Statistics: Analysis and Presentation of Safety Data

S. J. W. Evans

Introduction and background

All effective medicines have unwanted effects; it has been said that all medicines have two effects: the one you intend and one you don’t want. The consequence is that there must always be a continuing assessment of the balance of the risks and benefits of all medicines. Statistical methods and statistical thinking contribute at all stages in the process of drug development. In this chapter, the statistical issues relating to detecting adverse drug reactions (ADRs), both in clinical trials and in observational data (including spontaneous ADR reports), will be considered. These adverse effects may be clinical diagnoses or the signs and symptoms that potentially lead to a diagnosis. They may also be the results of laboratory tests or investigations, e.g. of X-rays or electrocardiograms.

Many medicines are licensed on the basis of the effect on a surrogate variable for efficacy, whereas adverse effects are usually not surrogates but are responses of immediate clinical relevance to a patient. For example, a drug may be licensed on the basis of its effect on blood pressure or cholesterol level, although these variables are not in themselves of direct clinical importance to the patient. The clinically important effects are those of, for example, stroke or heart attack. Adverse effects are usually noticed as being effects of clinical relevance, though occasionally some are identified because the variable is being measured for other purposes (e.g. a rise in blood pressure or increase in QT interval).

The title of the chapter refers to ‘safety data’, but safety is really an issue of absence of harm, and most data are collected on the occurrence of adverse effects. Some discussion of how ‘safety’ can be presented and discussed in summarizing the knowledge about a medicine is included. The main focus of this chapter will be on the analysis of randomized clinical trials, observational studies and spontaneous reports. There are statistical and public health issues involved in balancing risk and benefits, but these will not be covered here (see Chapter 8). There are also statistical issues arising from toxicology studies and from early phases in clinical drug development programme, which will not be dealt with specifically (see Chapter 3).
The statistical effort in relation to randomized trials has, over the years, been mainly devoted to assessment of efficacy. Medicines are licensed on the basis of their quality, safety and efficacy. Very early testing of a new drug will check its safety in animals and many drugs fail to progress to testing in humans. The early phases of drug testing will look for major safety problems in humans, but the main statistical effort in the late phases is placed on efficacy. It is clear that there is no point in introducing a new drug unless it has some efficacy in treating the disease or condition for which it is to be used. Major improvements have been seen in the preparation of protocols, and there are detailed guidelines that deal with the statistical approach to the analysis of efficacy. The details on safety are much less. At the design stage of a trial a key factor is to ensure that the statistical power will be sufficient to demonstrate clinically relevant efficacy if it exists. Such power calculations do not usually take safety considerations into account explicitly. A particular problem when concluding that efficacy exists when it does not (type 1 error) is multiplicity. Modern protocols pre-specify efficacy outcomes to be studied so that the possibility of testing many of them and choosing the most extreme is removed. With safety, which is the absence of ‘harm’, it is usually impossible to pre-specify a particular variable or outcome as being ‘of interest’ – there are many possible variables, so the potential for finding false positive effects (type 1 errors) is high. If the rate of false positives is minimized, then inevitably the rate of false negatives is increased. When concerns have been raised through previous studies, including animal toxicology and general knowledge of pharmacology, then pre-specified outcomes should be included in the protocol.

**Problems with clinical trials for detecting adverse reactions**

The first problem is that the sample size will often be too low to detect adverse reactions that are, although relatively rare, of considerable clinical relevance. The typical pre-licensing phase 3 trial used for assessment of efficacy of a new drug will have from about 50 to about 500 patients per treatment arm. Occasionally there will be both smaller and larger trials, and more than one trial will usually be needed to obtain a marketing authorization. This leads to problems with important but rare adverse events (AEs).

The situation is shown in diagrammatic form in Figure 6.1: the horizontal axis shows the rate of an AE in the control group (as a proportion, e.g. 0.001 is 1 in 1000, 0.1 per cent) and the vertical axis shows the relative risk of that AE in a treated group, compared with the control, that can be detected using different sample sizes. Three lines are shown using sample sizes per group of 50, 300 and 2000 with 80 per cent power. The smaller sample sizes are chosen because they are typical in many trials used to demonstrate efficacy. The largest sample size is that typical of very large trials used to demonstrate important effects on clinical outcomes (such as myocardial infarction or death) in cardiology. The uppermost line obviously relates to the smallest sample size – it is only very large relative risks that can be detected with small samples. Figure 6.1 shows that with a small trial of 50 per arm, unless the proportion of patients with the adverse effect in the control group is at least 0.2 (20 per cent), then it is very unlikely that the trial will have sufficient power to detect a relative risk of at least four. With 300 per arm then the background rate in the controls has to be at least 10 per cent to have a reasonable power to detect a doubling of the rate. In most practical situations where the AE is occurring at 5 per cent or less in the control group, then only relative risks of at least five will be detectable as statistically significant. In many
situations it is also clear that with important but rare ADRs the relative risks of 50 or 100 will be the only ones that can be detected in a single trial. In other words, only extreme levels of harm are detected in early trials, so that ‘safety’ is initially only the absence of extreme harm. Even with sample sizes of 2000 per group, detection of doubling of risks of 1 in 200 will not have good power.

This argument applies to a single adverse reaction. In practice, however, there is the potential for a large number of adverse reactions, so that the statistical analysis may need to take account of the problems caused by making multiple tests. The usual approach will result in adjustment to the significance test level so that the sample size necessary becomes very much greater. The use of adjustments will result in raising the rate of type 2 errors, i.e. real effects will tend to be missed. The need to combine data from several trials to achieve power is obvious and will be discussed later.

It is clear that, in typical phase 3 trials, the possibility of detecting even relatively frequent adverse reactions is small. It is important to use the most powerful statistical methods available to analyse information from such trials.

A further problem is that data on safety are reported erratically or unreported (Ioannidis and Contopoulos-Ioannidis, 1998). A more detailed investigation (Ioannidis and Lau, 2001) showed that the median amount of space on safety per article was less than one-third of a page and that less than half of reported trials gave specific reasons for the stopping of an individual’s treatment. The quality varied by medical subject area, with a tendency for reports on treatments for arthritis and HIV to give more careful attention to safety than cardiovascular trial reports. This variation may in part be due to the extent and the perceived importance of the drug’s expected toxicity.

**Figure 6.1** Relative risk detectable in studies of different size versus baseline rate of ADR
Issues of multiplicity

As has been noted above, there are difficulties when several response variables are analysed and tested for statistical significance. The same underlying problem occurs whether significance tests or confidence intervals are used. In the context of checking whether a new medicine has efficacy, it is most important to be conservative in the analysis so as to minimize chance findings of apparent efficacy when there is truly no benefit from a medicine. If several statistical tests are carried out on different response variables, then the probability that at least one of them becomes significant rises with the number of tests carried out. If, for example, 10 tests are carried out, each using a significance level of 0.05, then if the tests are independent of one another the overall probability that at least one of them is significant is

\[1 - (1 - 0.05)^{10} = 0.4\]

In order to preserve the probability of concluding that a difference has occurred when no difference truly exists (a type 1 error), then the significance level can be divided by 10 (the number of tests carried out), so that each test uses a significance level of 0.005. Now, the overall probability is \(1 - (1 - 0.005)^{10} = 0.049\) and the overall result still has a type 1 (false positive) error of approximately 0.05. This correction is called a Bonferroni correction. In practice, the various outcomes are not independent, so that the probability of finding at least one significant result is not as large as stated above. This also means that there is a tendency for a Bonferroni correction to overcorrect statistical tests; consequently, particularly when examining efficacy, other methods that are more powerful have been devised (Ludbrook, 1998).

The US Food and Drug Administration (FDA) have used rules (sometimes referred to as the Fairweather rules) that use 0.005 as the significance level for common tumours in toxicology testing in two-species–two-sex animal studies. They use 0.025 as the significance level for rare tumours, and note that the overall false positive rate is 10 per cent (Lin and Rahman, 1998). The consequence is that only extremely strong effects will be detected. This is part of the general problem of multiplicity that occurs in interpreting studies with many outcomes.

