{MH}.

$$95\%\ \text{CI} = RR_{MH}/\text{EF} \text{ to } RR_{MH} \times \text{EF}, \text{ where}$$
$$\text{the error factor EF} = \exp(1.96 \times \text{s.e.}_{MH})$$

The simplest formula for the **standard error of $\log RR_{MH}$** (Clayton and Hills 1993) is:

$$\text{s.e.}_{MH} = \sqrt{\left(\frac{V}{Q \times R}\right)}, \text{ where}$$
$$V = \Sigma V_i, \text{ and } V_i = \frac{d_i \times T_{1i} \times T_{0i}}{T_i^2}$$

V is the sum across the strata of the variances V_i for the number of exposed individuals experiencing the outcome event, i.e. the variances of the d_{1i}'s. Note that the formula for the variance V_i of d_{1i} for stratum i gives the same value regardless of which group is considered as exposed and which is considered as unexposed.

Example 23.2 (continued)

Using the results of the calculations for Q, R and V shown in Table 23.5, we find that:

$$\text{s.e.}_{\text{MH}} = \sqrt{\left(\frac{V}{Q \times R}\right)} = \sqrt{\left(\frac{8.27}{8.89 \times 7.61}\right)} = 0.35$$

so that $\text{EF} = \exp(1.96 \times 0.35) = 1.98$, $\text{RR}_{\text{MH}}/\text{EF} = 1.17/1.98 = 0.59$, and $\text{RR}_{\text{MH}} \times \text{EF} = 1.17 \times 1.98 = 2.32$. The 95% confidence interval is therefore:

$$95\%\ \text{CI for}\ \text{RR}_{\text{MH}} = 0.59\ \text{to}\ 2.32$$

Mantel–Haenszel χ^2 test of the null hypothesis

Finally, we test the null hypothesis that $\text{RR}_{\text{MH}} = 1$ by calculating the **Mantel–Haenszel χ^2 test statistic**:

$$\chi^2_{\text{MH}} = \frac{(\Sigma d_{1i} - \Sigma E_{1i})^2}{\Sigma V_i} = \frac{(O - E)^2}{V} = \frac{U^2}{V};\ \text{d.f.} = 1$$

This is based on a comparison in each stratum of the number of exposed individuals observed to have experienced the disease event (d_{1i}) with the expected number in this category (E_{1i}) if there were no difference in the rates between the exposed and unexposed. The expected numbers are calculated in the same way as for the standard χ^2 test described in Chapter 17.

$$E_{1i} = \frac{d_i \times T_{1i}}{T_i}$$

The formula has been simplified by writing O for the sum of the observed numbers, E for the sum of the expected numbers and U for the difference between them:

$$O = \Sigma d_{1i},\ E = \Sigma E_{1i}\ \text{and}\ U = D - E$$

Note that χ^2_{MH} *has just 1 degree of freedom irrespective of how many strata are summarized.*

Example 23.2 (continued)

From the data presented in Table 23.5, a total of $O = 33$ children living in poor housing experienced acute lower respiratory infections, compared with an

expected number of 31.72, based on assuming no difference in rates between poor and good housing. Thus the Mantel–Haenszel χ^2 statistic is:

$$\chi^2_{\mathrm{MH}} = \frac{U^2}{V} = \frac{(33 - 31.72)^2}{8.27} = 0.20 \text{ (1 d.f., } P = 0.655)$$

After controlling for type of cooking stove, there is no evidence of an association between quality of housing and incidence of lower respiratory infections.

Test for effect modification (interaction)

Use of Mantel–Haenszel methods to control for confounding assumes that the exposure–outcome association is the same in each of the strata defined by the levels of the confounder, in other words that the confounder does not modify the effect of the exposure on the outcome event. If this is true, $RR_i = \mathrm{RR}_{\mathrm{MH}}$, and it follows that:

$$(d_{1i} \times T_{0i} - \mathrm{RR}_{\mathrm{MH}} \times d_{0i} \times T_{1i}) = 0$$

The **χ^2 test for heterogeneity** is based on a *weighted* sum of the squares of these differences:

$$\chi^2 = \Sigma \frac{(d_{1i} \times T_{0i} - \mathrm{RR}_{\mathrm{MH}} \times d_{0i} \times T_{1i})^2}{\mathrm{RR}_{\mathrm{MH}} \times V_i \times T_i^2}, \text{ d.f.} = c - 1$$

where V_i is as defined above, and c is the number of strata. The greater the differences between the stratum-specific rate ratios and $\mathrm{RR}_{\mathrm{MH}}$, the larger will be the heterogeneity statistic.

Example 23.2 (continued)

The rate ratios in the two strata were very similar (1.16 in houses with wood-burning stoves and 1.18 in houses with kerosene or gas stoves). We do not, therefore, expect to find evidence of effect modification. Application of the formula for the test for heterogeneity gives $\chi^2 = 0.0005$ (1 d.f.), $P = 0.98$. There is thus no evidence that type of cooking stove modifies the association between quality of housing and rates of respiratory infections.

CHAPTER 24

Poisson regression

24.1 INTRODUCTION

In this chapter we introduce **Poisson regression** for the analysis of rates. This is used to estimate **rate ratios** comparing different exposure groups in the same way that logistic regression is used to estimate *odds ratios* comparing different exposure groups. We will show how it can be used to:

- compare the rates between two exposure (or treatment) groups
- compare more than two exposure groups
- examine the effect of an ordered or continuous exposure variable
- control for the confounding effects of one or more variables
- estimate and control for the effects of **exposures that change over time**

We will see that Poisson regression models comparing two exposure groups give identical rate ratios, confidence intervals and *P*-values to those derived using the methods described in Section 23.2. We will also see that Poisson regression to control for confounding is closely related to the Mantel–Haenszel methods for rate ratios, described in Section 23.3. Finally, we will show how to estimate and control for the effects of variables that change over time, by splitting the follow-up time for each subject.

Like logistic regression models, Poisson regression models are fitted on a *log scale*. The results are then antilogged to give rate ratios and confidence intervals. Since the principles and the approach are exactly the same as those outlined for logistic regression in Part C, a more concise treatment will be given here; readers are referred to Chapters 19 and 20 for more detail. More general issues in regression modelling are discussed in Chapter 29.

24.2 POISSON REGRESSION FOR COMPARING TWO EXPOSURE GROUPS

Introducing the Poisson regression model

The *exposure rate ratio* is defined as:

$$\text{Exposure rate ratio} = \frac{\text{rate in exposed group}}{\text{rate in unexposed group}}$$

If we re-express this as:

$$\text{Rate in exposed group} = \text{Rate in unexposed group} \times \text{Exposure rate ratio}$$

then we have the basis for a model which expresses the rate in each group in terms of two **model parameters**. These are:

1 The **baseline rate**. As in Chapters 19 and 20, we use the term **baseline** to refer to the exposure group against which all the other groups are compared. When there are just two exposure groups, then the baseline rate is the rate in the unexposed group. We use the parameter name Baseline to refer to the rate in the baseline group.
2 The **exposure rate ratio**. This expresses the effect of the exposure on the rate of disease. We use the parameter name Exposure to refer to the exposure rate ratio.

As with logistic regression, Poisson regression models are fitted on a log scale. The two equations that define this model for the *rate* of an outcome event are shown in Table 24.1, together with the corresponding equations for the *log rate*. The equations for the rate can be abbreviated to:

$$\text{Rate} = \text{Baseline} \times \text{Exposure}$$

The two equations that define the Poisson regression model on the log scale can be written:

$$\log(\text{Rate}) = \log(\text{Baseline}) + \log(\text{Exposure rate ratio})$$

Table 24.1 Equations defining the Poisson regression model for the comparison of two exposure groups.

Exposure group	Rate	Log rate
Exposed (*group 1*)	Baseline rate × exposure rate ratio	Log(baseline rate) + log(exposure rate ratio)
Unexposed (*group 0*)	Baseline rate	Log(baseline rate)

In practice, we abbreviate it to:

$$\log(\text{Rate}) = \text{Baseline} + \text{Exposure}$$

since it is clear from the context that output on the log scale refers to log rate and log rate ratios. Note that whereas the exposure effect on the rate ratio scale is *multiplicative*, the exposure effect on the log scale is *additive*.

Example 24.1

All the examples in this chapter are based on a sample of 1786 men who took part in the Caerphilly study, a study of risk factors for cardiovascular disease. Participants were aged between 43 and 61 when they were first examined, and were followed for up to 19 years. The first examinations took place between July 1979 and October 1983, and the follow-up for the outcome (myocardial infarction or death from heart disease) ended in February 1999. Further information about the study can be found at www.epi.bris.ac.uk/mrc-caerphilly.

The first ten lines of the dataset are shown in Table 24.2. Variable '*cursmoke*', short for *current smoker* at recruitment, was coded as 1 for subjects who were smokers and 0 for subjects who were non-smokers, and variable 'MI' was coded as 1 for subjects who experienced a myocardial infarction or died from heart disease during the follow-up period and 0 for subjects who did not. Variable '*years*' is the years of follow-up for each subject (the time from *examdate* to *exitdate*); it was derived using a statistical computer package, as described in Section 22.2.

There were 990 men who were current smokers at the time they were recruited into the study, and 796 men who had never smoked or who were ex-smokers. Table 24.3 shows rates of myocardial infarction in these two groups. The *rate ratio* comparing smokers with never/ex-smokers is $16.98/9.68 = 1.700$.

Table 24.2 First ten lines of the computer dataset from the Caerphilly study. Analyses of the Caerphilly study are by kind permission of the MRC Steering Committee for the Management of MRC Epidemiological Resources from the MRC Epidemiology Unit (South Wales).

id	dob	examdate	exitdate	years	MI	cursmoke
1	20/May/1929	17/Jun/1982	31/Dec/1998	16.54	0	1
2	9/Jul/1930	10/Jan/1983	24/Dec/1998	15.95	0	0
3	6/Feb/1929	23/Dec/1982	26/Nov/1998	15.93	0	1
4	24/May/1931	7/Jul/1983	22/Nov/1984	1.38	1	0
5	9/Feb/1934	3/Sep/1980	19/Dec/1998	18.29	0	0
6	14/Mar/1930	17/Nov/1981	31/Dec/1998	17.12	0	0
7	13/May/1933	30/Oct/1980	27/Dec/1998	18.16	0	1
8	23/May/1924	24/Apr/1980	24/Jan/1986	5.75	1	1
9	20/Jun/1931	11/Jun/1980	12/Dec/1998	18.50	0	1
10	12/May/1929	17/Nov/1979	20/Jan/1995	15.18	1	0

Table 24.3 Rates of myocardial infarction among men who were and were not current smokers at the time they were recruited to the Caerphilly study.

Current smoker at entry to the study	Number of myocardial infarctions	Person-years at risk	Rate per 1000 person-years
Yes (exposed)	$d_1 = 230$	$T_1 = 13\,978$	$\lambda_1 = 230/13.978 = 16.98$
No (unexposed)	$d_0 = 118$	$T_0 = 12\,183$	$\lambda_0 = 118/12.183 = 9.68$
Overall	$d = 348$	$T = 26\,161$	$\lambda = 348/26.161 = 13.30$

We will now show how to use Poisson regression to examine the association between smoking and rates of myocardial infarction in these data. To use a **computer package to fit a Poisson regression model**, it is necessary to specify three items:

1 The *name of the outcome* variable, which in this case is MI. If each line of the dataset represents an individual (as is the case here) then the outcome variable is coded as 1 for individuals who experienced the event and 0 for individuals who did not experience the event. If *data have been grouped according to the values of different exposure variables* then the outcome contains the total number of events in each group.
2 The *total exposure time*, for the individual or the group (depending on whether each line in the dataset represents an individual or a group). As will be explained in Section 24.3, this is used as an **offset** in the Poisson regression model.
3 The *name of the exposure* variable(s). In this example, we have just one exposure variable, which is called *cursmoke*. The *required convention for coding* is that used throughout this book; thus *cursmoke* was coded as 0 for men who were never/ex-smokers at the start of the study (the *unexposed* or *baseline* group) and 1 for men who were current smokers at the start of the study (the *exposed* group).

The **Poisson regression model** that will be fitted is:

$$\text{Rate of myocardial infarction} = \text{Baseline} \times \text{Cursmoke}$$

Its two parameters are:

1 Baseline: the rate of myocardial infarction in the baseline group (never/ex-smokers), and
2 Cursmoke: the rate ratio comparing current smokers with never/ex-smokers.

Output on the ratio scale

Table 24.4 shows the computer output obtained from fitting this model. The two *rows* in the output correspond to the two *parameters* of the logistic regression model; cursmoke is our exposure of interest and the **constant** term refers to the

Table 24.4 Poisson regression output for the model relating rates of myocardial infarction with smoking at the time of recruitment to the Caerphilly study.

