Statistical Methods in Medical Research

P. Armitage
MA, PhD
Emeritus Professor of Applied Statistics
University of Oxford

G. Berry
MA, PhD
Professor in Epidemiology and Biostatistics
University of Sydney

J.N.S. Matthews
MA, PhD
Professor of Medical Statistics
University of Newcastle upon Tyne

FOURTH EDITION

Blackwell
Science

assumptions that may not be easily testable. In particular, it is assumed that the regressions of response on compliance for the two groups have the same intercept at zero compliance, so that a non-compliant patient would have the same expected response whatever the assignment (as in the model of Example 18.1). Dunn (1999) discusses the effect of error in the compliance measurement, and concurs with a view expressed earlier by Pocock and Abdalla (1998) that complex models for compliance should not replace ITT analyses, but should be regarded as additional explanatory descriptions.

The journal issue edited by Goetghebeur and van Houwelingen (1998) contains several useful papers on this evolving body of methodology.

We note finally that non-compliance may affect any calculations of sample-size requirements made before the start of a trial. In an ITT analysis, the effect of non-compliance is likely to be to bias the difference in mean response between two treatments towards zero. This will be especially so in a trial to compare a new treatment against a placebo, if non-compliance results in patients assigned to the new treatment switching to the placebo. Suppose the expected difference is reduced by a fraction φ. Then the number of patients required to provide the intended power against this reduced treatment effect must be increased by a multiple $1/(1-\varphi)^2$. This can be taken into account in the planning of the trial if φ can be estimated reasonably accurately. Unfortunately, this will often not be possible. As we have seen, non-compliance can take many forms and have many consequences, the nature and extent of which will usually be unknown at the outset. It may be possible to estimate the proportion, θ, of non-compliant patients from previous experience of similar trials, but the effects of these protocol departures on the outcomes for the different groups may be less obvious. The best plan may be to make a simple, plausible assumption and err, if at all, in the direction of overstating the required trial size. One such assumption, in a two-treatment trial, might be that the outcome for a non-compliant patient is, on average, the same as that in the non-assigned group. The consequence of that assumption is that $\varphi = \theta$, and hence the intended trial size should be increased by a factor $1/(1-\theta)^2$ (Donner, 1984).

18.7 Data monitoring

In any large trial, the investigators should set up a system of *administrative monitoring*, to check that high standards are maintained in the conduct of the trial. Such a system will check whether the intended recruitment rate is being met, detect violations in entry criteria and monitor the accuracy of the information being recorded. Administrative monitoring may reveal unsatisfactory features of the protocol, leading to its revision. If the rate of recruitment is below expectation, the investigators may seek the cooperation of other medical centres or perhaps liberalize the entry criteria.

Administrative monitoring will normally make no use of the outcome data for patients in the trial. In contrast, *data monitoring* is concerned with the evidence emerging from the accumulating data on the safety and efficacy of the treatments under trial.

Safety will be an important issue in almost every trial. Most medical treatments and procedures produce minor side-effects, which will often have been anticipated from the results of earlier studies and may not cause serious concern. *Serious adverse events* (SAEs), especially when potentially life-threatening, must be carefully monitored (and perhaps reported to a central agency). A high incidence of unexpected SAEs, not clearly balanced by advantages in patient survival, may lead to early termination of the trial, or at least the modification or abandonment of the suspect treatment.

Differences in efficacy may arise during the course of the trial and give rise to ethical concerns. The investigators will probably have started the trial from a position of ethical equipoise, regarding all the rival treatments as potentially acceptable. If the emerging evidence suggests that one treatment is inferior to another, the investigators may feel impelled to stop the trial or at least drop the offending treatment.

The mechanisms for conducting data monitoring are discussed in the following subsections.

The Data Monitoring Committee

The responsibility for early termination or changes in protocol rests with the investigators (who, in a multicentre or other large trial, will normally form a *Steering Committee*). However, in a double-masked trial they will be unaware of the treatment assignments and unable to monitor the results directly. It is usual for the task to be delegated to an independent group, known as the *Data* [*and Safety*] *Monitoring Committee* (D[S]MC) or some similar title. The DMC will typically comprise one or more statisticians, some medical specialists in the areas under investigation and perhaps some lay members. It will not normally include investigators or commercial sponsors, although there is some variation of opinion on this point (Harrington *et al.*, 1994; Meinert, 1998).