In the protocol, where efficacy is usually emphasized there is a pre-specified plan to study certain response variables. In studying new adverse reactions, no such pre-specified plan can usually be devised. The number of possible adverse effects is very large. There are also problems of how these effects are to be classified. In the medical dictionaries that are routinely used there are several levels of a hierarchy of terms, and at each level there are a number of terms used to group reactions. For example, in MedDRA there are 26 system organ classes, nearly 1700 high-level terms, and over 14,000 preferred terms (see Chapter 12). This means that any particular adverse reaction has problems of classification, and if even the coarsest classification level of system organ class is used then the potential for multiple significance tests is considerable. If AEs are classified at a high level, then grouping them may result in hiding a particular problem. There will always be a trade-off between grouping AEs in order to obtain sufficiently large numbers in any particular category and splitting them in order to find a specific problem. There is no general answer to this problem, since several classifications at different levels may need to be considered.
Analysis and presentation of data from trials

In a randomized trial comparing a new drug, in which new adverse reactions might be detected, there will be a comparator group that may be placebo or an active drug. In practice, during the development of a new drug, each AE of a serious nature will be examined individually to assess likely causality by staff from the company developing the drug. If necessary, new information will be released warning investigators of the possibility of this being an adverse reaction (see Chapter 11). The strength of the randomized trial is that causality can be inferred readily because randomization means that groups will, on average, be similar in all characteristics, both those observed and those unobserved. The statistical analysis can then be used to indicate that there is either a genuine effect associated with treatment or that an extremely unlikely chance effect has occurred. It is still possible that there are biases in execution of the trial, but these are less likely than in observational studies. When the control group is a placebo and evidence is convincing that an excess of the AE occurs in the new drug group, then this will be taken as strong evidence that it is causally related to the drug. The use of veiling (blinding), especially when it veils both patient and health professional who ascertains adverse effects, is very helpful in this context.

When the comparator group is an active drug, then the excess occurrence of an AE for the new drug will also be taken as evidence that this is an adverse reaction caused by the new drug. There are circumstances when the comparator drug is active and has a known adverse reaction, so occurrence of a similar rate of the AE will be taken as evidence that it is also an adverse reaction for this drug, though it might be the events are caused by neither drug.

For efficacy, as noted above, surrogate measures are often used; for safety data in respect of new ADRs, surrogate variables are rare. In many randomized trials there is monitoring of laboratory variables, such as liver function tests, that are used to check on safety (see Chapter 5). The analysis of these laboratory variables proceeds as for other continuous measurements used for efficacy. This should include analysis that allows for repeated measures over time. They do not themselves detect a new adverse reaction, but it is sensible to ensure that the most powerful statistical analysis is used so that if liver damage or renal damage starts to occur then this is detected at an early stage. The trend in population average values will often be a surrogate for infrequent larger changes that occur in individuals. It is sensible to look for trends in these measures rather than only to look for rates of occurrence of clinically relevant change in an individual, such as three times the upper limit of normal, or 10 times the upper limit of normal. The rates of clinically significant change are certainly important, but they may occur so rarely that statistical power is too low. Measures of efficacy are not taken as the rate of occurrence of a clinically significant change in an individual patient in blood pressure, but efficacy can be based on population changes. Statistical significance is too often used to substitute for clinical relevance in such situations. A similar argument should apply to liver function and other laboratory values measured for safety purposes. The rest of this chapter discusses binary data, i.e. whether an AE has occurred or not. A good description of issues relating to continuous data can be found in Chuang-Stein et al. (2001).

Statistical measures of the occurrence of adverse events

In the usual randomized trial where individuals are allocated to the new treatment or control in parallel groups, there are $n_t$ participants in the treatment group who are followed and
there are \( x_t \) participants who have a particular AE. The simplest measure of occurrence is the proportion of participants who have that event during follow-up, i.e. \( x_t/n_t \). Similarly, with the same notation, the proportion in the control group who have that event will be \( x_c/n_c \). There are standard statistical tests for the comparison of proportions, with the most obvious being a one degree of freedom chi-squared test or Fisher’s exact test. These test the null hypothesis that the proportion with that AE is equal in the two groups. The null hypothesis is that the difference in proportions is zero. The data for the chi-squared test are laid out in a two-by-two table in Table 6.1.

<table>
<thead>
<tr>
<th></th>
<th>With AE</th>
<th>Without AE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated</td>
<td>( x_t )</td>
<td>( n_t - x_t )</td>
<td>( n_t )</td>
</tr>
<tr>
<td>Control</td>
<td>( x_c )</td>
<td>( n_c - x_c )</td>
<td>( n_c )</td>
</tr>
</tbody>
</table>

Both the chi-squared and Fisher tests result in a statistical significance level (\( P \) value) that gives the probability of obtaining a difference in proportions as large as, or larger than the one observed when there is no true difference in the proportion. The measure of the magnitude of the difference between the treated and control groups may be given as a difference in proportions or as a relative measure. The difference in proportions is often expressed as a percentage, but proportion is the statistically preferred value.

The two obvious relative measures are the odds ratio (OR) and relative risk. The odds of the adverse event in the treated group are

\[
x_t: n_t - x_t
\]

and the odds in the control group are

\[
x_c: n_c - x_c
\]

Differences in odds do not have any obvious meaning, unlike their ratio, called the OR. This is

\[
(x_t: n_t - x_t) / (x_c: n_c - x_c)
\]

It is possible to take the ratio of the proportions rather than their difference, and this is then described as a risk ratio or relative risk (RR). For both these measures, the null value (no difference) is unity. The difference in proportions is sometimes described as the risk difference, or absolute difference in risk.

In public health terms it is always absolute differences that are important. A rate of an ADR of 1 in 1000 untreated patients compared with 2 in 1000 who are treated is a relative risk of 2. A rate of an ADR of 1 in 1 000 000 untreated patients compared with 2 in 1 000 000 is also a relative risk of 2, but the public health implications differ by a factor of 1000. Relative measures are more sensitive to causal effects; very high RRs (say >10) will usually be causal, but if the background rate is very low indeed then they may not result in major action being necessary. These considerations are very relevant to risk–benefit balance.
(see Chapter 8). The presentation of absolute effects of benefit have sometimes used what appears to be an absolute number, the NNT (‘number needed to treat’ in order than one treated person gets a benefit they would not otherwise have had). This is not an absolute number, it is the reciprocal of a difference in rates, so the time over which the outcome has been measured (often 1 year) needs to be given. ‘NNT’ is implicitly the NNT to obtain benefit. Some authors have used ‘NNH’ as the ‘number needed to harm’, but it should be ‘NNT/H’, i.e. the number needed to be treated so that one person gets a harm they would not otherwise have had.

To take a specific example from a large trial, the Women’s Health Initiative (WHI) study (Rossouw et al., 2002), the comparison between oestrogen + progestin (HRT for simplicity) with placebo for incidence of invasive breast cancer is shown in Table 6.2. Here, the proportions with breast cancer are \( \frac{166}{8506} = 0.0195 \) with HRT and \( \frac{124}{8102} = 0.0153 \) with placebo. The difference in proportions is 0.0042 (0.42 per cent). The odds of having breast cancer in the treated group were 0.0199 (166/8340) and 0.0155 (124/7978) in the placebo group. The odds ratio is 1.28, and the RR is 1.275. The simple chi-squared test here is 4.29 with a \( P \) value of 0.0384. It should be noted that it is always better to give exact \( P \) values rather than just, for example, \( <0.05 \). Fisher’s exact test has a \( P \) value of 0.044. Usually, the chi-squared test will have a smaller \( P \) value than Fisher’s exact test, particularly when there are small expected values for the chi-squared test, and the exact test is to be preferred.

| Table 6.2 Occurrence of invasive breast cancer by treatment group in the WHI study |
|-----------------------------------|----------------|----------|
|                                   | With breast cancer | Without breast cancer | Total |
| HRT                               | 166              | 8340     | 8506   |
| Placebo                           | 124              | 7978     | 8102   |

This illustrates that odds are always larger than proportions and that ORs are always further from the null value of unity than RRs. ORs have desirable mathematical properties and their use should be encouraged.