	Rate ratio	z	$P > \|z\|$	95% CI
Cursmoke	1.700	4.680	0.000	1.361 to 2.121
Constant	0.00969	−50.37	0.000	0.00809 to 0.0116

baseline group. The same format is used for both parameters, and is based on what makes sense for interpretation of the effect of exposure. This means that some of the information presented for the constant (baseline) parameter is not of interest.

The column labelled 'Rate Ratio' contains the **parameter estimates**:

1 For the first row, labelled 'cursmoke', this is the *rate ratio* (1.700) comparing smokers at recruitment with never/ex-smokers. This is identical to the rate ratio that was calculated directly from the raw data (see Table 24.3).
2 For the second row, labelled 'constant', this is the *rate of myocardial infarction in the baseline group* (0.00969 = 118/12 183, see Table 24.3). As we explained in the context of logistic regression, this apparently inconsistent labelling is because output from regression models is labelled in a uniform way.

The remaining columns present z statistics, P-values and 95% confidence intervals corresponding to the model parameters. They will be explained in more detail after the explanation of Table 24.5 below.

Output on the log scale

Table 24.5 shows Poisson regression output, on the log scale, for the association between smoking and rates of myocardial infarction. The model is:

$$\text{Log(Rate)} = \text{Baseline} + \text{Cursmoke}$$

where

- Baseline is the log rate of myocardial infarction in never/ex-smokers, and
- Cursmoke is the log rate ratio comparing the rate of myocardial infarction in smokers with that in never/ex-smokers.

Table 24.5 Poisson regression output (log scale) for the association between smoking and rates of myocardial infarction.

	Coefficient	s.e.	z	$P > \|z\|$	95% CI
Cursmoke	0.530	0.113	4.680	0.000	0.308 to 0.752
Constant	−4.64	0.092	−50.37	0.000	−4.82 to −4.45

The interpretation of this output is very similar to that described for logistic regression in Chapter 19; readers are referred there for a more detailed discussion of all components of the output.

1 The *first* column gives the results for the **regression coefficients** (corresponding to the parameter estimates on a log scale). For the row labelled 'cursmoke' this is the *log rate ratio* comparing smokers with non-smokers. It agrees with what would be obtained if it were calculated directly from Table 24.3:

$$\text{log rate ratio} = \log(16.98/9.68) = \log(1.70) = 0.530$$

For the row labelled 'constant', the regression coefficient is the *log rate in the baseline group*, i.e. the log rate of myocardial infarction among non-smokers:

$$\text{log rate} = \log(118/12\,183) = \log(0.00969) = -4.637$$

2 The *second* column gives the *standard errors* of the regression coefficients. For a binary exposure variable, these are exactly the same as those derived using the formulae given in Section 23.2. Thus:

$$\text{s.e.(log rate ratio)} = \sqrt{(1/d_1 + 1/d_0)} = \sqrt{(1/118 + 1/230)} = 0.113$$

$$\text{s.e.(log rate in never/ex-smokers)} = \sqrt{(1/d_0)} = \sqrt{(1/118)} = 0.092$$

3 The *95% confidence intervals* for the regression coefficients in the *last* column are derived in the usual way. For the log rate ratio comparing smokers with never/ex-smokers, the 95% CI is:

$$\begin{aligned} 95\%\ \text{CI} &= (0.530 - (1.96 \times 0.113)) \text{ to } (0.530 + (1.96 \times 0.113)) \\ &= 0.308 \text{ to } 0.752 \end{aligned}$$

4 Each *z statistic* in the *third* column is the regression coefficient divided by its standard error. They can be used to derive a Wald test of the null hypothesis that the corresponding regression coefficient = 0.

5 The *P-values* in the *fourth* column are derived from the *z* statistics in the usual manner (see Table A1 in the Appendix) and can be used to test the null hypothesis that the true (population) value for the corresponding population parameter is zero. For example the *P*-value of 0.000 (i.e. < 0.001) for the log rate ratio comparing smokers with never/ex-smokers indicates that there is strong evidence against the null hypothesis that rates of myocardial infarction are the same in smokers as in non-smokers.

As previously explained in the context of logistic regression, we are usually not interested in the *z* statistic and corresponding *P*-value for the *constant* parameter.

Relation between outputs on the ratio and log scales

As with logistic regression, the results in Table 24.4 (output on the original, or ratio, scale) are derived from the results in Table 24.5 (output on the log scale). Once the derivation of the ratio scale output is understood, it is rarely necessary to refer to the log scale output: the most useful results are the rate ratios, confidence intervals and *P*-values displayed on the ratio scale, as in Table 24.4. Note that the output corresponding to the constant term (baseline group) is often omitted from computer output, since the focus of interest is on the parameter estimates (rate ratios) comparing the different groups.

1 In Table 24.4, the column labelled 'Rate Ratio' contains the *exponentials* (antilogs) of the Poisson regression coefficients shown in Table 24.5. Thus the rate ratio comparing smokers with never/ex-smokers = exp(0.530) = 1.700.

2 The *z* statistics and *P*-values are derived from the regression coefficients and their standard errors, and so are identical in the two tables.

3 The 95% confidence intervals in Table 24.4 are derived by antilogging (exponentiating) the confidence intervals on the log scale presented in Table 24.5. Thus the 95% CI for the rate ratio comparing smokers with never/ex-smokers is:

$$95\%\ \text{CI} = \exp(0.308)\ \text{to}\ \exp(0.752) = 1.361\ \text{to}\ 2.121$$

This is identical to the 95% CI calculated using the methods described in Section 23.2.

$$95\%\ \text{CI for rate ratio} = \text{rate ratio/EF to rate ratio} \times \text{EF}$$

where the *error factor* $\text{EF} = \exp(1.96 \times \text{s.e.}(\log\ \text{rate ratio}))$. Note that since the calculations are multiplicative:

$$\frac{\text{Rate ratio}}{\text{Lower confidence limit}} = \frac{\text{Upper confidence limit}}{\text{Rate ratio}}$$

This can be a useful check on confidence limits presented in tables in published papers.

24.3 GENERAL FORM OF THE POISSON REGRESSION MODEL

The general form of the Poisson regression model is similar to that for logistic regression (Section 19.3) and that for multiple regression (Section 11.4). It relates the **log rate** to the values of one or more exposure variables:

$$\log(\text{rate}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

The quantity on the right hand side of the equation is known as the **linear predictor** of the log rate, given the particular value of the p exposure variables x_1 to x_p. The β's are the **regression coefficients** associated with the p exposure variables.

Since $\log(\text{rate}) = \log(d/T) = \log(d) - \log(T)$, the general form of the Poisson regression model can also be expressed as:

$$\log(d) = \log(T) + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

The term $\log(T)$ is known as an **offset** in the regression model. To use statistical packages to fit Poisson regression models we must specify the outcome as the *number* of events and give the exposure time T, which is then included in the offset term, $\log(T)$.

We now show how this general form corresponds to the model we used in Section 24.2 for comparing two exposure groups. The general form for comparing two groups is:

$$\text{Log rate} = \beta_0 + \beta_1 x_1$$

where x_1 (the exposure variable) equals 1 for those in the *exposed* group and 0 for those in the *unexposed* (*baseline*) group.

Using a similar argument to that given in Section 19.3 in the context of logistic regression models, it is straightforward to show that:

1 β_0 (the *intercept*) corresponds to the log rate in the unexposed (baseline) group, and

2 β_1 corresponds to the log of the rate ratio comparing exposed and unexposed groups (the exposure rate ratio).

The equivalent model on the ratio scale is:

$$\text{Rate of disease} = \exp(\beta_0) \times \exp(\beta_1 x_1)$$

In this *multiplicative model* $\exp(\beta_0)$ corresponds to the rate of disease in the baseline group, and $\exp(\beta_1)$ to the exposure rate ratio.

24.4 POISSON REGRESSION FOR CATEGORICAL AND CONTINUOUS EXPOSURE VARIABLES

We now consider Poisson regression models for categorical exposure variables with more than two levels, and for ordered or continuous exposure variables. The principles have already been outlined in detail in Chapter 19, in the context of logistic regression. The application to Poisson regression will be illustrated by

examining the association between social class and rates of myocardial infarction in the Caerphilly study.

Poisson regression to compare more than two exposure groups

To examine the effect of **categorical exposure variables** in Poisson and other regression models, we look at the effect of each level compared to a *baseline* group. This is done using *indicator variables*, which are created automatically by most statistical packages, as explained in more detail in Box 19.1 on page 200.

Example 24.2

In the Caerphilly study, a Poisson regression model was fitted to investigate the evidence that rates of myocardial infarction were higher among men in less privileged social classes. Table 24.6 shows the output, with the social class variable, *socclass*, coded from 1 = social class I (most affluent) to 6 = social class V (most deprived). The model was fitted with social class group III non-manual as the baseline group, since this was the largest group in the study, comprising 925 (51.8 %) of the men. The regression confirms that there is a pattern of increasing rates of myocardial infarction in more deprived social classes. This trend is investigated further in Table 24.7 below.

Note that some statistical computer packages will allow the user to specify which exposure group is to be treated as the baseline group. In other packages, it may be necessary to recode the values of the variable so that the group chosen to be the baseline group has the lowest coded value.

Table 24.6 Poisson regression output for the effect of social class on the rate of myocardial infarction. The model has six parameters: the rate in the baseline group (rate not shown in the table) and the five rate ratios comparing the other groups with this one. It can be written in abbreviated form as: Rate = Baseline × Socclass.

	Rate ratio	z	$P > \|z\|$	95% CI
Socclass(1), I	0.403	−2.36	0.018	0.190 to 0.857
Socclass(2), II	0.759	−1.75	0.080	0.557 to 1.034
Socclass(3), III non-manual	1 (baseline group)			
Socclass(4), III manual	0.956	−0.25	0.802	0.675 to 1.355
Socclass(5), IV	0.965	−0.21	0.836	0.693 to 1.344
Socclass(6), V	1.316	1.14	0.253	0.821 to 2.109

Poisson regression for ordered and continuous exposure variables

Example 24.2 (continued)

To investigate further the tendency for increasing rates of myocardial infarction with increasing deprivation, we can perform a **test for trend** by fitting a Poisson regression model for the linear effect of social class. This will assume a constant

Table 24.7 Poisson regression output for the model for the linear effect of social class on rates of myocardial infarction: Rate = Baseline × [Socclass], where [Socclass] is the rate ratio per unit increase in social class.

	Rate ratio	z	$P > \|z\|$	95% CI
Socclass	1.117	2.411	0.016	1.021 to 1.223

increase in the log rate ratio for each unit increase in social class, and correspondingly a constant rate ratio per increase in social class. The results are shown in Table 24.7. The estimated rate ratio per unit increase in social class is 1.117 (95% CI 1.021 to 1.223, $P = 0.016$). There is some evidence of an association between increasing social deprivation and increasing rates of myocardial infarction.

24.5 POISSON REGRESSION: CONTROLLING FOR CONFOUNDING

Readers are referred to Chapter 20 for a detailed discussion of how regression models control for confounding in a manner that is analogous to the stratification procedure used in Mantel–Haenszel methods. Both methods assume that the true exposure effect comparing exposed with unexposed individuals is the same in each of the strata defined by the levels of the confounding variable.

Example 24.3

In Section 24.4 we found evidence that rates of myocardial infarction in the Caerphilly study increased with increasing social deprivation. There was also a clear association (not shown here) between social class and the prevalence of smoking at the time of recruitment, with higher smoking rates among men of less privileged social classes. It is therefore possible that social class confounds the association between smoking and rates of myocardial infarction. We will examine this using both Mantel–Haenszel and Poisson regression analyses to estimate the rate ratio for smoking after controlling for social class. We will then compare the results.

Table 24.8 shows the rate ratios for smokers compared to non-smokers in strata defined by social class, together with the Mantel–Haenszel estimate of the rate ratio for smoking controlling for social class. This equals 1.65 (95% CI 1.32 to 2.06), only slightly less than the crude rate ratio of 1.70 (see Table 24.4). It appears therefore that social class is not an important confounder of the association between smoking and rates of myocardial infarction.

Table 24.9 shows the output (on the rate ratio scale) from the corresponding Poisson regression. This model *assumes* that the rate ratio for smoking is the same regardless of social class, and (correspondingly) that the rate ratios for social class are the same regardless of smoking. The estimated rate ratio for smoking controlled for social class is 1.645, almost identical to the Mantel–Haenszel estimate (see Table 24.8). There is also little difference between the crude effect of social class (Table 24.6) and the effect of social class controlling for smoking.