The DMC should meet at approximately regular intervals, and receive unmasked data summaries presented by the trial statisticians. It will normally report to the Steering Committee, avoiding explicit descriptions of the data but presenting a firm recommendation for or against early termination or protocol modification.

In assessing the evidence, the DMC will need to bear in mind the difficulties arising from the repeated analysis of accumulating data. These are discussed in general terms in the next subsection, which is followed by a more explicit description of methods of analysis.

Sequential analysis

A *sequential* investigation is one in which observations are obtained serially and the conduct, design or decision on termination depend on the data so far observed. We are particularly concerned here with the possibility of early termination. *Sequential analysis* provides methods for analysing data in which the decision whether to terminate at any point depends on the data obtained. It was originally developed in pioneering work by A. Wald (1902–50).

Implicit in any sequential analysis is the concept of a *stopping rule*, defining the way in which the termination decision depends on the results obtained. A simple example arises in *sequential estimation*, where the purpose of an investigation might be to estimate a parameter to a specified level of precision. Suppose that, in a random sample of size n from a distribution with mean μ and variance σ^2, the estimated mean is $\bar{x}_n$ and the estimated standard deviation is s_n (the subscript indicating that these statistics will change randomly as n increases). The estimated standard error of $\bar{x}_n$ is $s_n/\sqrt{n}$ and, although s_n will fluctuate randomly, this standard error will tend to decrease as n increases.

A possible stopping rule for the sequential estimation of μ might therefore be to continue sampling until the standard error falls to some preassigned low value, and then to stop.

Standard methods of analysis, such as those described in the early chapters of this book, have assumed a fixed sample size, n. The question then arises whether these methods are valid for sequential studies in which n is not preassigned but depends on the accumulating data. The question can be answered at two different levels. From a frequentist point of view, the properties of a statistical procedure are affected by sequential sampling, in that the long-run properties have to be calculated for hypothetical repetitions of the data with the same sequential stopping rule rather than with the same sample size. However, for sequential estimation, as in the simple case described above, the effect is rather small. For instance, the probability that a confidence interval based on the estimated standard error covers the parameter value is not greatly affected, particularly in large samples. However, we shall see in the next subsection that other procedures, such as significance tests, may be more seriously affected.

From another standpoint we may wish to make inferences from the likelihood function, or, with a Bayesian interpretation, from the posterior distribution with some appropriate prior distribution. The stopping rule is now irrelevant, since likelihoods for different parameter values assume the same ratios whatever the stopping rule. This important result is called the *strong likelihood principle*.

In the data monitoring of a clinical trial, the case for early termination is likely to arise because there is strong evidence for an effect in favour of, or against, one treatment, and the standard way of examining such evidence in a

non-sequential experiment is by means of a significance test. Suppose that a significance test to compare the mean effects of two treatments is carried out repeatedly on the accumulating data, either at the occasional meetings of a DMC or more frequently by the trial statisticians. It is easy to see that the Type I error probability exceeds the nominal level of the significance test, because the investigator has a number of opportunities to find a 'significant' effect purely by chance, if the null hypothesis is true.

This effect is similar to that of multiple comparisons (§8.4), although the two situations are conceptually somewhat different. A hypothetical example will show that the effect of repeated significance tests is not negligible.

Suppose that, in a double-masked cross-over trial (§18.9), each patient receives two analgesic drugs, A and B, in adjacent weeks, in random order. At the end of the 2-week period each patient gives a preference for the drug received in the first week or that received in the second week, on the basis of alleviation of pain. These are then decoded to form a series of preferences for A or B.

It seems reasonable to test the cumulative results at any stage to see whether there is a significant preponderance of preferences in favour of A or B. The appropriate conventional test, at the nth stage, would be that based on the binomial distribution with sample size n and with $\pi = \frac{1}{2}$ (see §4.4). Suppose the tests are carried out at the two-sided 5% significance level. The investigator, proceeding sequentially, might be inclined to stop if at some stage this significance level were reached, and to publish the results claiming a significant difference at the 5% level. Indeed, this is a correct assessment of the evidence *at this particular stage*. The likelihood principle enunciated in the last subsection shows that the relative likelihoods of different parameter values (in this case different values of π, the probability of a preference for A) are unaffected by the stopping rule. However, some selection of evidence has taken place. The investigator had a large number of opportunities to stop at the 5% level. Even if the null hypothesis is true, there is a substantial probability that a 'significant' result will be found in due course, and this probability will clearly increase the longer the trial continues.