A confidence interval is a measure of the amount of statistical uncertainty around a value known as the point estimate. If a large number of 95 per cent confidence intervals (CIs) are constructed, then the true value of the parameter being estimated will be included within 95 per cent of such intervals. It is possible to construct CIs for summaries of data such as proportions, differences or relative risks. The CIs for the RR and OR are usually based on taking their logarithms and the CIs will be symmetric on a log scale. (The null value for log OR and log RR is zero.)

For the data from the WHI trial, the 95 per cent CI for the risk difference of 0.0042 is 0.00024 to 0.0082. This is an excess rate of 4 in 1000 patients who receive HRT and are followed up for an average of 5.2 years. This is an NNT/H of 250 for 5.2 years follow-up, i.e. an NNT/H of 1300 for one extra case of breast cancer per year over that first 5 years. The 95 per cent CI for the OR of 1.28 is 1.013 to 1.618. The 95 per cent CI for the RR of 1.275 is 1.012 to 1.606. None of these intervals (for the risk difference, OR or RR) contains the null value for the relevant summary (zero or one). This is consistent with the statistical
significance tests that are taken to be statistically significant; the \( P \) value is less than 0.05. Rejection of the null hypothesis \( H_0 \) at an \( x \) per cent level and the \( (100 - x) \) per cent CI excluding the value for the \( H_0 \) will usually be equivalent. Further details of how to calculate the statistical tests and confidence intervals are given in intermediate level textbooks in medical statistics (Altman, 1991) or in epidemiology (Rothman and Greenland, 1998).

Measures that take time into account

In the discussion above it has been assumed that each patient has been followed up until the end of the study, provided that the AE did not occur in that patient. In general this will not be true and the time that each patient is ‘at risk’ of having the AE will need to be taken into account. Even where there is a single dose, such as a vaccine, the follow-up period is still relevant. Immediate ADRs may not need to take time into account explicitly, but it is clear that the use of the word ‘immediate’ indicates that a time period over which ADRs are ascertained is relevant.

The usual summary used is to add up the total time at risk for each patient and sum that for the treatment and control groups separately. This then gives the total person time at risk for each group, usually expressed as say thousand person-years. Then, assuming that all the individuals did not have the AE at start of follow-up, the total number of those having the event during follow-up divided by the person-years gives the incidence rate. The ratio of incidence rates is referred to as a rate ratio.

In the WHI study there was a mean of 5.2 years of follow-up, so that in the HRT group there were 44 075 participant years (P-years) and in the placebo group there were over 41 289 P-years at risk: the rate per 10 000 person-years for breast cancer was 38 and 30 in the treated and placebo groups respectively. These are average rates over the follow-up period of the trial. The rate ratio is 1.27 (the same as the risk ratio to two significant figures). A risk has individuals as the denominators for the risks, whereas a rate has person-time as the denominator. Like the RR, the rate ratio is also described as a relative risk measure (Rothman and Greenland, 1998: 49).

It may be noted that the assumption made when using person-years as the denominator is that the risk of having an AE is constant at all times during the follow-up period. The risk per unit time is called the hazard rate, and using total person-years as the denominator assumes that this rate is constant over time. With some types of adverse reaction this assumption may be reasonable, but often this is not the case. For example, most hypersensitivity reactions are relatively rapid in onset and if they do not occur early in continuous treatment then their likelihood of occurring later is very much less. At the other extreme, any causal effect on cancer is likely to take at least a year, and more usually at least 3 years, before it could be detected. This is illustrated by the data shown in Table 6.3 taken from the WHI report (Rossouw et al., 2002). A different assumption is that the ratio of the hazard rates in two groups is constant. This may be more realistic, and analysis methods that utilize this assumption are given below.

It could be argued that the expected effect of HRT on breast cancer should only start to appear after 2 to 3 years, so using the total person time as the denominator is very misleading. In the summary of trials submitted for licensing of a new medicine it is often found, that the total person time in the treated group across all the trials, or even the total number treated, is the denominator used in giving the rate of occurrence of AEs. This is
rarely the best way of presenting or summarizing the data, and it must be treated with great caution. The correct ways of dealing with this issue have been set out (O’Neill, 1987) but are often ignored.

It is usually very much better to present the cumulative hazard of the AEs, and good examples are seen in Figure 6.2, regarding data from the WHI study (Rossouw et al., 2002).

Figure 6.2 illustrates the cumulative risk (hazard) in each of the treatment groups for each of four classes of AE. At each time point when an AE occurs, the risk of occurrence is calculated based on the number of AEs occurring at that time divided by the number of participants at risk of having that event at that time. Those who have dropped out of the trial by that time point, for whatever reason, are not counted in the denominator. The method can also be used to examine benefits, and the original figure also showed benefit for two categories of clinical outcome: colo-rectal cancer and hip fracture.

The cumulative risk is obtained by calculating the probability of not having the event at that time point – this is sometimes referred to as the probability of ‘survival’. This may be applied to AEs, though its original use was in looking at death rates. This cumulative ‘survival’ probability is obtained by multiplying the cumulative survival up to the previous time point by the current survival probability. The cumulative hazard is \( 1/C_0 \) (cumulative survival). The process is started by assuming the survival probability at the start time is unity. The method is known as a Kaplan–Meier estimate of survival or cumulative hazard. Kaplan–Meier curves for AEs are best shown as cumulative hazard plots, as in Figure 6.2; this is a curve that goes upwards rather than the conventional survival curve, which goes downwards.

The calculation of the cumulative risk is simple and is given in most introductory medical statistics books, e.g. Altman (1991: 368).

The curves derived from the Kaplan–Meier method can themselves be misleading if too much attention is paid to the data at longer times. This is where the estimates are at their most uncertain, since the numbers ‘at risk’ may be rather small. Good practice truncates these curves so that data based on very few observations are not included. Figure 6.2 gives the numbers at risk (which is a good example to follow), but it can be seen that the numbers fall off after 4 years of follow-up, so that by year 6 of follow-up less than 25 per cent, and by year 7 less than 10 per cent of those randomized are at risk of having events.

<table>
<thead>
<tr>
<th>Year</th>
<th>HRT P-years</th>
<th>Placebo P-years</th>
<th>HRT BC</th>
<th>Placebo BC</th>
<th>HRT rate</th>
<th>Placebo rate</th>
<th>HRT/placebo rate ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8435</td>
<td>8050</td>
<td>11</td>
<td>17</td>
<td>13</td>
<td>21</td>
<td>0.62</td>
</tr>
<tr>
<td>2</td>
<td>8353</td>
<td>7980</td>
<td>26</td>
<td>30</td>
<td>31</td>
<td>38</td>
<td>0.82</td>
</tr>
<tr>
<td>3</td>
<td>8268</td>
<td>7888</td>
<td>28</td>
<td>23</td>
<td>34</td>
<td>29</td>
<td>1.17</td>
</tr>
<tr>
<td>4</td>
<td>7926</td>
<td>7562</td>
<td>40</td>
<td>22</td>
<td>50</td>
<td>29</td>
<td>1.72</td>
</tr>
<tr>
<td>5</td>
<td>5964</td>
<td>5566</td>
<td>34</td>
<td>12</td>
<td>57</td>
<td>22</td>
<td>2.59</td>
</tr>
<tr>
<td>6+</td>
<td>5129</td>
<td>4243</td>
<td>27</td>
<td>20</td>
<td>53</td>
<td>47</td>
<td>1.13</td>
</tr>
<tr>
<td>Total</td>
<td>44 075</td>
<td>41 289</td>
<td>166</td>
<td>124</td>
<td>38</td>
<td>30</td>
<td>1.27</td>
</tr>
</tbody>
</table>

\( ^a \) BC = number of cases of invasive breast cancer.