Table 24.8 Rate ratios for the association of smoking with rates of myocardial infarction in the Caerphilly study, separately in social class strata, together with the Mantel–Haenszel estimate of the rate ratio for smoking controlling for social class.

Social class stratum	Rate ratio (95% CI) for smokers compared to non-smokers
I (most affluent)	2.07 (0.46 to 9.23)
II	1.49 (0.86 to 2.58)
III non-manual	1.68 (1.23 to 2.30)
III manual	1.38 (0.73 to 2.62)
IV	1.75 (0.91 to 3.35)
V (least affluent)	2.15 (0.77 to 5.96)
Mantel–Haenszel estimate of the rate ratio for smokers compared to non-smokers, controlling for social class	1.65 (1.32 to 2.06)
χ^2 for heterogeneity of rate ratios = 0.82 (d.f. = 5, $P = 0.98$)	

Table 24.9 Poisson regression output for the model including both current smoking and social class. The model can be written in abbreviated form as Rate = Baseline × Cursmoke × Socclass, where the baseline group are non-smokers in Socclass (3).

	Rate ratio	z	$P > \|z\|$	95% CI
Cursmoke	1.645	4.351	0.000	1.315 to 2.058
Socclass(1)	0.445	−2.103	0.035	0.209 to 0.946
Socclass(2)	0.830	−1.176	0.240	0.608 to 1.133
Socclass(4)	1.014	0.075	0.940	0.715 to 1.437
Socclass(5)	0.976	−0.142	0.887	0.701 to 1.359
Socclass(6)	1.333	1.194	0.232	0.832 to 2.136

Note the different forms of the output for the Mantel–Haenszel and Poisson regression approaches. The Mantel–Haenszel output shows us stratum-specific effects of the exposure variable, which draws our attention to differences between strata and reminds us that when we control for smoking we assume that the effect of smoking is the same in different social classes. The Poisson regression output shows us the effect of smoking controlled for social class, and the effect of social class controlled for smoking. However, we should be aware of the need to test the underlying assumption that the effect of each variable is the same regardless of the value of the other: that is that there is no **effect modification**, also known as **interaction**. For Mantel–Haenszel methods this was described in Section 23.3. We see how to examine interaction in regression models in Chapter 29.

24.6 SPLITTING FOLLOW-UP TO ALLOW FOR VARIABLES WHICH CHANGE OVER TIME

In any long-term study the values of one or more of the exposure variables may change over time. The most important such change is in the **age** of subjects in the

study. Since rates of most disease outcomes are strongly associated with age, we will usually wish to control for age in our analysis.

To allow for changes in age, or for any exposure variable whose value changes during the study, we simply divide the follow-up time for each person into distinct periods, during which the variable does not change. Since age, of course, changes constantly we divide the follow-up time into **age groups**. For example, in the Caerphilly study we might use five-year age groups: 40–44, 45–49, 50–54 and so on. Note that age 50–54 means 'from the date of the 50th birthday to the day before the 55th birthday'. The underlying assumption is that rates do not differ much *within* an age group, so that for example it assumes that the rate of myocardial infarction will be similar for a 54-year-old and a 50-year-old. Narrower age bands will be appropriate when rates vary rapidly with age; for example in a study of infant mortality.

Table 24.10 and Figure 24.1 illustrate the division of the follow-up period into 5-year age bands for subject numbers 1 and 2 in the Caerphilly dataset. Subject 1 was aged 58.52 years when he was recruited, and therefore started in the 55–59 age group. He passed through the 60–64, 65–69 and 70–74 age groups, and was in the 75–79 age group at the end of the study (at which time he was aged 75.36). Subject 2 was also in the 55–59 age group when he was recruited. He was in the 60–64 age group when he experienced a myocardial infarction on 27 Feb 1985, at which time he was aged 61.81.

It is important to note that the value of MI (myocardial infarction, the outcome variable) is equal to 0 for every interval unless the subject experienced an MI at the end of the interval, in which case it is 1. Thus for subject 1, the value of MI is 0 for every interval, and for subject 2 it is 0 for the first interval and 1 for the second interval. In general, the value of the outcome variable for a subject who experienced the outcome will be zero for every interval except the last.

Having divided the follow-up time in this way, we may now use **Mantel–Haenszel** or **Poisson regression** methods to examine the way in which disease rates change with age group, or to examine the effects of other exposures having

Table 24.10 Follow-up time split into 5-year age bands for the first two subjects in the Caerphilly study.

Date at start of interval	Date at end of interval	Age group	Age at start of interval	Age at end of interval	Years in interval	MI
Subject 1, born 22 Aug 1923, recruited 1 Mar 1982, exit (at end of follow-up) 31 Dec 1998						
1 Mar 1982	21 Aug 1983	55–59	58.52	60	1.48	0
22 Aug 1983	21 Aug 1988	60–64	60	65	5	0
22 Aug 1988	21 Aug 1993	65–69	65	70	5	0
22 Aug 1993	21 Aug 1998	70–74	70	75	5	0
22 Aug 1998	31 Dec 1998	75–79	75	75.36	0.36	0
Subject 2, born 8 May 1923, recruited 30 May 1982, exit (on date of MI) 27 Feb 1985						
30 May 1982	7 May 1983	55–59	59.06	60	0.94	0
8 May 1983	27 Feb 1985	60–64	60	61.81	1.81	1

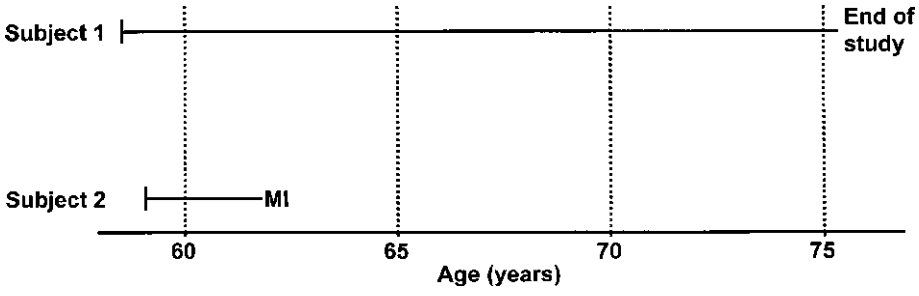

Fig. 24.1 Age of subjects 1 and 2 during the Caerphilly study. The dotted vertical lines denote 5-year age bands.

controlled for the effects of age group. Perhaps surprisingly, we analyse the contributions from the different time periods from the same individual in exactly the same way as if they were from different individuals. See Clayton and Hills (1993) for the reasons why this is justified. Also, note that if we analyse this expanded data set (with follow-up split into age groups) but omit age group from the analysis we will get exactly the same answer as in the analysis using the original intervals. This is because the number of events and the total follow-up time are exactly the same in the original and expanded datasets.

Table 24.11 shows the total number of events (d) and person-years (T) in the different age groups in the Caerphilly study, together with the rates per 1000 person-years and corresponding 95% confidence intervals. Rates of myocardial infarction generally increased with increasing age.

Table 24.11 Rates of myocardial infarction in different age groups in the Caerphilly study.

Age group	d	T	Rate per 1000 person-years	95% CI
45–49	12	1 627	7.376	4.189 to 12.989
50–54	42	4 271	9.833	7.267 to 13.305
55–59	73	6 723	10.858	8.632 to 13.657
60–64	102	7 115	14.336	11.807 to 17.406
65–69	76	4 287	17.726	14.157 to 22.195
70–74	30	1 872	16.029	11.207 to 22.926
75–79	13	266	48.958	28.428 to 84.315

This same approach may be used to examine any effect that may change over time. For example:

- if **repeat measurements of exposures** are made on different occasions after baseline, we may divide the follow-up time into the periods following each measurement, with *time-updated* values of the exposure measured at the beginning of each period.
- **secular changes** can be analysed by dividing time into different **time periods** (for example, 1970 to 1974, 1975 to 1979, etc.).

Joint effects may be investigated by dividing the period of follow-up according to the values of two variables. Note that the way in which individuals move through different categories of age group and time period may be displayed in a **Lexis diagram** (see Clayton and Hills, 1993 or Szklo and Nieto, 2000).

In Section 27.5, we explain how Poisson regression with follow-up time split into intervals is related to **Cox regression** analysis of survival data, and in Section 27.4 we discuss the criteria for choice of the time axis.

CHAPTER 25

Standardization

25.1 INTRODUCTION

Death rates and disease incidence rates are usually strongly related to age, and often differ for the two sexes. Population mortality and incidence rates therefore depend critically on the age–sex composition of the population. For example, a relatively older population would have a higher crude mortality rate than a younger population even if, age-for-age, the rates were the same. It is therefore misleading to use overall rates when comparing two different populations unless they have the same age–sex structure. We saw in Chapter 23 how to use Mantel–Haenszel methods and in Chapter 24 how to use Poisson regression to compare rates between different groups after controlling for variables such as age and sex.

We now describe the use of **standardization** and **standardized rates** to produce comparable measures between populations or sub-groups, adjusted for major confounders, such as any age–sex differences in the composition of the different populations or subgroups. **Mantel–Haenszel** or **regression methods** should be used to make formal comparisons between them.

There are two methods of standardization: *direct* and *indirect*, as summarized in Table 25.1. Both use a **standard population**.

Table 25.1 Comparison of direct and indirect methods of standardization.

	Direct standardization	Indirect standardization
Method	Study rates applied to standard population	Standard rates applied to study population
Data required		
Study population(s)	Age–sex specific rates	Age–sex composition + total deaths (or cases)
Standard population	Age–sex composition	Age–sex specific rates (+ overall rate)
Result	Age–sex adjusted rate	Standardized mortality (morbidity) ratio (+ age–sex adjusted rate)

- In **direct standardization**, the age–sex specific rates from each of the populations under study are applied to a *standard* population. The result is a set of **standardized rates**.
- In **indirect standardization**, the age–sex specific rates from a *standard* population are applied to each of the study populations. The result is a set of **standardized mortality (or morbidity) ratios (SMRs)**.

The choice of method is usually governed by the availability of data and by their (relative) accuracy. Thus, direct standardization gives more accurate results when there are small numbers of events in any of the age–sex groups of the study populations. The indirect method will be preferable if it is difficult to obtain national data on age–sex specific rates.

Both methods can be extended to adjust for other factors besides age and sex, such as different ethnic compositions of the study groups. The direct method can also be used to calculate **standardized means**, such as age–sex adjusted mean blood pressure levels for different occupational groups.

25.2 DIRECT STANDARDIZATION

Example 25.1

Table 25.2 shows the number of cases of prostate cancer and number of person-years among men aged ≥ 65 living in France between 1979 and 1996. The data are shown separately for six 3-year time periods. Corresponding rates of prostate cancer per 1000 person-years at risk (pyar) are shown in Table 25.3

Table 25.3 shows that the crude rates (those derived from the total number of cases and person-years, ignoring age group) increased to a peak of 2.64/1000 pyar in 1988–90 and then declined. However Table 25.2 shows that the age-distribution of the population was also changing during this time: the number of person-years in the oldest (≥ 85 year) age group more than doubled between 1979–81 and 1994–96, while increases in other age groups were more modest. The oldest age group also experienced the highest rate of prostate cancer, in all time periods.

Table 25.2 Cases of prostate cancer/1000 person-years among men aged ≥ 65 living in France between 1979 and 1996.

	Time period					
Age group	1979–81	1982–84	1985–87	1988–90	1991–93	1994–96
65–69	2021/2970	1555/2197	1930/2686	2651/3589	2551/3666	2442/3764
70–74	3924/2640	3946/2674	3634/2272	2842/1860	3863/2703	4158/3177
75–79	5297/1886	5638/1946	6018/1980	6211/2028	4640/1598	4253/1659
80–84	4611/985	5400/1134	6199/1189	6844/1294	6926/1393	6412/1347
≥ 85	3273/478	3812/539	4946/616	6581/764	7680/878	8819/1003
Total	19126/8959	20351/8490	22727/8743	25129/9535	25660/10238	26084/10950

Table 25.3 Rates of prostate cancer (per 1000 person-years) in men aged ≥ 65 living in France between 1979 and 1996.

	Time period					
Age group	1979–81	1982–84	1985–87	1988–90	1991–93	1994–96
65–69	0.68	0.71	0.72	0.74	0.70	0.65
70–74	1.49	1.48	1.60	1.53	1.43	1.31
75–79	2.81	2.90	3.04	3.06	2.90	2.56
80–84	4.68	4.76	5.21	5.29	4.97	4.76
≥ 85	6.85	7.07	8.03	8.61	8.75	8.79
Crude rate	2.13	2.40	2.60	2.64	2.51	2.38
Standardized rate	2.35	2.40	2.60	2.64	2.54	2.39

This means that the overall rates in each time period need to be adjusted for the age distribution of the corresponding population before they can meaningfully be compared. We will do this using the method of direct standardization.