In this example, the discrete nature of the binomial distribution means that it is impossible to get a result significant at the two-sided 5% level until $n = 6$ preferences have been recorded, and then the Type I error probability is (again because of the discreteness) less than 5%, namely 0·031. With $n = 50$, the Type I error probability has risen to 0·171, and with $n = 100$ it is 0·227. In fact, it rises continually as n increases, eventually approaching 1.

For repeated t tests on a continuous response variable, assumed to be normally distributed with unknown variance, the effect is even more striking, since the error probabilities are not reduced by discreteness. For $n = 100$, the Type I error probability is 0·39 (McPherson, 1971).

To control the Type I error probability at a low value, such as 5%, a much more stringent significance level (that is, a *lower* probability) is required for assessing the results at any one stage. Suppose that the stopping rule is to stop the trial if the cumulative results at any stage show a significant difference at the nominal two-sided $2\alpha'$ level, or to stop after N stages if the trial has not stopped earlier. To achieve a Type I error probability, 2α, of 5%, what value should be chosen for $2\alpha'$, the significance level at any one stage? The answer clearly depends on N: the larger the value of N, the smaller $2\alpha'$ must be. Some results for binomial responses, and for normally distributed responses with known variance, are shown in Table 18.2.

The choice of N, the maximum sample size in a sequential test, will depend on much the same considerations as those outlined in §4.6. In particular, as in criterion **3** of that section, one may wish to select a sequential plan which not only controls the Type I error probability, but has a specified power, say, $1 - \beta$, of providing a significant result when a certain alternative to the null hypothesis is true. In the binomial test described earlier, a particular alternative hypothesis might specify that the probability of a preference for drug A, which we denote by π, is some value π_1 different from $\frac{1}{2}$. If the sequential plan is symmetrical, it will automatically provide the same power for $\pi = \pi_{-1}(= 1 - \pi_1)$ as for π_1.

Table 18.3 shows the maximum sample sizes, and the significance levels for individual tests, for binomial sequential plans with Type I error probability $2\alpha = 0{\cdot}05$ and a power $1 - \beta = 0{\cdot}95$ against various alternative values of π. These are examples of *repeated significance test* (RST) *plans*. More extensive tables are given in Armitage (1975, Tables 3.9–3.12).

Table 18.2 Repeated significance tests on cumulative binomial and normal observations; nominal significance level to be used for individual tests for Type I error probability $2\alpha = 0{\cdot}05$.

Number of stages, N	Nominal significance level (two-sided), $2\alpha'$, for individual tests	
	Binomial	Normal
1	—	0·050
5	—	0·016
10	0·031	0·010
15	0·023	0·008
20	0·022	0·007
50	0·013	0·005
100	0·008	0·004
150	0·007	0·003

Table 18.3 Maximum sample size, N, and nominal significance levels for individual tests, $2\alpha'$, for binomial RST plans with Type I error probability $2\alpha = 0{\cdot}05$ and power $1 - \beta = 0{\cdot}95$ against various values of π differing from the null value of 0·5.

π_1	$2\alpha'$	N
0·95	0·0313	10
0·90	0·0225	16
0·85	0·0193	25
0·80	0·0147	38
0·75	0·0118	61
0·70	0·0081	100

We have assumed so far that sequential monitoring of the data is carried out continuously, after every new observation. This is very unlikely to be the case in any large-scale trial, where data summaries are normally prepared at intervals, in preparation for meetings of the DMC. However, the effect of periodic monitoring may be much the same as for continuous monitoring, in that, when a periodic review suggests that a termination point may be encountered in the near future, a more intensive review is likely to be conducted for new data arriving in the immediate future. In that case, the trial is likely to be terminated at the same time as if continuous monitoring had been in place.

The book by Armitage (1975) presents plans for continuous monitoring, but is largely superseded by the more comprehensive book by Whitehead (1997). The latter book also concentrates mainly on continuous monitoring, but with a wide range of data types and alternative stopping rules.

The extensive use of DMCs, with interim data analyses at a relatively small number of times, has led to the widespread use of stopping rules based on *group sequential* schemes, whereby only a small number of repeated analyses are considered. We discuss these in the next subsection.