\( ^b \) Rate per 10 000 participant-years.
The Kaplan–Meier method does not directly provide significance tests or confidence intervals for comparisons between groups. It is possible to treat the data as comparisons of proportions as discussed above, but these do not take into account differences over time and do not fully utilize the data. The simplest method of comparing the curves is the log rank test (Peto et al., 1977). Although the result of this test can be expressed as a chi-squared statistic, it is often more informative to calculate the hazard ratio (HR) and its 95% confidence interval (95% CI) for each comparison. The HR represents the ratio of the hazard rates in one group compared to another, and the 95% CI gives an estimate of the uncertainty around this ratio.

**Figure 6.2** Examples of Kaplan–Meier estimates of cumulative risks (derived from Rossouw et al. (2002))

Statistical tests utilizing time since start of treatment

The Kaplan–Meier method does not directly provide significance tests or confidence intervals for comparisons between groups. It is possible to treat the data as comparisons of proportions as discussed above, but these do not take into account differences over time and do not fully utilize the data. The simplest method of comparing the curves is the log rank test (Peto et al., 1977). Although the result of this test can be expressed as a chi-squared statistic, it is often more informative to calculate the hazard ratio (HR) and its 95% confidence interval (95% CI) for each comparison. The HR represents the ratio of the hazard rates in one group compared to another, and the 95% CI gives an estimate of the uncertainty around this ratio.
value, it is not the same as the simple chi-squared test presented above. The log rank test treats the data in a similar way to calculating a Kaplan–Meier estimate. At each time point where an AE (failure) occurs, it is assumed that the rate should be the same in the treated as in the control group. An overall rate across both groups is calculated so that an expected number of failures is obtained for each group at that time point. The cumulative difference between the observed number of failures (O) and the expected number (E) for the whole time period under consideration is obtained and \((O - E)^2/E\) can be compared with a chi-squared distribution on one degree of freedom for testing the difference between the curves. This is a test of the null hypothesis that the two curves are identical. It does not assume anything about the hazard rate itself – it does not have to be constant, but it does assume that the ratio of the hazards is always constant and equal to one. There are various subtle modifications of the log rank test that apply different weights to the information at the beginning of follow-up compared with that at the end of follow-up. Further details on survival analysis can be found in Collett (1994).

A more complex method for comparing time to event data is that known as proportional hazards regression or Cox regression. This, like the log rank test, compares an entire survival curve without making assumptions about the form of the hazard rate at any particular time, but it does assume that the ratio of the hazard rates between two groups is proportional at all times. This method can be used to adjust for other prognostic factors, as well as for making a comparison between a treated and a control group. It may be used both for data from randomized trials and observational cohort studies. The result of the Cox model is a hazard ratio, which is analogous to a relative risk averaged over all the time points considered. It also allows for a confidence interval around the hazard ratio to be calculated.

In the WHI study described above, the estimated hazard ratio for invasive breast cancer was 1.26 with 95 per cent CI 1.00–1.59, derived from a Cox model analysis. This is similar to the point estimate of relative risk calculated above as 1.28 with CI 1.01–1.61. The Cox model took into account the clinical centre where patients were being treated, age, prior disease and their treatment group in a simultaneous low-fat diet trial. These adjustments are usually less relevant in a clinical trial than in an observational study.

The log rank test and the Cox model can be described as semi-parametric. This is because they do not assume a parametric form for the hazard rate over time, but they both effectively assume proportional hazard ratios. It is possible to have fully parametric models that assume a particular form for the hazard rate. For example, the exponential model assumes a constant hazard rate. It is possible to allow for hazard rates that increase or decrease or are even J-shaped. Some of these methods are described by Collett (1994). There are also methods available for checking the assumptions of survival analysis and these should be used when examining the difference between groups in rates of occurrence of adverse outcomes.

When comparing rates with the number of cases with events as the numerator and person-time as the denominator, the basic assumption is that the number of cases follows a Poisson distribution. Analysis of these rates uses Poisson regression; see Clayton and Hills (1993). The results of these analyses can be expressed as incidence rate ratios.

The results from a Cox model analysis are always presented as relative measures of the effect rather than as absolute measures. It is not possible to obtain either absolute measures of rates or relative risks at a specific time point directly from the analysis. With parametric methods it is possible to obtain absolute measures, and this approach may therefore be used more often in the future.
Combining data from several trials: meta-analysis

A major problem with clinical trials is that they tend to be too small to detect uncommon or rare ADRs. There are obvious benefits to be gained from putting all the available information together to increase statistical power. In principle, this is more important for analysis of ADRs than for analysis of efficacy. However, most of the problems with trials are not solved by combining data. Important problems that remain relate to the classification of ADRs and in ensuring that all the relevant data have been captured. If the trials have excluded those likely to be treated in clinical practice then meta-analysis might give a false sense of reassurance. A major problem with the standard systematic review (the process of defining the problem, searching for all data and setting them out) is that the data may be derived from published papers. These are prone to ‘publication bias’ (Egger et al., 2001). At the stage of applying for a marketing authorization for a new drug, both regulators and the company will have access to complete data on the drug prior to its being licensed. This means that there is no problem with ‘publication’ bias, since all the data are available (though unpublished at this stage).

The greatest strength of a meta-analysis of trials is that the results which are being combined are the within-study, between-treatment group differences. It means that the different studies themselves are not assumed to have similar results, but it is assumed that the between-treatment differences are relatively similar across studies. One of the consequences is that it is important that the scale on which the differences are measured shows consistency across studies. If the (absolute) baseline risk varies across studies, then it may be that the (absolute) risk difference differs markedly across studies, but the OR is consistent. Therefore, pooling the ORs across studies may be the best approach. Methods that assume the between-treatment differences are constant across trials are called fixed-effect models; where an allowance is made for some heterogeneity in the between-treatment differences, then these are called random-effects models. If the variation is very large, then even a random-effects model may not be sensible, and the very idea of combining disparate results should be questioned. The detailed statistical methods are beyond the scope of this chapter, but they may be found in Sutton et al. (2000) or Egger et al. (2001).

A frequently used but weaker, and in some instances flawed, approach to combining data is simply to add up the numbers across all trials of all the AE in the treated group divided by the number of patients randomized to treatment. The same is done for the control group and the overall rates compared between treated and control. In some instances this will give a similar result to a proper meta-analysis, but in most cases it will have less precision. This is particularly likely when there has been unequal randomization to treated and control in some of the trials, and such combination can be very misleading. Over- or under-estimation of between-treatment rates of events can occur. The method should not be used routinely.

A systematic review should be a routine part of the drug development process so that ADRs are able to be detected as far as possible (Koch et al., 1993; Lee and Lazarus, 1997).

Analysis and presentation of data from observational studies

All the statistical methods that are used in clinical trials may be used in observational studies, but the interpretation is much less easy. Randomized controlled trials (RCTs) generally (but not inevitably) result in the formation of similar groups, so that ‘like is compared with like’. In a particular trial there is no guarantee that the groups are similar, but the statistical
significance test used to compare the overall results gives the probability that an observed difference results from chance imbalance in both measured and unmeasured prognostic factors. This does not relate to any tests comparing the measured prognostic factors, and it should be emphasized that tests comparing values at baseline are not generally sensible (Altman, 1998). In observational studies, groups can differ in relation to many factors other than the treatment comparison of interest; these include patient characteristics, follow-up, ascertainment of treatment or of medical outcomes. Some of the factors may not have been measured. This means that the differences observed may be due to chance or many other factors that could be systematically different, so the interpretation of statistical significance tests for observational studies does not have the same interpretation as in an RCT. Patients change treatments and the problems of classifying exposure when patients switch categories are considerable. This can be problematic in deciding whether an ADR is caused by the treatment a patient is receiving when the diagnosis is made, or whether some previous treatment led to the patient having symptoms that caused a change in medication, with these symptoms being a precursor of the disease that is only diagnosed later. For example, bleeding may lead to a patient changing from one HRT to another, and subsequently endometrial cancer is found. Is the cancer caused by the first or the second HRT, or by neither?