1 The first step in direct standardization is to identify a standard population. This is usually one of the following:
- one of the study populations
- the total of the study populations
- the census population from the local area or country

The choice is to some extent arbitrary. Different choices lead to different summary rates but this is unlikely to affect the interpretation of the results unless the patterns of change are different in the different age group strata (see point 5). Here we will use the number of person-years for the period 1985–87.

2 Second, for *each of the time periods of interest*, we calculate what would be the overall rate of prostate cancer in our standard population if the age-specific rates equalled those of the time period of interest. This is called the *age standardized survival rate* for that time period.

$$\text{Age standardized rate} = \begin{array}{l}\text{Overall rate in standard population}\\ \text{if the age-specific rates were the same}\\ \text{as those of the population of interest}\end{array} = \frac{\Sigma(w_i \times \lambda_i)}{\Sigma w_i}$$

In the above definition, w_i is the person-years at risk in age group i in the standard population, $\lambda_i = d_i/pyar_i$ is the rate in age group i in the time period of interest and the summation is over all age groups. Note that this is simply a **weighted average** (see Section 18.3) of the rates in the different age groups in the time period of interest, weighted by the person-years at risk in each age group in the standard population.

Table 25.4 Calculating the age standardized rate of prostate cancer for 1979–81, using direct standardization with the person-years during 1985–87 as the standard population.

Age group	Standard population: thousands of person-years in 1985–87, w_i	Study population: Rates in 1979–81, λ_i	Estimated number of cases in standard population, $w_i \times \lambda_i$
65–69	2686	0.6805	1827.8
70–74	2272	1.4864	3377.1
75–79	1980	2.8086	5561.0
80–84	1189	4.6812	5565.9
≥ 85	616	6.8473	4217.9
All ages	$\Sigma w_i = 8743$		$\Sigma(w_i \times \lambda_i) = 20549.8$
		Age adjusted rate = 2.35	

For example, Table 25.4 shows the details of the calculations for the age-standardized rate for 1979–81, using the person-years in 1985–87 as the standard population. In the 65 to 69-year age group, applying the rate of 0.6805 per 1000 person-years to the 2686 person-years in that age group in the standard population gives an estimated number of cases in this age group of $0.6805 \times 2686 = 1827.8$. Repeating the same procedure for each age group, and then summing the numbers obtained, gives an overall estimate of 20549.8 cases out of the total of 8743 thousand person-years in the *standard* population: an age-standardized rate for the *study* population of 2.35 per 1000 person-years.

3 The results for all the time periods are shown in the bottom row of Table 25.3. The crude and standardized rates of prostate cancer in the different time periods are plotted in Figure 25.1(a). This shows that the crude rate was lower than the directly standardized rate in the 1979–81 period, but similar thereafter. This is because, as can be seen in Table 25.2, in the 1979–81 period there were proportionally fewer person-years in the oldest age groups, in which prostate cancer death rates were highest.

4 The standard error for the standardized rate is calculated as:

Standard error of standardized rate	Standard error of standardized proportion
$\frac{1}{\Sigma w_i}\sqrt{\left(\sum \frac{w_i^2 d_i}{(pyar_i)^2}\right)}$	$\frac{1}{\Sigma w_i}\sqrt{\left(\sum \frac{w_i^2 p_i(1-p_i)}{n_i}\right)}$

where the left hand formula is used for standardized rates and the right hand formula for standardized proportions. In these formulae the weights w_i are the person-years or number of individuals in the standard population. Using this formula, the standard error of the standardized rate in 1979–81 is 0.017 per 1000 person-years, so that the 95% confidence interval for the standardized rate in 1979–81 is:

$$95\%\ \text{CI} = 2.35 - 1.96 \times 0.017 \text{ to } 2.35 + 1.96 \times 0.017$$
$$= 2.32 \text{ to } 2.38 \text{ per } 1000 \text{ person-years}$$

5 Finally, it is important to inspect the patterns of rates in the individual strata before standardizing, because when we standardize we assume *that the patterns of change in the rates are similar in each stratum*. If this is not the case then the choice of standard population will influence the observed pattern of change in the standardized rates. For example, in Figure 25.1(b) it can be seen that the rate in the ≥ 85 year age group increased more sharply than the rates in the other age groups. This means that the greater the proportion of individuals in the ≥ 85 year age group in the standard population, the sharper will be the increase in the standardized rate over time.

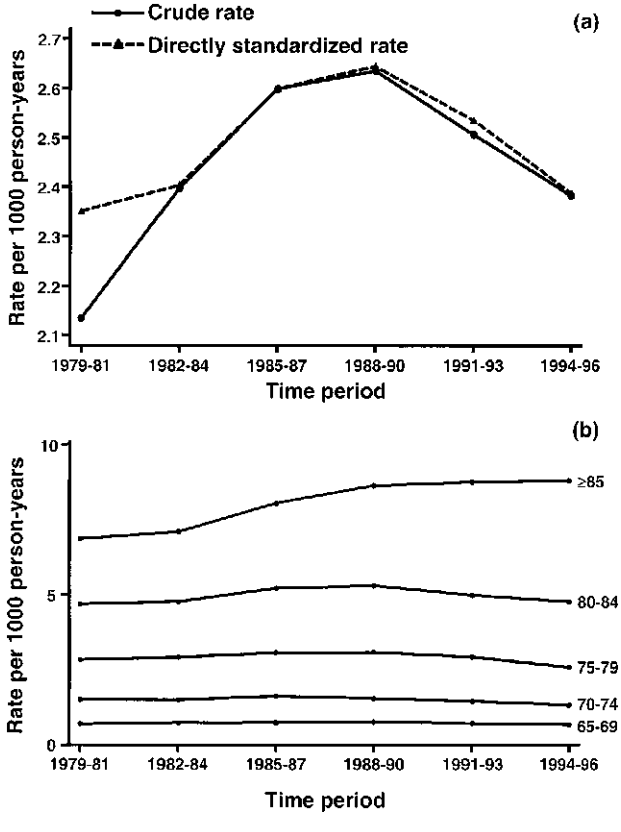

Fig. 25.1 (a) Crude and directly standardized rates of prostate cancer among men aged $\geq$ 65 years living in France between 1979 and 1986, with the population in 1985–87 chosen as the standard population. (b) Time trends in age-specific rates of prostate cancer, among men aged $\geq$ 65 years living in France between 1979 and 1986.

25.3 INDIRECT STANDARDIZATION

Example 25.2

Table 25.5 shows mortality rates from a large one-year study in an area endemic for onchocerciasis. One feature of interest was to assess whether blindness, the severest consequence of onchocerciasis, leads to increased death rates. From the results presented in Table 25.5 it can be seen that:

- not only does mortality increase with age and differ slightly between males and females, but
- the prevalence of blindness also increases with age and is higher for males than for females.

The blind sub-population is therefore on average older, with a higher proportion of males, than the non-blind sub-population. This means that it would have a higher crude mortality rate than the non-blind sub-population, even if the individual age–sex specific rates were the same. An overall comparison between the blind and non-blind will be obtained using the method of indirect standardization.

1 As for direct standardization, the *first* step is to identify a standard population. The usual choices are as before, with the restrictions that age–sex specific mortality rates are needed for the standard population and that the population chosen for this should therefore be large enough to have reliable estimates of these rates. In this example the rates among the non-blind will be used.

2 These standard rates are then applied to the population of interest to calculate the *number of deaths* that would have been *expected* in this population if the mortality experience were the same as that in the standard population.

For example, in stratum 1 (males aged 30–39 years) one would expect a proportion of 19/2400 of the 120 blind to die, if their risk of dying was the same as that of the non-blind males of similar age. This gives an expected 0.95 deaths for this age group. In total, 22.55 deaths would have been expected among the blind compared to a total observed number of 69.

3 The ratio of the observed to the expected number of deaths is called the **standardized mortality ratio (SMR)**. It equals 3.1 (69/22.55) in this case. Overall, blind persons were 3.1 times more likely to die during the year than non-blind persons.

$$\text{Standardized mortality ratio (SMR)} = \frac{\text{observed number of deaths}}{\text{expected number of deaths if the age–sex specific rates were the same as those of the } \textit{standard} \text{ population}} = \frac{\Sigma d_i}{\Sigma E_i}$$

The SMR measures how much more (or less) likely a person is to die in the study population compared to someone of the same age and sex in the standard population. A value of 1 means that they are equally likely to die, a value larger

Table 25.5 Use of indirect standardization to compare mortality rates between the blind and non-blind, collected as part of a one-year study in an area endemic for onchocerciasis. The mortality rates among the non-blind have been used as the standard rates.

		Non-blind persons			Blind persons				Expected number of deaths among blind if rates were the same as those of the non-blind ($E_i = \lambda_i \times T_i$)
Age (yrs)	Stratum (i)	Number of person-years	Number of deaths	Deaths/1000/yr (λ_i)	% blind	Number of person-years (T_i)	Number of deaths (d_i)	Deaths/1000/yr	
Males									
30–39	1	2400	19	7.9	4.8	120	3	25.0	0.95
40–49	2	1590	21	13.2	9.7	171	7	40.9	2.26
50–59	3	1120	20	17.9	17.9	244	13	53.3	4.36
60+	4	610	20	32.8	28.0	237	24	101.3	7.77
Females									
30–39	5	3100	23	7.4	2.6	84	2	23.8	0.62
40–49	6	1610	22	13.7	4.1	69	3	43.5	0.94
50–59	7	930	16	17.2	15.3	168	8	47.6	2.89
60+	8	270	8	29.6	25.6	93	9	96.8	2.76
Total		11630	149	12.8	9.3		69	58.2	22.55
SMR				1.0				3.1 (69/22.5)	
Age–sex adjusted mortality rate				12.8				39.7 (3.1 × 12.8)	

than 1 that they are more likely to die, and a value smaller than 1 that they are less likely to do so. The SMR is sometimes multiplied by 100 and expressed as a percentage. Since the non-blind population was used as the standard, its expected and observed numbers of deaths are equal, resulting in an SMR of 1.

4 The 95% confidence interval for the SMR is derived using an error factor (EF) in the same way as that for a rate ratio (see Section 23.2):

$$95\%\ \text{CI} = \text{SMR}/\text{EF}\ \text{to}\ \text{SMR} \times \text{EF},\ \text{where}\ \text{EF} = \exp(1.96/\sqrt{d_i})$$

In this example, $\text{EF} = \exp(1.96/\sqrt{69}) = 1.266$, and the 95% confidence interval for the SMR is:

$$95\%\ \text{CI} = \frac{\text{SMR}}{\text{EF}}\ \text{to}\ \text{SMR} \times \text{EF} = 3.06/1.266\ \text{to}\ 3.06 \times 1.266 = 2.42\ \text{to}\ 3.87$$

5 **Age–sex adjusted mortality rates** may be obtained by multiplying the SMRs by the crude mortality rate of the standard population, when this is known. This gives age–sex adjusted mortality rates of 12.8 and 39.7/1000/year for the non-blind and blind populations respectively.

$$\text{Age–sex adjusted mortality rate} = \text{SMR} \times \text{crude mortality rate of standard population}$$

25.4 USE OF POISSON REGRESSION FOR INDIRECT STANDARDIZATION

We may use Poisson regression to derive the SMR, by fitting a model with:

- each row of data corresponding to the strata in the study population;
- the number of events in the study population as the outcome. In Example 25.2 this would be the number of deaths in the blind population;
- *no* exposure variables (a 'constant-only' model);
- specifying the *expected number* of events in each stratum (each row of the data), instead of the number of person-years, as the offset in the model. In Example 25.2, these are the expected number of deaths given in the right hand column of Table 25.5.

Table 25.6 shows the output from fitting such a model to the data in Example 25.2. The output is on the log scale, so the SMR is calculated by antilogging the

Table 25.6 Poisson regression output (log scale), using the expected number of deaths in the blind population as the offset.

	Coefficient	s.e.	z	$P > \|z\|$	95% CI
Constant	1.1185	0.1204	9.29	0.000	0.8825 to 1.3544

coefficient for the constant term. It equals exp(1.1185) = 3.1, the same as the value calculated above.

$$\text{SMR} = \exp(\text{regression coefficient for constant term})$$

The 95% CI for the SMR is derived by antilogging the confidence interval for the constant term. It is exp(0.8825) to exp(1.3544) = 2.42 to 3.87. It should be noted that indirect standardization assumes that the age–sex specific rates in the standard population are known without error. Clearly this is not true in the example we have used: the consequence of this is that confidence intervals for the SMR derived in this way will be somewhat too narrow. For comparison, a standard Poisson regression analysis of the association between blindness and death rates for the data in Table 25.5 gives a rate ratio of 3.05, and a 95% CI of 2.24 to 4.15.