Group sequential schemes

A group sequential plan with specified Type I error probability could be obtained as a particular case of repeated significance tests with a small value of N, with the understanding that each of the N 'observations' is now a statistic derived from a group of individual observations. We need, though, to consider a more general framework. In the schemes for repeated significance tests illustrated in Tables 18.2 and 18.3, the proviso that the nominal significance level is the same for all the individual tests is unnecessarily restrictive. Some wider possibilities are illustrated in Example 18.2.

Example 18.2

In Example 4.14 (p. 140), we determined the sample size required to achieve specified power in a comparison of means of two groups of measurements from a lung-function test. Observations were assumed to be normally distributed with known variance, and, with a two-sided significance level of 0·05, a power of 0·8 was required for an alternative hypothesis that the standardized difference in means was 0·5. (The specified difference was $\delta_1 = 0{\cdot}251$, and the standard deviation was $\sigma = 0{\cdot}51$, but for present purposes it is the standardized difference, δ_1/σ, that matters.) The solution was that $n = 63$ individuals were needed in each group, a total of 126. We now consider various alternative group sequential plans that provide the same power.

Table 18.4 gives details of three group sequential plans, achieving the same power requirements. They all require five equally spaced inspections of the data. The table shows the standardized normal deviate (z value), calculated from a comparison of the means of the data so far, which would indicate termination of the trial at each stage. These are plotted in Fig. 18.1, and form sequential boundaries for the z value. Alternative methods of plotting boundaries use either (i) the bounds for the parameter estimate (in this case the difference in means); or (ii) those for a cumulative sum (in this case the difference between the two totals). These are easily obtained from the bounds for z, after n' observations in each group, by multiplying by $\sigma\sqrt{(2/n')}$ and $\sigma\sqrt{(2n')}$, respectively.

The characteristics of the three schemes illustrated in Table 18.4 and Fig. 18.1 are described below.

The *Pocock* boundaries (Pocock, 1977) are based on repeated significance tests at a fixed level, as described earlier. (The z value of 2·41 shown in Table 18.4 corresponds to the two-sided nominal significance level of 0·016 shown in Table 18.2.) The bounds for z determine the Type I error probability, while the power is controlled by the terminal sample size shown in the lower part of Table 18.4. Note that the terminal sample size, 155, exceeds the fixed sample size of 126, but

Table 18.4 Three group sequential schemes for the trial described in Example 18.2. Bounds for the standardized normal deviate (z value) at interim and final inspections. Entries in this table are derived from EaSt for Windows (1999) and Geller and Pocock (1987).

	Pocock	O'Brien–Fleming	Haybittle–Peto
Interim inspection			
1	2·41	4·56	3·29
2	2·41	3·22	3·29
3	2·41	2·63	3·29
4	2·41	2·28	3·29
5	2·41	2·04	1·97
Sample size			
Terminal	155	130	126
Mean on H_0	151	129	126
Mean on H_1	101	103	113

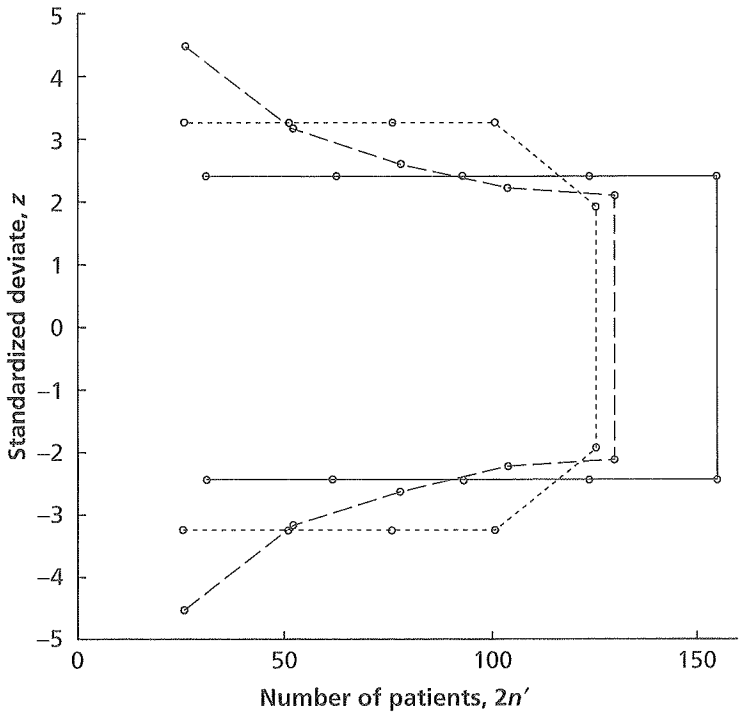

Fig. 18.1 Boundaries for three group sequential schemes with five interim inspections, as detailed in Table 18.4. (——) Pocock; (– – –) O'Brien–Fleming; (- - -) Haybittle–Peto.