Bias can occur in trials, but it is much more of a problem in observational comparisons. When trying to assess whether a new drug is associated with an adverse reaction in an observational study, there will be a comparison of the rate of occurrence of the adverse effect between those exposed to the drug and a control group. Confounding occurs when a third factor is associated both with the treatment and with the adverse effect. For example, age will affect the rate of occurrence of many AEs (e.g. myocardial infarction), and if the treated group is older than the control group then age becomes a confounding factor. The problem of confounding hardly exists in trials because with randomization there is no tendency for any ‘third factor’ to be associated with treatment. Confounding has a major impact in non-randomized studies.

A major problem in drug safety is called ‘confounding by indication’. This is a form of selection bias, also known as ‘channelling’. It occurs when those who receive one treatment are more severe in their condition or, for other reasons, are at higher risk for the outcome being reported or occurring than those in the comparison group. The change to a second form of HRT can be a form of this type of bias. A study examining the differences between patients with arthritis prescribed a Cox-2 inhibitor and those not prescribed one provides an empirical example of this bias (Wolfe et al., 2002).

Methods for dealing with confounding

Confounding may be addressed in the study design or in the analysis. The general principle is to remove one of the two associations that give rise to confounding. First, the association of confounding factor(s) with the outcome and secondly the association of treatment with the confounding factor(s). If either of these associations is no longer present in the data as analysed, then confounding does not occur. This applies to the two main types of observational study, the case-control and cohort studies. In cohort studies, exposed and unexposed groups are followed and the rate of occurrence of the outcome is compared between the groups. In case-control studies the logic is reversed, and the comparison groups are formed of those who have had the outcome event and a similar group who have not had the outcome event. Then the previous exposure is compared in the two groups.
Design: matching to reduce confounding

Matching is frequently used in case-control studies to attempt to remove the association between the disease outcome and possible third factors that could also be associated with treatment. It has been routine to match on age and sex in case-control studies; and when studies are done in a database of general practice medical records, then matching by general practice has also been routine. This is usually done on a case-by-case basis (individual matching) or it may be done for groups of subjects (frequency matching). A full discussion of matching is given by Rothman and Greenland (1998). Here, it is made clear that matching does not reduce confounding unless the matching factor is also taken into account in the analysis. Matching may improve efficiency (statistical power) if the matching factor is a true confounder, but it may harm efficiency if it is not.

Overmatching has various effects. If a matching variable is associated with exposure but not disease then this usually results in loss of statistical efficiency but does not introduce bias. If matching is on a variable affected by exposure or by disease (or even worse, both) then bias can be introduced and the results become unreliable. Matching on symptoms or signs of disease is not advisable, because they can be associated with both the disease and the exposure under study, and hence the associations within the strata formed by the matching will be biased (Rothman and Greenland, 1998, 157).

In cohort studies, two groups of patients are followed up after receipt of treatment or the opportunity to receive it. The exposed group is compared with an unexposed group, which may be no treatment or an alternative treatment. They are each followed to see if an adverse effect occurs, with the methods of analysis very similar to those used in randomized trials. In order to deal with confounding, the objective is to make a comparison of similar groups. Matching may also be used in cohort studies. The details of analysis are complex, and the gains from matching are not great in many circumstances (Greenland and Morganstern, 1990).

Design: restriction

A variable that is constant within a study cannot be a confounder. Restricting a study to, say, only males means that gender cannot cause confounding. This means that the number of available subjects is markedly reduced, and it also causes difficulty for generalization of the results. The main use of restriction may be as part of a programme of studies when demonstration of an effect in, say, a high-risk group provides a basis for a wider study. A wider study of similar size would have greater variation and could fail to demonstrate an adverse effect.

Analysis: stratification to reduce confounding

An alternative approach to matching in the design can be used at the analysis stage by dividing participants into strata. The strata are defined by different levels of the confounding variable. For example, if age is a confounder then the data are split into several age bands so that comparisons between exposed and unexposed people within strata are made between those of similar age. The results are then combined across strata to give a single result that can be described as ‘age-adjusted’. With this approach it is important to have sufficient strata so that within each stratum the rate of occurrence of the AE is consistent. Merely
dividing the data into two strata will rarely be sufficient to adjust for confounding. Natural
strata, such as hospital of treatment or general practice, are commonly used without there
necessarily being strong evidence that such factors are confounders. Stratification may be by
several different variables, with a consequence that, if many variables are used, the
individual strata will contain few observations. This means that analysis methods for sparse
data, such as Mantel–Haenszel, should be used. This treats the data within each stratum as a
two-by-two table, obtaining ORs from the table and calculating a weighted average OR
across all the strata, having weights that are proportional to the amount of information in
each stratum (Rothman and Greenland, 1998).

Stratification can apply both to case-control and cohort studies. It is useful to examine the
data with stratum-specific results, even if more complex analysis is used subsequently. Care
must be taken not to overinterpret findings in one stratum compared with another, as
variation is expected from one stratum to another just by chance.

For practical reasons, either relating to computing or to the cost of obtaining data,
appropriate analysis to address confounding in a large cohort may be difficult. One approach
used in these circumstances is to analyse a subset by forming a case-control study within a
cohort. This is known as a nested case-control study, and has been much used within large
databases such as the UK General Practice Research Database (Wood and Coulson, 2001).

Regression to reduce confounding

In most instances where there is a continuous confounding variable it is better to use a
regression method to adjust for the confounding. There are two main approaches. First, the
traditional method of adjusting for the effect of the confounding variable on the AE. This
will usually involve logistic regression in a case-control study. In a cohort study, where the
time taken for the AE to occur is analysed, then a survival analysis method such as
proportional hazards regression is used. In all forms of regression analysis, issues such as
linearity of effects and choice of which variables (called covariates) to include in the model
lead to difficulties. The pre-specification of the analysis is not always done, so there is
potential for finding the most ‘interesting’ results by doing very many analyses and reporting
only one. A second and newer, perhaps controversial, method is to calculate what is called a
propensity score.

Propensity scores

This method uses logistic regression to examine the factors that are associated with exposure
to treatment. The propensity is a score that measures the probability of individuals with
different characteristics using a drug, and is derived from the measured variables that are
assumed to be associated with treatment. It does not necessarily assume that these factors
are associated with the (adverse) effect being studied. Patients are then divided into strata
based on their propensity score, or the propensity score is used in a regression model to
analyse the occurrence of the AE. Adjustment using propensity score has some advantages,
particularly when dealing with relatively rare AEs in observational data (Braitman and
Rosenbaum, 2002). The precision of any regression equation is dependent on having a
sufficiently large number of occurrences of the response variable. The numbers who are
treated are usually very much larger than the number of occurrences of an adverse effect.
This means that the equation relating a propensity score to the probability of treatment is
able to be more precisely determined than an equation relating to the occurrence of the adverse effect (Wang and Donnan, 2001). This means that the method allows the association between treatment and a rare outcome to be modelled using many potential measured prognostic variables. It does not adjust for unmeasured variables that are not associated with the measured variables, but then no observational method is able to do that.

A recent example used propensity scores to allow for possible differences between a group treated with rifampicin and pyrazinamide compared with a control group receiving isoniazid (Jasmer et al., 2002). Treatment was given based on alternating weeks in treating latent tuberculosis infection. The outcome being studied was hepatotoxicity. There were 18 cases of grade 3 or 4 hepatotoxicity in 411 patients in the trial; 16 were in the combined treatment group and two in the control. The analysis stratified patients into five groups using the quintiles of the propensity score, to allow for possible differences between the groups. The crude OR was 8.5 (95 per cent CI 1.9–76.5), and adjusting for propensity the OR was 7.8 (1.7–71.3). In this study, little difference would be expected between the groups, but the method can be of considerable value. Further details may be found in Rosenbaum (2002).