Extension to several SMRs

It is fairly straightforward to extend this procedure to estimate, for example, the SMRs for each area in a geographical region by calculating the observed and expected number of deaths in each age–sex stratum in each area, and fitting a Poisson regression model including indicator variables for each area, and *omitting* the constant term. The SMRs would then be the antilogs of the coefficients for the different area indicator variables.

CHAPTER 26

Survival analysis: displaying and comparing survival patterns

26.1 INTRODUCTION

The methods described so far in this part of the book assume that rates are constant over the period of study, or within time periods such as age groups defined by splitting follow-up time as described in Section 24.6. However, in longitudinal studies in which there is a clear event from which subjects are followed, such as diagnosis of a condition or initiation of treatment, it may not be reasonable to assume that rates are constant, even over short periods of time. For example:

- the risk of death is very high immediately after heart surgery, falls as the patient recovers, then rises again over time;
- the recurrence rate of tumours, following diagnosis and treatment of breast cancer, varies considerably with time.

Methods for **survival analysis** allow analysis of such rates without making the assumption that they are constant. They focus on:

1 the **hazard** $h(t)$: the instantaneous rate at time t. They do not assume that the hazard is constant within time periods;
2 the **survivor function** $S(t)$, illustrated by the **survival curve**. This is the probability that an individual will survive (i.e. has not experienced the event of interest) up to and including time t.

We start by describing two ways of estimating the survival curve; life tables and the Kaplan–Meier method. We will then explain the proportional hazards assumption, and discuss how to compare the survival of two groups using Mantel–Cox methods. In the next chapter we will discuss regression analysis of survival

data. We will see that these methods are closely related to, and often give similar results to, the Mantel–Haenszel and Poisson regression methods for the analysis of rates.

In Chapter 22 we stated that survival times for subjects who are known to have survived up to a certain point in time, but whose survival status past that point is not known, are said to be **censored**. Throughout this and the next chapter we will assume that the probability of being censored (either through loss to follow-up or because of death from causes other than the one being studied) is unrelated to the probability that the event of interest occurs. If this assumption is violated then we say that there is **informative censoring**, and special methods must be used.

26.2 LIFE TABLES

Life tables are used to display the survival pattern of a community when we do not know the exact survival time of each individual, but we do know the number of individuals who survive at a succession of time points. They may take one of two different forms. The first, a *cohort life table*, shows the actual survival of a group of individuals through time. The starting point from which the survival time is measured may be birth, or it may be some other event. For example, a cohort life table may be used to show the mortality experience of an occupational group according to length of employment in the occupation, or the survival pattern of patients following a treatment, such as radiotherapy for small-cell carcinoma of bronchus (Table 26.1). The second type of life table, a *current life table*, shows the expected survivorship through time of a hypothetical population to which current age-specific death rates have been applied. Historically, this was more often used for actuarial purposes and was less common in medical research. In recent times, this approach has been used to model the burden of disease due to different causes and conditions (Murray & Lopez, 1996).

Example 26.1

Table 26.1 shows the survival of patients with small-cell carcinoma of bronchus, month by month following treatment with radiotherapy. This table is based on data collected from a total of 240 patients over a 5 year period. The data themselves are summarized in columns 1–4 of the life table; the construction of a **cohort life table** is shown in columns 5–8.

Column 1 shows the number of months since treatment with radiotherapy began. Columns 2 and 3 contain the number of patients alive at the beginning of each month and the number who died during the month. For example, 12 of the 240 patients died during the first month of treatment, leaving 228 still alive at the start of the second month. The number of patients who were censored during each month (known to have survived up to month i but lost to follow-up after that time) is shown in column 4. The total number of persons at risk of dying during the month, adjusting for these losses, is shown in column 5. This equals the

Table 26.1 Life table showing the survival pattern of 240 patients with small-cell carcinoma of bronchus treated with radiotherapy.

(1) Interval (months) since start of treatment i	(2) Number alive at beginning of interval a_i	(3) Deaths during interval d_i	(4) Number censored (lost to follow-up) during interval c_i	(5) Number of persons at risk $n_i = a_i - c_i/2$	(6) Risk of dying during interval $r_i = d_i/n_i$	(7) Chance of surviving interval $s_i = 1 - r_i$	(8) Cumulative chance of survival from start of treatment $S(i) = S(i-1) \times s_i$
1	240	12	0	240.0	0.0500	0.9500	0.9500
2	228	9	0	228.0	0.0395	0.9605	0.9125
3	219	17	1	218.5	0.0778	0.9222	0.8415
4	201	36	4	199.0	0.1809	0.8191	0.6893
5	161	6	2	160.0	0.0375	0.9625	0.6634
6	153	18	7	149.5	0.1204	0.8796	0.5835
7	128	13	5	125.5	0.1036	0.8964	0.5231
8	110	11	3	108.5	0.1014	0.8986	0.4700
9	96	14	3	94.5	0.1481	0.8519	0.4004
10	79	13	0	79.0	0.1646	0.8354	0.3345
11	66	15	4	64.0	0.2344	0.7656	0.2561
12	47	6	1	46.5	0.1290	0.8710	0.2231
13	40	6	0	40.0	0.1500	0.8500	0.1896
14	34	4	2	33.0	0.1212	0.8788	0.1666
15	28	5	0	28.0	0.1786	0.8214	0.1369
16	23	7	1	22.5	0.3111	0.6889	0.0943
17	15	12	0	15.0	0.8000	0.2000	0.0189
18	3	3	0	3.0	1.0000	0.0000	0.0000

number alive at the beginning of the month minus half the number lost to follow-up, assuming that on average these losses occur half-way through the month.

Column 6 shows the risk of dying during a month, calculated as the number of deaths during the month divided by the number of persons at risk. Column 7 contains the complementary chance of surviving the month.

Column 8 shows the cumulative chance of surviving. This is calculated by applying the rules of conditional probability (see Chapter 14). It equals the chance of surviving up to the end of the previous month, multiplied by the chance of surviving the month. For example, the chance of surviving the first month was 0.9500. During the second month the chance of surviving was 0.9605. The overall chance of surviving two months from the start of treatment was therefore $0.9500 \times 0.9605 = 0.9125$. In this study all the patients had died by the end of 18 months.

More generally, the cumulative chance of surviving to the end of month i is given by:

$$S(i) = \text{chance of surviving to month } (i-1) \times \text{chance of surviving month } i$$
$$= S(i-1) \times s_i \text{ or } s_1 \times s_2 \times \ldots \times s_i$$

These are the **probabilities *S(i)* of the survivor function**. The survival curve is illustrated in Figure 26.1.

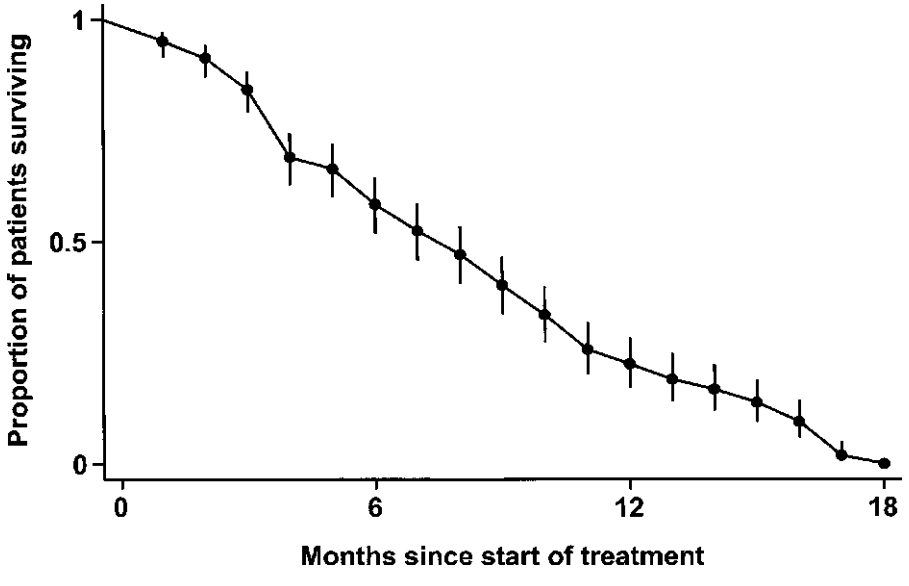

Fig. 26.1 Survival curve for patients with small-cell carcinoma of the bronchus treated with radiotherapy, drawn from life table calculations presented in Table 26.1.

Confidence interval for the survival curve

The 95% confidence interval for each $S(i)$ is derived using an error factor (see Kalbfleisch & Prentice, 1980, pp. 14, 15 for details) as follows:

$$95\%\ \text{CI} = S(t)^{(1/\text{EF})} \text{ to } S(t)^{\text{EF}}, \text{ where}$$
$$\text{EF} = \exp\left[1.96 \times \frac{\sqrt{[\Sigma d/(n(n-d))]}}{[\Sigma \log((n-d)/n)]^2}\right]$$

In this formula, the summations are over all the values of d and n, up to and including time interval i. Figure 26.1 includes the 95% confidence intervals calculated in this way, using the data in columns 3 and 5 of Table 26.1. Because derivation of such confidence intervals involves a substantial amount of calculation, it is usually done using a statistical computer package.

Life expectancy

Also of interest is the average length of survival, or **life expectancy**, following the start of treatment. This may be crudely estimated from the survival curve as the time corresponding to a cumulative probability of survival of 0.5, or it may be calculated using columns 1 and 8 of the life table. For each interval, the length of the interval is multiplied by the cumulative chance of surviving. The total of these values plus a half gives the life expectancy. (The addition of a half is to allow for the effect of grouping the life table in whole months and is similar to the continuity corrections we have encountered in earlier chapters.)

$$\text{Life expectancy} = 0.5 + \sum \left(\begin{matrix}\text{length of} \\ \text{interval}\end{matrix} \times \begin{matrix}\text{cumulative chance} \\ \text{of survival}\end{matrix} \right)$$

In Table 26.1 all the intervals are of 1 month and so the life expectancy is simply the sum of the values in column 8 plus a half, which equals 7.95 months.

26.3 KAPLAN–MEIER ESTIMATE OF THE SURVIVAL CURVE

In many studies we know the exact follow up time (for example, to within 1 day) for each individual in the study, and may therefore wish to estimate the survivor function $S(t)$ using this information rather than by dividing the survival time into discrete periods, as is done in the life table method. This avoids the assumption that individuals lost to follow-up are censored half way through the interval. The

difference between the approaches is likely to be minimal if the periods in the life table are short, such as 1 month, but for longer periods (such as 1 year) information is likely to be lost by grouping.

The estimate using exact failure and censoring times is known as the **Kaplan–Meier** estimate, and is based on a similar argument to that used in deriving life tables. To derive the Kaplan–Meier estimate, we consider the **risk sets** of individuals still being studied at *each time, t, at which an event occurs*. If there are n_t individuals in the risk set at time t, and d_t events occur at that precise time then the estimated *risk*, r_t, of an event at time t is d_t/n_t, and so the estimated **survival probability at time *t*** is:

$$s_t = 1 - r_t = \frac{n_t - d_t}{n_t}$$

At *all* times at which no event occurs, the estimated survival probability is 1.

To estimate the survivor function, we use a similar conditional probability argument to that used in deriving life tables. We number the times at which disease events occur as t_1, t_2, t_3 and so on. Since the estimated survival probability until just before t_1 is 1:

$$S(t_1) = 1 \times s_{t_1} = s_{t_1}$$

The survival probability remains unchanged until the next disease event, at time t_2. The survivor function at this time t_2 is:

$$S(t_2) = S(t_1) \times s_{t_2} = s_{t_1} \times s_{t_2}$$

In general, the survival probability up to and including event j is:

$$S(t_j) = S(t_{(j-1)}) \times s_{t_j} = s_{t_1} \times s_{t_2} \times \ldots \times s_{t_j}$$

This is known as the **product-limit formula**. Note that loss to follow-up does not affect the estimate of survival probability: the next survival probability is calculated on the basis of the new denominator, reduced by the number of subjects lost to follow-up since the last event.