if the alternative hypothesis is true (in this case a standardized difference of 0·5) the mean sample size is reduced substantially, to 101. Note also that the terminal bound for z substantially exceeds the fixed sample-size value of 1·96, so, if the trial continued to the fifth stage with a final value of z between 1·96 and 2·41, the interpretation would be difficult to explain. As Piantadosi (1997) remarks, 'this is an uncomfortable position for investigators'.

The other two schemes illustrated in Table 18.4 largely avoid this difficulty. The *O'Brien–Fleming* (OBF) scheme (O'Brien & Fleming, 1979) uses constant bounds for the cumulative sum that are very close to those appropriate for a fixed sample-size test at the final inspection. A consequence is that the bounds for z at the kth inspection ($k = 1, 2, \ldots, K$) are equal to those for the final inspection ($k = K$) multiplied by $\sqrt{(K/k)}$; $K = 5$ in this example. They are thus initially very wide, as shown in Table 18.4 and Fig. 18.1, and converge to final values close to those appropriate for a non-sequential test. When H_0 is true, trial results will usually lead to termination at the Kth inspection, and, as indicated by Table 18.4, the mean sample size is close to the maximum, lower for the OBF scheme than for the Pocock scheme. When H_1 is true, the two schemes have very similar properties. In some variants of OBF, the excessively wide bounds for $k = 1$ are pulled in to the standardized normal deviate corresponding to the 0·001 level, $z = 3{\cdot}29$.

Haybittle (1971) and Peto *et al.* (1976) suggested that a constant, but high, value of z should be used for all the interim analyses, which (as in OBF) permits a value close to the fixed sample-size value to be used at the final stage. For a Type I error probability of 0·05, different variants of the *Haybittle–Peto* scheme use 3·29 (as in Table 18.4) or 3·00 for the interim bounds.

In choosing between these three schemes, an informal Bayesian approach may be helpful. The Pocock scheme will lead to earlier termination than the others if the treatment effect is very large. If the initial view is that such large differences are plausible, it may be wise to adopt this scheme. Otherwise, the O'Brien–Fleming or Haybittle–Peto schemes will be preferable, avoiding, as they do, undue reliance on results from a small number of early observations, and removing the ambiguity about the bound for the final inspection. Further general comments about the use of group sequential schemes are contained in the final subsection.

The schemes described so far require the number of inspections to be decided in advance, and their timing to be at equally spaced intervals. These conditions are rarely achievable. Flexibility is provided by the *alpha-spending function* approach (Lan & DeMets, 1983; Kim & DeMets, 1987; DeMets & Lan, 1994), whereby the predetermined Type I error probability can be 'spent' in a flexible way, the schedule being decided for the convenience of the DMC, although independently of the trial results.

We have assumed so far that the observations are normally distributed with known variance. This is, of course, unlikely to be true, but, as in many of the methods described in this book, the normal distribution methodology often provides a useful approximation for a wide range of other situations. However, special methods have been developed for many other data types, including binary observations and survival times. A comprehensive account of group sequential methods is given by Jennison and Turnbull (2000), and other useful surveys are those by Kittleson and Emerson (1999) and Whitehead (1999). Geller and Pocock (1987) tabulate boundaries for various schemes. Useful software is provided by EaSt for Windows (1999) and the PEST system (MPS Research Unit, 2000).

Whitehead (1997) develops a very general system of sequential designs for continuous monitoring, implemented in PEST. For the standard normal model with known variance, they possess boundaries which are linear for the cumulative sum plot. Group sequential designs are handled by providing so-called *Christmas tree corrections* to the continuous boundaries.

Stochastic curtailment

The schemes described above permit early stopping when convincing evidence arises for a difference in efficacy between treatments. The main motivation there is the ethical need to avoid the continued use of an inferior treatment.