Meta-analysis of observational studies

Systematic reviews with an accompanying meta-analysis are of greatest strength in the context of randomized trials; however, they can also have a place when examining observational studies. For example, meta-analysis has been used in the controversy over whether oral contraceptives containing desogestrel and gestodene have a higher rate of occurrence than those containing levonorgestrel (Henessy et al., 2001; Kemmeren et al., 2001). In these examples the cohort and case-control studies were combined to obtain an estimated overall effect.

The chapter in Egger et al. (2001) on observational studies is of particular relevance to issues of safety. Brewer and Colditz (1999) give a summary of the potential for the use of meta-analysis in assessing ADRs and post-marketing surveillance. In the same issue, Temple (1999) discusses some of the limitations. Meta-analysis is highly relevant in the regulatory setting, when all safety data from randomized trials and observational studies of a new drug are available. There are obvious limitations from using only published papers, both because of the general problem of publication bias (where positive results are more likely to be published) and because of the poor quality of data reporting in relation to possible adverse reactions (Ioannidis and Lau, 2001).

There are methods for combining data from observational studies and including randomized trial data (Sutton and Abrams, 2001). Unfortunately, observational studies may each be subject to the same type of bias, and the usual problem of low statistical power may not be the main issue. Allowing for non-sampling errors will generally be necessary.

Use of statistical methods for signal detection with spontaneous reports

A major method of detecting new ADRs has been the use of spontaneous reports (SRs) of suspected ADRs coming from health professionals (and in some countries from patients). The ‘Yellow Card’ system in the UK has been one of the best known methods (Davis and Raine, 2002). Inferring causality on an individual basis for each report is rarely possible. It
is sensible to use all the data coming from SRs, so that, even though the individual reports are not very reliable, taken as a whole they provide insight. The object in the first place is to detect a signal of a possible new ADR. Further detailed evaluation of the signal will then need to take place to test whether it is truly a real ADR, caused by the drug. The problem is that although there may be suspicion of an AE being caused by the drug, it is possible that the AE is merely part of the background – i.e. most diseases induced by drugs may also occur in the absence of any drug treatment. Statistical methods described below can help to sort the true ADRs from those that are just background.

The standard method for a long time was to take the rate of occurrence of the SRs and compare this with what might be expected as ‘background’. The numerator in the rate is the number of reports and the denominator is the number of prescriptions. This has many problems: the denominator is only a variable proxy for numbers of users. It is also well known that SRs will usually be an underestimate of the true number of ADRs and the extent of this under-reporting is not known. The figures that are often quoted, such as 10 per cent of ADRs are reported, are not strongly evidence-based. Under-reporting is very variable and is dependent on factors such as the seriousness of the reaction. At best, a sensitivity analysis can be done to examine a plausible range of rates of SRs, but this is not usually helpful. The data on prescriptions may also be delayed in time, so that immediate assessment of a possible new ADR is also delayed.

### Disproportionality methods

An alternative approach is to use the data on SRs without using prescription data. This is analogous to using a proportional mortality ratio, as is done in population epidemiology when accurate denominators are unavailable. The basic idea is to compare reports for a particular medical term, as the proportion of all reports, making the comparison between a particular drug and, for example, all other drugs in a database of SRs. This is illustrated in Table 6.4. The database may be the entire database for a country’s regulatory authority, the worldwide database held by the World Health Organization (WHO), or it can be the database of a particular company. Many of the reports will be background, so that those that are caused by the drug are expected to occur at a higher rate with this drug than in the rest of the database. Various disproportionality methods have been described, and their utility has been reviewed by van Puijenbroek et al. (2002). They are described below.

<table>
<thead>
<tr>
<th></th>
<th>Particular reaction</th>
<th>Other reactions</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug of interest</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>Other drugs</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
<td>b + d</td>
<td></td>
</tr>
</tbody>
</table>

The ratio of the proportions has been described as a proportional reporting ratio (PRR; Evans et al., 2001). The principle of the method was first suggested by Patwary (1969). It has been used by regulatory authorities and also by some companies. A closely related approach is to use the OR rather than the ratio of proportions. This has been referred to as a
reporting odds ratio (ROR) and used initially for consideration of particular ADRs for evaluation rather than simply detection of a signal (Stricker and Tijessen, 1992; Egberts et al., 1997; Moore et al., 1997).

\[
\text{PRR} = \frac{a/(a + b)}{c/(c + d)} \quad \text{ROR} = \frac{ad}{bc}
\]

The PRR and ROR will be approximately the same if \( b \gg a \) and \( d \gg c \), as will usually be the case.

The expected value in cell \( a \), assuming that the proportion of reports for this reaction for this drug is the same as this reaction for all drugs, is obtained by

\[
E = \frac{(a + c)(a + b)}{a + b + c + d} \quad (6.1)
\]

The observed count \( O = a \). An alternative is to use all other drugs to determine the expected ratio, so

\[
E = \frac{c(a + b)}{c + d} \quad (6.2)
\]

The analysis of the table should strictly be done on independent observations, so that the rates should use ‘all other drugs’ as the category, rather than ‘all drugs’. If the size of the database is very large compared with the number of reports for the drug of interest, then including this drug’s reports in the total will make virtually no difference to the result.

An alternative way of looking at this approach is comparing the number of observations of a particular drug–reaction combination \( O \) (cell \( a \) in Table 6.4) with the number expected \( (E) \), based on the row and column marginal totals. This is the way that the expected counts are derived for a chi-squared test. The generality of this approach will be discussed in another context below. The PRR is then simply \( O/E \). It has been suggested that a cut-off of \( \text{PRR} = 2 \) or more, the number of reports for a particular drug–reaction combination \( >2 \) and a \( \chi^2 > 4 \) can be used to define a signal (Evans et al., 2001). A \( \chi^2 > 3.84 \) is the exact value for a significance test at 0.05. This is then evaluated in further detail to see if there is evidence to suggest that the ADR is indeed causal. This approach will detect new ADRs and allow a focus on important signals when a large number of reports are received. However, the cut-off of \( \text{PRR} > 2 \) results in a large number of potential signals to be evaluated and higher cut-offs may be used.

There are a number of issues that need to be addressed when using disproportionality methods. The analysis can be performed at any level of classification. The dictionaries used for these classifications may have a fully hierarchical system, e.g. MedDRA (Chapter 12). The term actually recorded by a health professional is at a very low level in the hierarchy of terms, such as ‘heart attack’. This will possibly have other synonyms, and the ‘preferred term’, at a higher level, might be ‘myocardial infarction’. A still higher level term could be ‘ischaemic heart disease’, and the highest might be ‘cardiovascular disease’. The problem is that the lower level of terms may not have many reports and there can be misclassification, so that a grouping at a higher level may be preferable in this context. Most usage of PRR methods has been at the ‘preferred term’ level, but this has the consequence that a large
number of reports will be for a single reaction (at PT level)—drug combination. In a UK database of over 400,000 SRs, there are about 110,000 different drug—reaction combinations. About 70,000 of these only occur once in the database. There has been little work to show what is the most sensitive and specific level at which to detect new ADRs. It seems likely that the best strategy is to choose a level for maximum sensitivity, which will require a fairly high level, then to investigate in more detail those drug—reaction combinations that are suggestive of a signal. Any automated system should allow for searching for new ADRs at more than one level in the hierarchy. In a system for automatic screening, the only feasible drug comparator is all drugs; but, for a detailed investigation, other groupings, such as same class of drug or same indication, may provide insight.

The next problem is whether to adjust for other factors such as age and sex. There is no doubt that background events will be age and sex dependent, so that it is sensible to stratify the data by age and sex provided that age and sex are known for most of the reports. This approach has recently been used by the FDA (see DuMouchel method below). It is likely that, in some circumstances, this will increase sensitivity and specificity, but it is important that the method used works well with sparse data, since stratifying will mean that the numbers in each 2×2 table will tend to be small. It is also possible to adjust for other factors, such as calendar time. This latter factor is more relevant when scanning an entire database, but it is less sensible when using the method for monitoring new reports as they arrive.