Example 26.2

The examples for the rest of this chapter are based on data from a randomized trial (see Chapter 34) of Azathioprine for primary biliary cirrhosis, a chronic and eventually fatal liver disease (Christensen *et al.*, 1985). The trial was designed to

compare an active treatment, Azathioprine, against placebo. Between October 1971 and December 1977, 248 patients were entered into the trial and followed for up to 12 years. A total of 184 patients had the values of all prognostic variables measured at baseline. Of these, 31 had central cholestasis (a marker of disease severity) at entry. Among these 31 patients there were 24 deaths, and 7 losses to follow-up, as shown in Table 26.2.

The first death was at 19 days, so the risk of death at 19 days was $r_{19} = 1/31 = 0.0323$. The survival probability at 19 days is therefore $s_{19} = 1 - 0.0323 = 0.9677$, and the survivor function $S(19) = s_{19} = 0.9677$. The next death was at 48 days; at this point 30 patients were still at risk. The risk of death at 48 days was $r_{48} = 1/30 = 0.0333$. The survival probability at 48 days is therefore $s_{48} = 1 - 0.0333 = 0.9667$, and the survivor function $S(48) = s_{19} \times s_{48} = 0.9355$. Similarly, the estimate of the survivor function at 96 days is $s_{19} \times s_{48} \times s_{96} = 0.9677 \times 0.9667 \times 0.9655 = 0.9032$, and so on.

Displaying the Kaplan–Meier estimate of *S(t)*

The conventional display of the Kaplan–Meier estimate of the survival curve for the 31 patients with central cholestasis is shown in Figure 26.2. The survival curve is shown as a **step function**; the curve is horizontal at all times at which there is no outcome event, with a vertical drop corresponding to the change in the survivor function at each time when an event occurs. At the right-hand end of the curve, when there are very few patients still at risk, the times between events and the drops in the survivor function become large, because the estimated risk ($r_t = d_t/n_t$) is large at each time t at which an event occurs, as n_t is small. The survivor function should be interpreted cautiously when few patients remain at risk.

Confidence interval for the survival curve

Confidence intervals for $S(t)$ are derived in the same way as described earlier for life tables.

26.4 COMPARISON OF HAZARDS: THE PROPORTIONAL HAZARDS ASSUMPTION

The main focus of interest in survival analysis is in comparing the survival patterns of different groups. For example, Figure 26.3 shows the Kaplan–Meier estimates of the survivor functions for the two groups of patients with and without central cholestasis at baseline. It seems clear that survival times for patients without central cholestasis at baseline were much longer, but how should we quantify the difference in survival? The differences between the survival curves are obviously not constant. For example both curves start at 1, but never come together

Table 26.2 Derivation of the Kaplan–Meier estimate of the survivor function *S(t)*, for 31 patients with primary biliary cirrhosis complicated by central cholestasis. Analyses of this study are by kind permission of Dr E. Christensen.

Time (days) t	Number at risk at time of event(s) n_t	Number of deaths at time t d_t	Number lost to follow-up at time t c_t	Risk of death $r_t = d_t/n_t$	Probability of survival $s_t = 1 - r_t$	Survivor function $S(t) = S(t_{previous}) \times s_t$
19	31	1	0	0.0323	0.9677	0.9677
48	30	1	0	0.0333	0.9667	0.9355
96	29	1	0	0.0345	0.9655	0.9032
150	28	1	0	0.0357	0.9643	0.8710
177	27	1	0	0.0370	0.9630	0.8387
193	26	1	0	0.0385	0.9615	0.8065
201	25	1	0	0.0400	0.9600	0.7742
245	24	1	0	0.0417	0.9583	0.7419
251	23	1	0	0.0435	0.9565	0.7097
256	22	1	0	0.0455	0.9545	0.6774
302	21	0	1	0	1	0.6774
341	20	1	0	0.0500	0.9500	0.6435
395	19	1	0	0.0526	0.9474	0.6097
421	18	1	0	0.0556	0.9444	0.5758
464	17	1	0	0.0588	0.9412	0.5419
578	16	1	0	0.0625	0.9375	0.5081
582	15	0	1	0	1	0.5081
586	14	0	1	0	1	0.5081
688	13	1	0	0.0769	0.9231	0.4690
828	12	0	1	0	1	0.4690
947	11	1	0	0.0909	0.9091	0.4263
1159	10	0	1	0	1	0.4263
1219	9	1	0	0.1111	0.8889	0.3790
1268	8	1	0	0.1250	0.8750	0.3316
1292	7	0	1	0	1	0.3316
1693	6	1	0	0.1667	0.8333	0.2763
1881	5	1	0	0.2000	0.8000	0.2211
1940	4	1	0	0.2500	0.7500	0.1658
1975	3	1	0	0.3333	0.6667	0.1105
2338	2	0	1	0	1	0.1105
2343	1	1	0	1	0	0

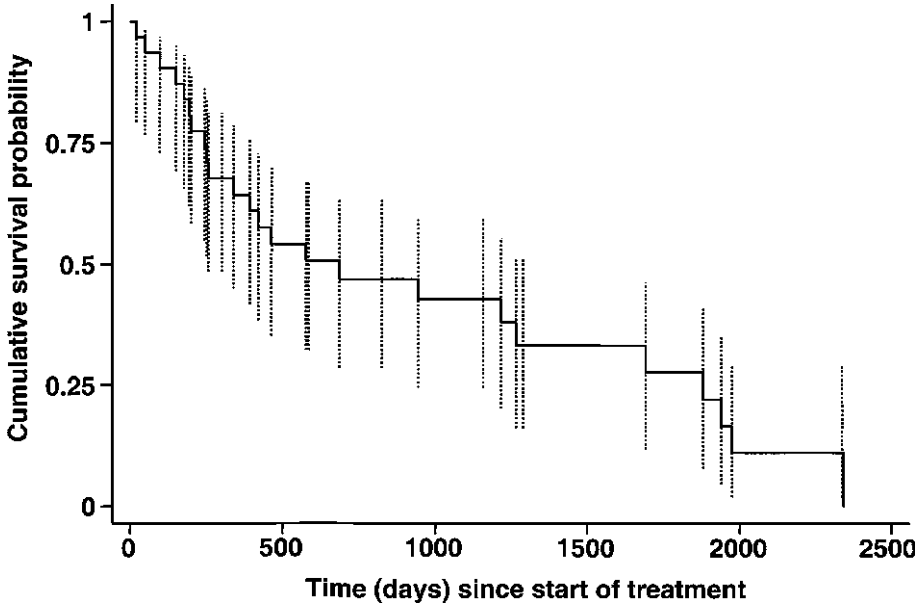

Fig. 26.2 The Kaplan–Meier estimate of the survivor function, *S(t)*, together with upper and lower confidence limits, for 31 patients with primary biliary cirrhosis and central cholestasis.

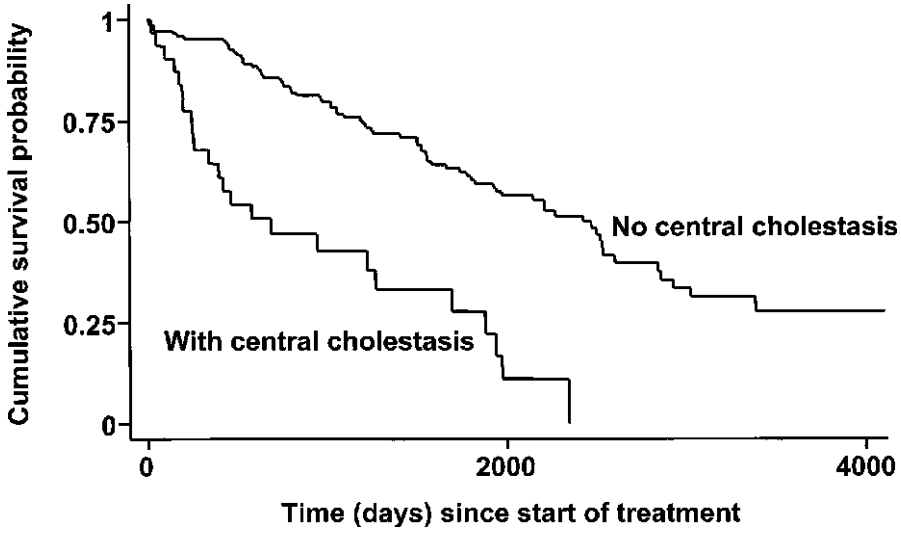

Fig. 26.3 Kaplan–Meier estimates of the survivor function, *S(t)*, for primary biliary cirrhosis patients with and without and central cholestasis at baseline.

again. With two groups followed until everyone has died, both survival curves will also finish at 0; yet one group may have survived on average much longer than the other.

We solve the problem of allowing for differences in survival time by comparing the **hazards** in the two groups over the duration of follow-up. As noted at the beginning of this chapter, in survival analysis we avoid the assumption that the hazards of the event of interest are constant over the study period. Instead, we assume that the *ratio* of the hazards in the two groups remains constant over time, even if the underlying hazards change. In other words, we assume that at all times t:

$$\frac{h_1(t)}{h_0(t)} = \text{constant}$$

where $h_1(t)$ is the hazard in the exposed group at time t and $h_0(t)$ is the hazard in the unexposed group at time t. This important assumption is known as the **proportional hazards** assumption.

Examining the proportional hazards assumption

We now see how this assumption may be examined graphically. It is difficult to estimate the hazard directly from data, since this would give a series of 'spikes' when an event occurs, interspersed with zeros when there is no disease event. Instead we use the **cumulative hazard function**, $H(t)$. This is the total hazard experienced up to time t, and is estimated by the sum of the risks at each time i at which an event occurs.

$$H(t) = \Sigma \frac{d_i}{n_i}, \text{ summed over all times up to and including } t$$

This is known as the **Nelson–Aalen estimate of the cumulative hazard function**. It follows from the definition of the cumulative hazard that the hazard function is the slope in a graph of cumulative hazard against time, so we can examine the way in which the hazard varies with time by examining how the slope of the cumulative hazard varies with time.

If the ratio of the hazards in the exposed and unexposed groups is constant over time, it follows that the ratio of the cumulative hazard functions must also equal this constant:

$$\frac{H_1(t)}{H_0(t)} = \frac{h_1(t)}{h_0(t)} = \text{constant}$$

And that, applying the rules of logarithms:

$$\log(H_1(t)) - \log(H_0(t)) = \log(\text{constant})$$

Therefore, if the proportional hazards assumption is correct then graphs of the log of the cumulative hazard function in the exposed and unexposed groups will be parallel.

Figure 26.4 shows the log cumulative hazard against time since start of treatment for primary biliary cirrhosis patients with and without central cholestasis at baseline. It suggests that there is no major violation of the proportional hazards assumption, since the lines appear to be reasonably parallel. In this example time has been plotted on a log scale to stretch out the early part of the time scale, compared to the later, because more events occur at the beginning of the study than near the end. Note, however, that this does not affect the relative positioning of the lines; they should be parallel whether time is plotted on a log scale or on the original scale.

It can be shown mathematically that that the cumulative hazard is related to the survival function by the following formulae:

$$H(t) = -\log(S(t)), \text{ or equivalently}$$
$$S(t) = e^{-H(t)}$$

Because of this, graphs of $\log(-\log(S(t)))$ are also used to examine the proportional hazards assumption.

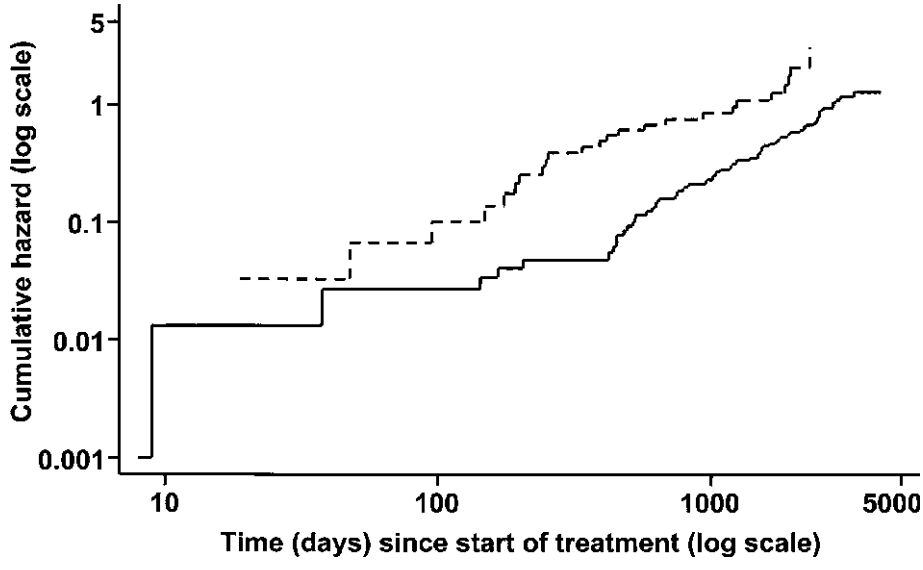

Fig. 26.4 Cumulative hazard (log scale) against time (log scale) for primary biliary cirrhosis patients with and without central cholestasis at baseline, in order to check the proportional hazards assumption.