A somewhat different situation may arise if the interim results for, say, two treatments are very similar, and when it can be predicted that the final difference would almost certainly be non-significant. Methods for curtailing a trial under these circumstances have been proposed by many authors (Schneiderman & Armitage, 1962; Lan *et al.*, 1982 (using the term *stochastic curtailment*); Ware *et al.*, 1985 (using the term *futility*); Spiegelhalter & Freedman, 1988 (using Bayesian methods)). Boundaries permitting stochastic curtailment can be incorporated into schemes permitting early stopping for efficacy effects and are easily implemented with EaSt or PEST.

Although this approach may be useful in enabling research efforts to be switched to more promising directions, there is a danger in placing too much importance on the predicted results of a final significance test. Data showing non-significant treatment effects may nevertheless be valuable for estimation, especially in contributing to meta-analyses (see §18.10). It may be unwise to terminate such studies prematurely, particularly when there is no treatment difference to provide an ethical reason for stopping.

Other considerations

The methods described in this section have been developed mainly from a non-Bayesian point of view. As indicated earlier, in the Bayesian approach the stopping rule is irrelevant to the inferences to be made at any stage. A trial could reasonably be stopped whenever the posterior distribution suggested strong evidence of a clear advantage for one treatment. This approach to the design, analysis and monitoring of clinical trials has been strongly advocated, for instance, by Berry (1987) and Spiegelhalter *et al.* (1994). Grossman *et al.* (1994) have discussed the design of group sequential trials which preserve Type I error probabilities and yet involve boundaries determined by a Bayesian formulation, the prior distribution representing initial scepticism about the possible treatment effect.

We have assumed, in describing repeated significance tests, that the null hypothesis specifies a lack of difference in efficacy between treatments. It may be useful to base a stopping rule on tests of a specific non-zero difference (Meier, 1975; Freedman *et al.*, 1984; see also §4.6, p. 140, and the discussion of equivalence trials in §18.9). All the sequential methods outlined here can be adapted by basing the boundaries on tests of the required non-zero values.

The rather bewildering variety of methods available for data monitoring can perhaps be put into perspective by the widely held view that all such rules should be treated flexibly, as guidelines rather than rigid prescriptions. Many authors would argue that a DMC should define a stopping rule at the outset, even though its implementation is flexible. Others (Armitage, 1999) have favoured a more open approach, without a formal definition of the stopping rule, but with a

realization of the effect of repeated inspections of data on the Type I error. An intermediate attitude is perhaps to use the Haybittle–Peto approach, whereby differences of less than about three times their standard error are generally ignored during the interim analyses.

The reason for this sort of flexibility is that a decision to stop will usually depend on more than the analysis of a single response variable. There may be several primary endpoints, for both efficacy and safety, and, in a follow-up study, these may be measured at various times during follow-up. Effects seen at an early stage of follow-up might not persist over a longer period. Results from other relevant trials may suggest the need to terminate the current trial or amend the protocol. Changes in clinical practice or in evidence from other studies may change the views of the investigators about the importance or otherwise of significant but small effects. No single stopping rule would take account of all these features. Finally, it should be remembered that the decision whether or not to stop rests with the investigators: the DMC will make recommendations, but these need not necessarily be followed by the Steering Committee.

18.8 Interpretation of trial results

As noted in §18.3, the results of a clinical trial do not necessarily have an immediate impact on clinical practice. Other practitioners may have stronger prior convictions than the trial investigators, the refutation of which requires stronger evidence than that produced by the trial. Or there may be concern about long-term adverse effects or changes in efficacy. In this section we note a number of issues that affect the acceptability or interpretation of trial results.

Number needed to treat (NNT)

A new treatment may be more expensive or less acceptable to patients than a standard treatment. An important general question is whether the benefit apparently conferred by a new treatment justifies its use on the large number of patients falling into the relevant category. To some extent this should have been taken into account by the trial investigators, in designing the study, and discussed in the published report of the trial. In trials with a binary response variable, such as the incidence of stroke within a 3-year period, a useful index is available to indicate the balance between future usage and benefit.

Suppose that the probabilities of a specified adverse response in patients receiving a new or a control treatment are, respectively, π_T and π_C, estimated from the trial by p_T and p_C. Then, the number of patients needed to be treated to prevent one single adverse outcome (the NNT) is $1/(\pi_C - \pi_T)$, and this is estimated by $1/(p_C - p_T)$. Note that this is the reciprocal of the absolute risk reduction, and is not expressible purely in terms of the relative risk reduction.