A further issue is whether to use all reports or only those of a serious nature. The term ‘serious’ has been used to mean those events that result in hospitalization (or its prolongation), death, and congenital malformation. At the UK MCA the word ‘serious’ is used in the ADROIT database to refer to a set of terms that are regarded as being serious whether they have one of those outcomes or not. Other dictionaries have different thresholds for terms of a serious nature, such as ‘critical terms’ in the WHO dictionary. The MCA has used PRR methods for examining serious events as part of a prioritization approach. The most important ADRs from a public health perspective are the serious ones, but this does not mean that lesser terms can be ignored.

When examining a class of drugs, it is reasonable to look at the PRR for each drug separately, but great care must be taken in making between-drug comparisons. There can be biased reporting for a particular drug—reaction combination, but much biased reporting applies to all suspected reactions for a specific drug. This implies that using prescriptions as a denominator will be more biased than the proportional methods. It is possible to use a statistical test for differential effects between drugs of the same class, but the danger is that these will lead to ‘over-interpretation’, since the underlying data are not of a high quality. The coding of drugs according to the ATC classification allows for sensible objective grouping of drugs for signal detection purposes. It must be emphasized that this process is one of scanning for signals rather than demonstrating causality.

A more important problem occurs when there are a very large number of reports for a particular drug—reaction combination. This is likely to be a true ADR, but it can result in reducing the PRRs for other reactions for that drug, and reducing the PRRs for that reaction in other drugs. This is because the numbers in cell \(a\) become large for this drug—reaction combination; when another drug is examined for this reaction then the total in the database is inflated and no longer provides what may be regarded as ‘background’. Similarly, the total of reports for this drug is then higher and the proportion of reports for other reactions is then less. This is illustrated by Figure 6.3 which shows the spectrum of the proportion of reports
for aspirin, derived from the MCA’s ‘Foreign’ database (i.e. reports sent by companies to the UK authority originating from outside the UK). A relatively smooth line is drawn giving the proportion of reports for the whole database as a reference. The horizontal axis has the type of ADR, and this is sorted in order of increasing frequency in the database. The numeric label is arbitrary. In Figure 6.3(a) it is shown on a linear scale and in Figure 6.3(b) it is shown on a log scale, so that the proportionate height of the point for aspirin and a particular reaction above the line for the database is a measure of the PRR. The very large number of reports of gastrointestinal bleeding reduces the PRR for other reactions. Incidentally, it is notable that there is a ‘trough’ for cardiovascular reactions, which might be expected in view of the known benefit of aspirin.
The remedy is to recalculate the PRRs, subtracting the data for the identified reaction from the entire database. This is not difficult for one particular instance, but it is difficult to do in a routine manner.

One statistical issue is whether the unit of reporting is taken as a report or as a reaction, since several reactions may appear on a single report. From a statistical point of view it is better to use a report as the unit, since there is possible dependence between reactions on the same report. However, precision is not the main requirement for a method of signal detection, and approximate answers will suffice. Apparent statistical precision is spurious, since sampling error is not the main problem; there are many biases in the data and their value is in signal detection, not in confirmation.

There is some potential for these methods to be used in assessing data from clinical trials. Control groups in trials can be a very useful source of the rate of occurrence of background AEs, and their potential has not yet been fully exploited.

A major problem with the PRR and similar methods (including ROR) occurs when the expected number is very small. This has two effects: first, the value of the PRR can be very high indeed with a small expected value; second, the PRR itself is not a reflection of how many reports have been received. Using a chi-squared test or a CI in combination with the PRR deals with this to some degree, but the chi-squared test (and CIs) will tend to be less than robust with very small expected values. A PRR of 200 is obtained with one report when 0.005 are expected and the same PRR occurs when 200 reports are observed with one expected. These have different implications for public health. The approaches discussed below seek to address this explicitly.

**WHO Bayesian confidence neural network**

This is a method based on the same principles described above but with more mathematical sophistication. It has been developed in Sweden and is in use by the WHO Centre for International Drug Monitoring in Uppsala, the Uppsala Monitoring Centre (UMC; Bate et al., 1998). It uses what is effectively log₂ of the PRR and calls this the **information component** (IC). It also calculates a CI around the IC so that a signal is detected if the lower limit of the 95 per cent CI (using a log scale) is above zero. The method has been validated by examining signals that would have been generated in the period 1993–2000 and testing whether they were in standard reference texts, Martindale and the Physician’s Desk Reference (Lindquist et al., 2000). The method showed a useful performance with 85 per cent negative predictive value and a 44 per cent positive predictive value. It was noted that there were 17 positive associations that could not be dismissed as false positive signals, even though they were missing from standard texts. The ‘gold standard’ is itself not truly a perfect standard and conclusions, therefore, must inevitably be limited.

A comparison between the Bayesian confidence neural network (BCNN) method and the PRR and ROR has been made by van Puijenbroek et al. (2002). This showed that each of the methods gives very similar answers when numbers of reports are greater than about three, but that, as expected, PRR and ROR are more unstable with small numbers of reports. They used the lower 95 per cent CI for the PRR, which is equivalent to using the PRR when chi-squared is statistically significant at the 5 per cent level, as discussed above.

The overall effect of the WHO method is similar to the PRR method used by the UK MCA, but based on a larger database derived from worldwide regulatory authorities. The
BCNN method has not generally been used with stratification of the data by age, sex and calendar period, though this would be possible.

**DuMouchel method**

The method that has been used in recent years by the US FDA is also a Bayesian method (DuMouchel, 1999). There is a public-domain version of the software that was used to carry out the analyses. It is similar to the PRR and other methods in utilizing the same type of $2 \times 2$ table, and emphasizes the $O/E$ approach. However, it has some advantage in that it is less vulnerable to small values of the expected count than the other methods, especially the PRR and ROR.

The method involves one pass through the entire database of SRs of (suspected) ADRs to estimate the parameters of the Bayesian prior distribution of the number of reports for all the drug–reaction combinations (hence, it can be seen as a form of empirical Bayes rather than based on subjective prior probabilities). The second pass then determines the departure from what might be expected given the number of reports for this reaction in total and the number of reports for this drug in total. This departure is approximately proportional to $O/E$ (the PRR) but shrinks the value towards unity, the null value, which makes a notable difference if the expected value is very small.

The assumption is that the prior distribution for the ratio of the observed counts in a particular drug–reaction combination to the expected counts is a mixture of two gamma distributions. Hence, the method is described as a ‘gamma Poisson shrinker’. The result presented is the geometric mean of the empirical Bayes estimate of the posterior distribution of the ‘true’ PRR. This is given the symbol EBGM, and the value of this is used to rank drug–reaction combinations. It is also possible to calculate a standard error for this estimate, and, as with the WHO method, the lower 95 per cent confidence limit is used for signal generation purposes.

Figure 6.4 shows a plot of the values derived from the DuMouchel method against the PRR for a selection of potential signals from the MCA’s ADROIT database. The symbol on the graph shows the number of reports for a particular drug–reaction combination. It shows that small numbers of observed reports can result in a high PRR (with a correspondingly very small expected number) but the EBGM is not as extreme. In most instances the number of reports is large enough not to make any notable difference, but with small expected numbers it is likely that the PRR will generate too many false positive signals. This is partly why the MCA have also used a cut-off requiring at least three reports for a drug–reaction combination.

Figure 6.4 shows that those drug–reaction combinations with small numbers of reports tend to move above a line of equality, and the EBGM shrinks them to much lower values.

The DuMouchel method is at its strongest in scanning an entire database to see if anything has been missed using the traditional case-by-case evaluation. It requires notable computer resources to carry out a regular monitoring of new reports, but this is not a major drawback since it is not difficult to run the whole process on a weekly or monthly basis.