Links between hazards, survival and risks when rates are constant

In Section 22.3 we described the relationship between risks and rates, and noted that when the event rate, λ, is constant over time then the proportion of the population event-free decreases exponentially over time. This proportion is exactly the same as the survivor function, $S(t)$. In the box below we extend the set of relationships to include the hazard, and cumulative hazard. Note that the hazard is constant over time, and that the cumulative hazard increases linearly over time. This is in contrast to the risk which does not increase at a steady pace; its rate of increase decreases with time.

When the event rate, λ, is constant over time:

$h(t) = \lambda$

$H(t) = \lambda t$

$S(t) = e^{-\lambda t}$

Risk up to time $t = 1 - e^{-\lambda t}$

Average survival time $= 1/\lambda$

26.5 COMPARISON OF HAZARDS USING MANTEL–COX METHODS: THE LOG RANK TEST

Mantel–Cox estimate of the hazard ratio

The Mantel–Cox method is a special application of the Mantel–Haenszel procedure, in which we construct a separate 2×2 table for each time at which an event occurs. It combines the contributions from each table, assuming that the hazard ratio is constant over the period of follow-up. We will use the same notation as that given in Table 18.3. Usually, there is only one event at a particular time, so in each table either d_{1i} is 0 and d_{0i} is 1 or vice-versa, but the procedure also works if there are ties (more than one event at a particular time). The **Mantel–Cox estimate of the hazard ratio** is given by:

$$\text{HR}_{\text{MC}} = Q/R, \text{ where}$$

$$Q = \Sigma \frac{d_{1i} \times h_{0i}}{n_i} \text{ and } R = \Sigma \frac{d_{0i} \times h_{1i}}{n_i}$$

Standard error and confidence interval of the Mantel–Cox HR

The standard error of log HR_{MC} is:

$$\text{s.e.}_{MC} = \sqrt{(V/(Q \times R))}, \text{ where}$$
$$V = \Sigma V_i = \Sigma \frac{d_i \times n_{0i} \times n_{1i}}{n_i^2}$$

V is the sum across the strata of the variances V_i for the number of exposed individuals experiencing the outcome event.

This may be used to derive a 95% confidence interval for HR_{MC} in the usual way:

$$95\%\ \text{CI} = HR_{MC}/EF \text{ to } HR_{MC} \times EF, \text{ where}$$
$$EF = \exp(1.96 \times \text{s.e.}_{MC})$$

Mantel–Cox χ^2 (or log rank) test

Finally, we test the null hypothesis that $HR_{MC} = 1$ by calculating the **Mantel–Cox χ^2 statistic**, which is based on comparisons in each stratum of the number of exposed individuals *observed* to have experienced the event (d_{1i}), with the *expected* number in this category (E_{1i}) if there were no difference in the hazards between exposed and unexposed. Note that χ^2_{MC} *has just 1 degree of freedom irrespective of how many events occur*.

$$\chi^2_{MC} = \frac{U^2}{V}; \text{ d.f.} = 1, \text{ where}$$
$$U = \Sigma(d_{1i} - E_{1i}), \text{ and } E_{1i} = \frac{d_i \times n_{1i}}{n_i}$$

This χ^2 test is also known as the **log rank test**; the rather obscure name comes from an alternative derivation of the test.

Example 26.3

In the trial of survival in primary biliary cirrhosis patients, there were 72 deaths among the 153 patients without central cholestasis at baseline, and 24 deaths among the 31 patients with central cholestasis at baseline. Table 26.3

Table 26.3 Calculations needed to derive the Mantel–Cox estimate of the hazard ratio and the corresponding (log rank) test statistic for survival in primary biliary cirrhosis patients, with and without central cholestasis at baseline.

Day, i	n_{0i}	d_{0i}	n_{1i}	d_{1i}	$Q_i = \frac{d_{1i} \times h_{0i}}{n_i}$	$R_i = \frac{d_{0i} \times h_{1i}}{n_i}$	$U_i = d_i - \frac{d_i \times n_{1i}}{n_i}$	$V_i = \frac{d_i \times n_{1i} \times n_{0i}}{n_i^2}$
9	152	2	31	0	0	0.3388	−0.3388	0.2814
19	150	0	31	1	0.8287	0	0.8287	0.1419
38	150	2	30	0	0	0.3333	−0.3333	0.2778
48	148	0	30	1	0.8315	0	0.8315	0.1401
96	148	0	29	1	0.8362	0	0.8362	0.1370
144	148	1	28	0	0	0.1591	−0.1591	0.1338
150	147	0	28	1	0.8400	0	0.8400	0.1344
167	147	1	27	0	0	0.1552	−0.1552	0.1311
177	145	0	27	1	0.8430	0	0.8430	0.1323
193	144	0	26	1	0.8471	0	0.8471	0.1296
201	144	0	25	1	0.8521	0	0.8521	0.1260
207	144	1	24	0	0	0.1429	−0.1429	0.1224
245	143	0	24	1	0.8563	0	0.8563	0.1231
251	143	0	23	1	0.8614	0	0.8614	0.1194
256	143	0	22	1	0.8667	0	0.8667	0.1156
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Totals					21.224	5.538	15.686	7.387

shows the calculations needed to derive the Mantel–Cox hazard ratio and associated log rank test statistic for the first 15 days on which one or more deaths occurred, together with the total values of U, V, Q and R for the whole dataset.

The estimated hazard ratio is $Q/R = 21.224/5.538 = 3.833$. The interpretation is that, on average, the hazard in patients with central cholestasis at baseline was 3.833 times the hazard in patients without central cholestasis.

The standard error of the log hazard ratio is

$$\sqrt{[V/(Q \times R)]} = \sqrt{[7.387/(21.224 \times 5.538)]} = 0.2507$$

The error factor is therefore $\exp(1.96 \times 0.2507) = 1.635$, so that the 95% CI for the hazard ratio is 2.345 to 6.264. The (log rank) χ^2 statistic is:

$$\chi^2_{MC} = \frac{15.686^2}{7.387} = 33.31, \; P < 0.001$$

There is thus strong evidence that the hazard rates, and hence survival rates, differed between the two groups.

These methods can also be extended to adjust for different compositions of the different groups, such as different sex ratios or different age distributions. For instance, we could stratify additionally on sex, and apply the method in the same way.

CHAPTER 27

Regression analysis of survival data

27.1 INTRODUCTION

We now describe **Cox regression**, also known as **proportional hazards regression**. This is the most commonly used approach to the regression analysis of survival data. It uses the same approach as the Mantel–Cox method described in Section 26.5:

- it assumes that the ratio of the hazards comparing different exposure groups remains constant over time. This is known as the **proportional hazards** assumption;
- it is based on considering the **risk sets** of subjects still being followed up at each time that an event occurred. At the time of each event, the values of the exposure variables for the subject who experienced the disease event are compared to the values of the exposure variables for all the other subjects still being followed and who did not experience the disease event.

After introducing Cox regression, we then consider:

- what to do when the proportional hazards assumption does not appear to hold;
- the way in which the choice of time axis influences the nature of the risk sets;
- the link between Cox and Poisson regression;
- the use of parametric survival models as an alternative approach.

General issues in regression modelling, including fitting linear effects and testing hypotheses, are discussed in more detail in Chapter 29.

27.2 COX REGRESSION

The mathematical form of the **Cox proportional hazards model** is:

$$\text{Log}(h(t)) = \log(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

where $h(t)$ is the hazard at time t, $h_0(t)$ is the **baseline hazard** (the hazard for an individual in whom all exposure variables = 0) at time t, and x_1 to x_p are the p exposure variables.

On the *ratio scale* the model is:

$$h(t) = h_0(t) \times \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)$$

When there is a single exposure variable (x_1) and just two exposure groups ($x_1 = 1$ for exposed individuals and 0 for unexposed individuals) the model is described by two equations, as shown in Table 27.1.

The **hazard ratio** comparing exposed with unexposed individuals at time t is therefore:

$$HR(t) = \frac{h_0(t)\exp(\beta_1)}{h_0(t)} = \exp(\beta_1)$$

The model thus assumes that the hazard ratio remains constant over time; it equals $\exp(\beta_1)$. It is this assumption that is highlighted in the name 'proportional hazards' regression. The regression coefficient β_1 is the estimated *log hazard ratio* comparing exposed with unexposed individuals.

Table 27.1 Equations defining the Cox regression model for the comparison of two exposure groups, at time *t*.

Exposure group	Log(Hazard at time *t*)	Hazard at time *t*
Exposed ($x_1 = 1$)	$\log(h_0(t)) + \beta_1$	$h_0(t) \times \exp(\beta_1)$
Unexposed ($x_1 = 0$)	$\log(h_0(t))$	$h_0(t)$

Example 27.1

Table 27.2 shows the output from a Cox regression analysis of the effect of central cholestatis at baseline (variable name *cencho0*) in primary biliary cirrhosis patients. There is clear evidence that this increased the hazard rate. The results are very similar to the Mantel–Cox estimate of the hazard ratio (3.833, 95% CI = 2.345 to 6.264), derived in Section 26.5. The square of the Wald z-test statistic is $5.387^2 = 29.02$, similar to but a little smaller than the log rank χ^2 statistic of 33.31, derived in Section 26.5. Three points should be noted:

1 Cox regression analysis is based on a **conditional likelihood** estimation procedure, in which the values of the exposure variables are compared between individuals *within* the risk sets of individuals being followed at each time at which

Table 27.2 Cox regression output for the model for the effect of central cholestasis at baseline in the study of survival in patients with primary biliary cirrhosis, introduced in Example 26.2.

	Hazard ratio	z	$P > \|z\|$	95% CI
cencho0	3.751	5.387	0.000	2.319 to 6.067

an event occurs. *The baseline hazard (which can vary over time) is therefore not estimated and is not displayed.*

2 As explained earlier, the model is based on the proportional hazards assumption. This assumption may be investigated graphically, as described in Section 26.4. Alternatively, statistical tests of the proportional hazards assumption are available, as discussed below.

3 As with all regression models, it is straightforward to estimate the effect of more than one exposure variable. As usual, we assume that the effects of different exposures combine in a *multiplicative* manner: this was explained in detail in Section 20.2, in the context of logistic regression. On the basis of this assumption, we may interpret the estimated effect of each exposure variable as the effect after controlling for the confounding effects of other exposure variables in the model. This assumption may be examined by fitting interaction terms (see Section 29.4).

27.3 NON-PROPORTIONAL HAZARDS

Non-proportional hazards correspond to an interaction between the exposure variable and time: in other words the exposure effect (hazard ratio) changes over time. In addition to the graphical examination of proportional hazards described in Section 26.4, many software packages provide statistical tests of the proportional hazards assumption. Three analysis options when evidence of non-proportional hazards is found are:

1 Extend the model to include an exposure-time interaction term. For example, for a single binary exposure variable, the model could assume:

$$\text{hazard ratio} = \exp(\beta_1 + \beta_2 t)$$

In theory, there is no reason that complex changes of the exposure hazard ratios over time should not be modelled. However, not all statistical software will allow this.

2 If the variable for which there is evidence of non-proportional hazards is a confounder, rather than the main exposure of interest, then the regression may be *stratified* according to the values of this confounding variable. This modifies the risk sets, so that they include only individuals with the same value of the confounding variable. The effect of the confounder is not estimated, but its effects are controlled for without assuming proportional hazards.

3 Split the follow-up time into different periods, as described in Section 24.6. It is then straightforward to fit models that allow the exposure effect to differ between time periods. Splitting follow-up time can also be used to derive tests of the proportional hazards assumption, by looking for interactions between exposure and time period (see Section 29.4 for a description of tests for interaction in regression models).

27.4 CHOICE OF TIME AXIS IN SURVIVAL ANALYSES

When following subjects after diagnosis or treatment of a disease, it may be reasonable to suppose that the major determinant of variation in the hazard will be the time since diagnosis or treatment. This was the assumption we made in the study of primary biliary cirrhosis, when we examined patients from the time they were treated. Our **risk sets** were constructed by considering all subjects who were at risk at the times after the start of treatment at which events occurred.

However, there are different options for the choice of time axis which may be more suitable in other situations. For example, consider the Caerphilly study of risk factors for cardiovascular disease, in which the dates of the first examinations took place between July 1979 and October 1983, and participants were aged between 43 and 61 when they were first examined. There are three possible choices for the time scale for construction of risk sets:

1 time since recruitment to the study;
2 time since birth (i.e. age);
3 year of the study (i.e. date).

Each of these choices will lead to different risk sets (sets of subjects at risk when an event occurred) at the times at which events occur. We illustrate the differences between these time scales using ten patients randomly chosen from the Caerphilly study. Their dates of birth, entry to, and exit from, the study, together with the corresponding ages and time in the study are shown in Table 27.3.

Table 27.3 Dates and ages of entry to, and exit from, the Caerphilly study for ten randomly selected subjects.

Subject number	Date of birth	Date of first examination	Date of exit	Age at entry	Age at exit	Years in study (T)	MI
151	20 Oct 1931	30 May 1980	18 Dec 1998	48.61	67.16	18.55	0
158	21 Mar 1933	2 Dec 1981	9 May 1984	48.70	51.13	2.43	1
658	12 Aug 1925	22 Oct 1981	18 Jul 1996	56.19	70.93	14.74	1
941	28 Oct 1933	29 May 1982	19 Dec 1998	48.58	65.14	16.56	0
1376	19 Sep 1935	21 Mar 1982	25 Nov 1998	46.50	63.18	16.68	0
1467	9 Jan 1930	6 Jul 1982	3 Aug 1993	52.49	63.56	11.08	0
1650	19 Nov 1927	24 Nov 1982	31 Dec 1998	55.01	71.12	16.10	0
1673	14 Feb 1926	3 Jul 1983	31 Dec 1998	57.38	72.88	15.50	0
1754	21 Jul 1921	1 Oct 1980	31 Dec 1998	59.20	77.45	18.25	0
1765	27 Mar 1924	30 Dec 1982	13 Dec 1998	58.76	74.71	15.95	0

The risk sets corresponding to the three different choices of time axis are illustrated in Figure 27.1. The horizontal lines represent the follow-up time for each subject. The follow-up line ends in a closed circle for subjects who experienced an MI (numbers 158 and 658). It ends in an open circle for subjects who were censored, either because they were lost to follow-up (subject 1467 on 3 August 1993), or because they were still healthy at the time of their end of study follow-up in November or December 1998 (the other seven subjects). Subjects whose follow-up is intersected by the dotted vertical lines, at the times of the MIs, are members of the risk set for that MI, i.e. those with whom the covariates of the patient who experienced the MI are compared.

1 *Risk sets* corresponding to *time from entry to the study*, Figure 27.1(a): at the time of the first MI all subjects were still being followed and are therefore in the risk set, while at the time of the second MI all subjects except 158 and 1467 are in the risk set.

   The majority of published applications of Cox regression use this choice, in which all subjects start at time 0. This is partly because Cox regression was originally developed for data on survival following a defined event, and also because until recently most computer programs for Cox regression insisted that all subjects enter at time 0. However, there is no reason why risk sets should not be constructed on the basis of delayed entry, and some statistical packages now allow flexible choices of time axis in Cox regression. In contrast, choices (2) and (3) both imply that subjects enter the study at different times, as well as having different periods of follow-up.

2 *Risk sets* corresponding to choosing *age as the time axis*, Figure 27.1(b): these consist of all subjects who were still being followed at a time when they were the same age as that of the subject who experienced the MI. Since subject 158 was relatively young when he experienced his MI, only three other subjects are members of this risk set. Similarly only four other subjects are members of the risk set for subject 658.

3 *Risk sets* corresponding to choosing *calendar time as the time axis*, Figure 27.1(c): in this example, because subjects were recruited over a relatively short period, the risk sets are the same as for (a), but in general this need not be the case.

Criteria for choice of time axis

In general, the best choice of time axis in survival analysis will be the scale over which we expect the hazard to change most dramatically. In studies of survival following diagnosis of a disease such as cancer, the best time axis is usually time since recruitment (start of study). Calendar time would be a sensible choice in studies of survival following an environmental disaster, such as the leak of poisonous fumes from a factory, which occurred at a particular time. In contrast, recruitment to the Caerphilly study did not depend on the participant experiencing

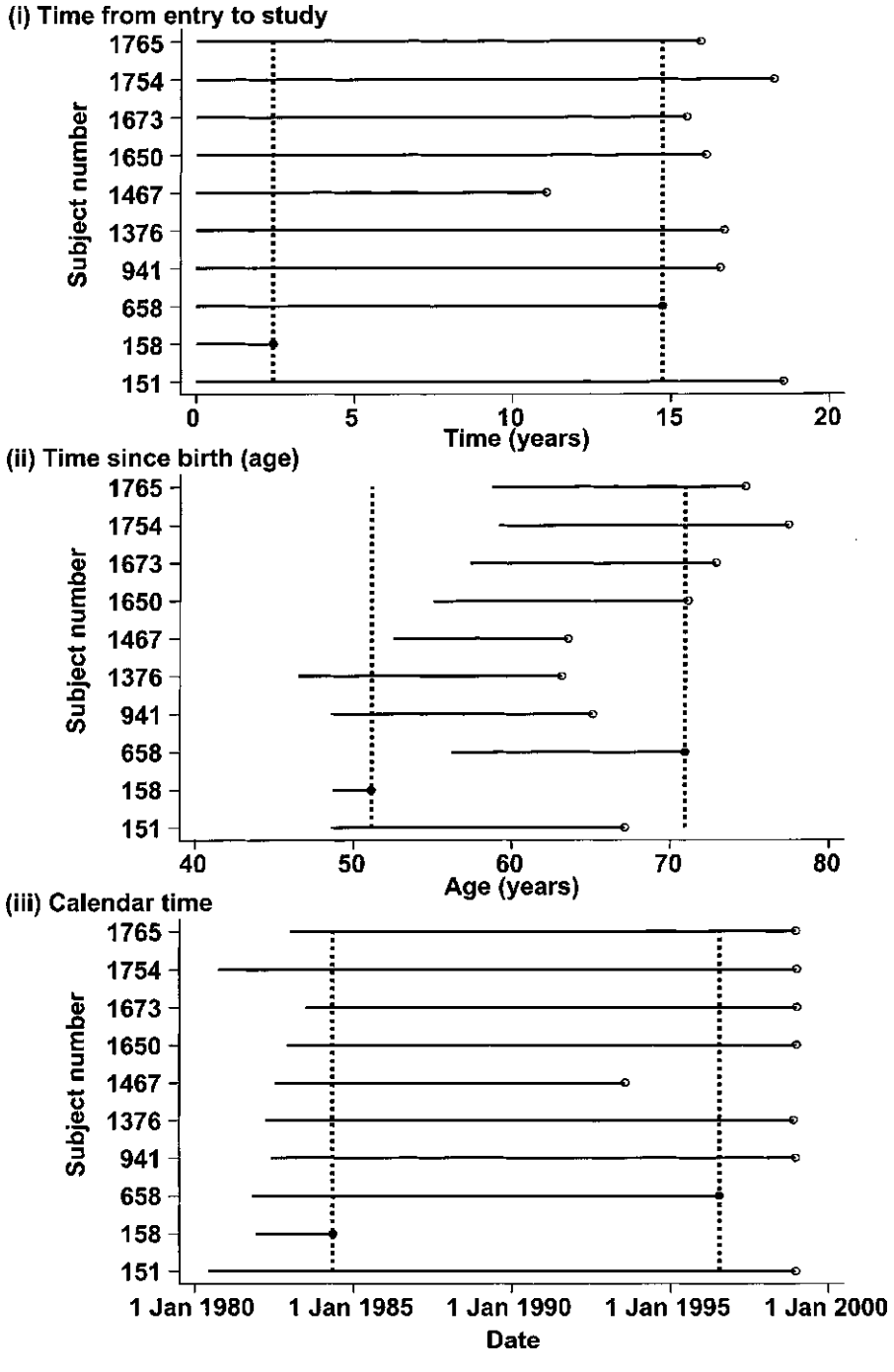

Fig. 27.1 Risk sets corresponding to three different choices of time axis, for ten patients randomly chosen from the Caerphilly study. The follow-up line ends in a closed circle for subjects who experienced an MI and an open circle for subjects who were censored. The dotted vertical lines show the risk sets at the time of each MI for the different choices of time axis.

a particular event: simply on the person living in Caerphilly and being in later middle age at the time the study was established. Therefore measuring time from recruitment to the study does not seem a sensible choice of time axis: in this case age is a better choice.

More than one time axis

Finally, we may wish to do a Cox regression that allows for the effect of more than one variable to change over time. There are two main reasons for doing this:

1. we may want to allow for changing rates of disease according to, say, age group, while keeping time since an event such as diagnosis of disease as the time axis used to define the risk sets;
2. we may want to allow for the effect of exposures which are measured more than once, and estimate the association of the most recent exposure measurement with rates of disease.

The procedure is the same in each case. We simply split the follow-up time for each subject into periods defined by (1) age group, or (2) the time between exposure measurements, in the same way as described at the end of Section 24.6. Providing that the software being used for Cox regression will allow for delayed entry, we then fit a standard Cox regression model, controlling for the effects of the time-varying exposures.

27.5 LINKS BETWEEN POISSON REGRESSION AND COX REGRESSION

We have described two different regression models for the analysis of longitudinal studies. In Poisson regression we assume that rates are constant within time periods, and estimate rate ratios comparing exposed with unexposed groups. In Cox regression we make no assumptions about how the hazard changes over time; instead we estimate hazard ratios comparing different exposure groups. This is done by constructing *risk sets*, which consist of all subjects being followed at the time at which each event occurs, and assuming that the hazard ratio is the same across risk sets.

At the end of Chapter 24 we saw that we may allow for variables which change over time in Poisson regression by splitting the follow-up time, for example into 5-year age groups, and estimating the rate ratio separately in each time period, compared to a baseline period. This is illustrated in Figure 27.2, using 5-year age groups, for the ten subjects from the Caerphilly study. We consider the total number of events, and total length of follow-up, in each age group. Now suppose that we make the age groups smaller (1-year, say). Only age groups in which an event occurs will contribute to the analysis, and the follow-up time within each of these groups will be approximately equal. As we make the time intervals progressively shorter, we will be left with the risk sets analysed in Cox regression.

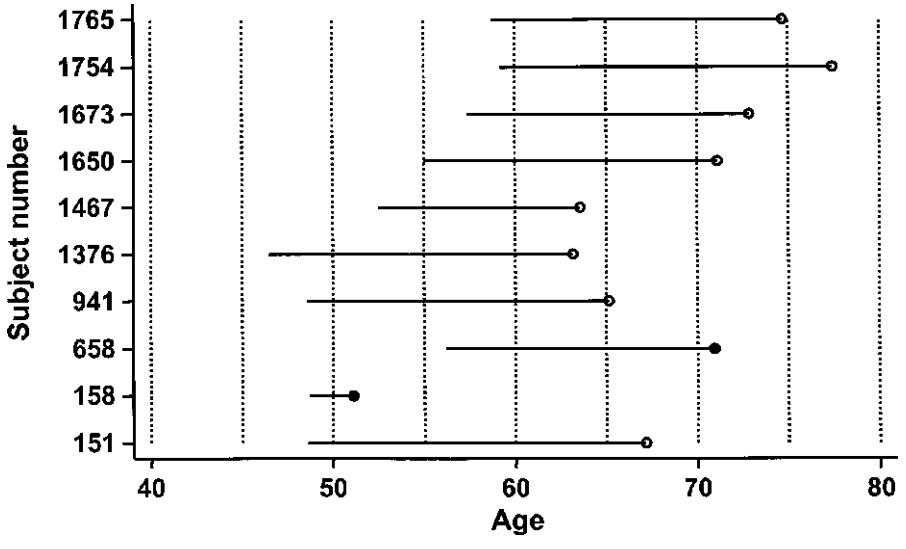

Fig. 27.2 Follow-up split into 5-year age groups, for ten subjects from the Caerphilly study.

27.6 PARAMETRIC SURVIVAL MODELS

Parametric survival models are an alternative regression approach to the analysis of survival data in which, instead of ignoring the hazard function, as in Cox proportional hazards models, we model the survivor function in the baseline group using one of a choice of mathematical functions. For example, we have already seen in Sections 22.3 and 26.4 that if the rate (hazard) is constant over time then the survivor function is exponential. This is exactly the assumption of Poisson regression, which means that it is therefore identical to a parametric survival model assuming an exponential survivor function. Other commonly used survivor function distributions are the Weibull, Gompertz, gamma, lognormal and log-logistic functions. **Weibull models** assume proportional hazards and usually give very similar estimated hazard ratios to those from Cox models. Because parametric survival models explicitly estimate the survivor function they may be of particular use when the aim of a study is to predict survival probabilities in different groups. For more details, see Cox and Oakes (1984) or Collett (2003).