There are new developments of DuMouchel’s method to allow for examination of drug interactions that have been applied within the FDA (Szarfman et al., 2002). This is described as a ‘multi-item gamma Poisson shrinker’ (MPGS). The software for this method is available commercially. This is very elegant and obtains signal scores for pairs, triplets and higher-order multiple numbers of drugs used by individual patients. The other methods using SR
databases have used each drug–reaction combination as a single entity rather than taking into account the fact that there may be more than one drug mentioned on a report. The potential for automated scanning of databases to look for drug interactions is beginning to be used in the FDA and in other systems. For a good review of the application of Bayesian methods see Gould (2003).

**Sequential probability ratio tests**

It is possible to compare two hypotheses based on the likelihood of observing the data given each of the hypotheses. The comparison of the likelihoods can be done in a sequential way with accumulating data, and are called sequential probability ratio tests (SPRTs). These tests discriminate between two hypotheses in the most powerful way and have been used in sequential medical trials (Armitage, 1958). They are used in industrial statistics for monitoring processes and were developed during World War II but not published until afterwards by Barnard (1946) and Wald (1947) independently. Much work has been done on this and similar methods in the context of monitoring clinical trials as data on efficacy accumulate (Whitehead, 1997). A new application has been in monitoring death rates in medical practice (Spiegelhalter et al., 2003). Here, it is suggested that these be applied to the problem of monitoring SRs to detect new ADRs.

The two hypotheses of interest are that the number of reports coming in for a particular drug and particular reaction is what is expected, i.e. the null, or alternatively that the number of reports is importantly divergent from the expected number. For example, the MCA has used a cut-off of two for PRRs, so a reasonable alternative hypothesis is to look for at least a doubling of the risk of a particular reaction.

The log-likelihood ratio (LLR) of the two hypotheses is calculated on the basis of...
accumulating data. This measures the relative likelihood of the two hypotheses given the data. It is for this process of dealing with accumulating data that this method is particularly suitable.

It is necessary to assume that SRs can be described using a Poisson model (which applies generally to counts of the number of events), and the alternative hypothesis is framed as the ratio of the rate for the observed to expected rate; in other words, effectively a relative risk (RR). The correct LLR to use, where $O$ is the observed number of SRs for a particular drug–reaction combination and $E$ is the number expected, is then

$$O \log(\text{RR}) - E(\text{RR} - 1)$$

If a doubling of risk is of interest, then $\text{RR} = 2$ and so the LLR becomes

$$O \log(2) - E$$

And so this suggests that for on-going monitoring

$$O \cdot 0.69 - E$$

is calculated as the LLR, which can be checked against a cut-off value.

The value of $E$ can be calculated from the existing SR database just as described for PRRs above, using Equations (6.1) or (6.2). In principle the value of $E$ could be obtained from other means, such as data on morbidity in a population similar to that treated by the drug, but the most straightforward way is to use the SR database. The use of this method, stratified by age and sex or other factors, has not yet been fully developed and tested in practice.

The threshold values allowing for continuous monitoring are given in Table 6.5. For risks of false positives and false negatives of 0.05, then, the boundary is 2.94.

<table>
<thead>
<tr>
<th>Type 1 error ($\alpha$)</th>
<th>Type 2 error ($\beta$)</th>
<th>LLR threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>2.94</td>
</tr>
<tr>
<td>0.05</td>
<td>0.1</td>
<td>2.89</td>
</tr>
<tr>
<td>0.05</td>
<td>0.2</td>
<td>2.77</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>2.20</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
<td>2.08</td>
</tr>
</tbody>
</table>

If a reaction has an extremely low background rate and occurs very rarely indeed, then $E$ will be close to zero. With zero expected events, then, the threshold is crossed when $O \cdot 0.69$ becomes $>2.94$, i.e. $O > 4.26$. This suggests that action should be taken when $O > 5$. This is a relatively conservative attitude towards false positive and false negative errors, and the most sensitive cut-off is when $O \cdot 0.69 > 2.08$, in other words approximately $O > 3$. This could be seen as a statistical justification of the old threshold of more than three spontaneous reports being regarded as a signal. The problem with that approach was that it assumed an extremely low background rate for that medical event.

The use of the SPRT in signal detection is still being developed, but it can be seen to place
more emphasis on the difference between the observed and expected than just on their ratio. The consequence is that, although the alternative hypothesis is effectively a PRR > 2, the magnitude of the number of reports is a major factor determining the generation of a signal. At the same time, it is clear that in public health terms there is a need for different thresholds for different possible reactions. Phocomelia or aplastic anaemia may need only a single report associating their occurrence with a particular drug to be sufficient to require immediate detailed investigation and follow up. There is further work to be done in setting standards for such monitoring.

Summary and conclusions

Trials

Standard methods of analysis may be used in analysing occurrence of adverse effects. In general, for ADRs (binary data) it is best to use survival analysis and to show plots of cumulative hazard over time in order to allow for dropout from trials and effects of medicines that are not necessarily constant over time. When an ADR is associated with continuous laboratory measurements, analysis should examine trends, just as is done for efficacy variables, in addition to converting them to binary variables using a cut-off, such as three-times the upper limit of normal (see Chapter 5).

Very large trials are needed to allow statistical methods to demonstrate the existence of adverse reactions, unless they are very common. The use of meta-analysis of all the available trials does help to increase power, but even then the power may remain limited. The consequence is that observational studies will continue to be necessary for detecting most unexpected ADRs.

Observational data

Again standard methods can be used. Bias is likely to be greater than in RCTs, but careful design and analysis can give reliable answers, although small effects must always be treated with caution. Propensity score methods have utility, especially in cohort studies when analysing rare outcomes, and many important ADRs are very rare.

The problem of studying multiple outcomes has not been solved. It is possible to use adjustments that reduce type 1 errors, but these automatically increase type 2 errors. It has been argued by some epidemiologists that adjustment is unnecessary, (Rothman, 1990). It seems sensible to treat new and unexpected findings that are inconsistent with previous knowledge or with other data from within a single study with caution, but there is an argument for approaching analyses of variables that are used for safety with a precautionary approach while being more stringent when multiple outcomes are examined for efficacy. The very use of the word ‘safety’ (as opposed to ‘harm’, which is what is actually studied) implies an asymmetry. Patients need reassurance over safety (Chapter 16), and there is a need to ensure that issues of safety are not dismissed too readily. Some studies are set up to test hypotheses in regard to safety that have been generated by individual case reports; when the response variable of interest is pre-specified, the problem of multiplicity is reduced.
Spontaneous reporting

The detection of new adverse reactions will depend on careful recording by health professionals, both of drugs used and medical events. Their suspicions, and in some cases those of patients, are vital to early recognition. Statistical methods can aid this process, provided they are used effectively at all stages in drug development from pre-licensing through to post-marketing surveillance. Disproportionality methods, such as the PRR, are effective tools for the analysis of data from spontaneous reports of suspected adverse reactions. Their mathematical sophistication must not be allowed to give undue weight to poor quality original data. These measures are to be used as signals of potential hazards and should not be misinterpreted as proof of causality.

Ongoing monitoring of new drugs may be helped by a statistical method like the SPRT, which takes the repeated examination of data into account. Statistical methods can be improved and introduced to support the never-ending vigilance required to extend knowledge about adverse reactions to drugs. Such knowledge should always be considered provisional and be open to re-evaluation in the light of further experience.

Acknowledgements

I am grateful to Professor Stuart Pocock, who made many valuable comments on this chapter. I wish to thank Dr Raine and the UK MCA for permission to use data derived from the ADROIT database. Dr David Spiegelhalter suggested the use of SPRTs for Poisson counts in monitoring death rates.

I am also grateful to Roussow et al. for taking the time to supply a modified version of their original figure (Figure 6.2).

Stata software version 7 (College Station, Texas, USA) was used for calculating power for Figure 6.1, and for processing data to create other figures.

References


