



Using evidence to guide practice — Supplement

Common study designs¹⁻⁴

Cross-sectional study — a survey of the frequency of a disease or risk factor in a defined population at a given time. Can assess prevalence, may generate hypotheses about associations between risk factors and diseases, but cannot evaluate hypotheses, as does not take into account how the timing of exposure to a risk factor relates to the development of a disease.

Cohort study — an observational study of a group of subjects with a specific disease or characteristic who are followed up over a period of time to detect complications or new events. This group may be compared with a control group. If follow-up is over several years, this study can be prone to loss of subjects. Can show associations, but cannot establish causality.

Case-control study — a type of observational study in which characteristics of subjects with a disease are compared with a selected group of control subjects without the disease. The validity of this type of study depends on the appropriate selection of control subjects. Can show associations, but cannot establish causality.

Controlled trial — an experimental study in which an intervention is applied to one group of subjects and the outcome of interest is compared with that in an otherwise identical control group who received another intervention (such as another active treatment or placebo) at the same time. Ideally, subjects are assigned to treatment in a random way such that neither the subject nor the investigator knows which intervention the subject is receiving (this would be termed a **double-blind randomised controlled trial**). These studies often employ strict criteria for the inclusion and exclusion of subjects, which can lead to difficulties in extrapolating their results to wider populations seen in clinical practice.

Crossover study — a type of controlled trial in which each subject acts as his own control, by receiving intervention "A" for one period of time and intervention "B" for another. This design is only appropriate for evaluating palliative treatments of chronic stable conditions. The order in which interventions A and B are received and the time interval between receiving the interventions can bias the results. An adequate washout period between treatments and splitting the group into two so that one group receives intervention A first and the other group receives intervention B first can reduce these effects.

Systematic review — a review that systematically identifies, critically appraises and synthesises the results of a number of studies so that an overall estimate of the effectiveness of an intervention can be made, based upon all the available evidence. Publication bias, where selective publication of studies with positive results occurs, can be a problem.

Meta-analysis — a statistical technique used to combine the results from different studies identified, for example, by systematic review. The results of individual studies are weighted in the analysis to reflect the scientific rigour and uncertainties of the estimates of intervention effects. The validity of a meta-analysis depends on the quality of the systematic review on which it is based and the extent to which the characteristics of the individual studies differ (heterogeneity).

| Hierarchy of strength of evidence used in NICE technology appraisals (strongest to weakest) | |
|---|--|
| Ia | evidence from systematic reviews or meta-analysis of randomised controlled trials |
| Ib | evidence from at least one randomised controlled trial |
| IIa | evidence from at least one controlled study without randomisation |
| IIb | evidence from at least one other type of quasi-experimental study |
| III | evidence from non-experimental descriptive studies, such as case-control studies |
| IV | evidence from expert committee reports or opinions or clinical experience of respected authorities |



This supplement is only available online and supports MeReC Briefing Issue No. 30

This publication was correct at the time of preparation: September 2005

Understanding how results of clinical studies are expressed and calculated

Clinical studies are replete with statistics and other numerical expressions that can appear complicated and intimidating to those who are unfamiliar with how they are used and their meaning. However, it is not necessary to be a statistician to understand and interpret the more common methods of expressing study results. A brief overview of these methods is provided in *MeReC Briefing* Issue No. 30 and the ways of calculating these are shown below, using results from the CURE randomised controlled trial⁵ for illustration only.

The table below shows the incidence of events in the study. There were 6,259 subjects randomised to the group receiving clopidogrel plus aspirin (a + b) and there were 6,303 subjects randomised to the group receiving placebo plus aspirin (c + d). In the clopidogrel plus aspirin group 582 subjects experienced the primary endpoint event (a), and in the placebo plus aspirin group 719 subjects experienced the primary endpoint event (c) over a mean of nine months.

Absolute risk refers to the simple event rate in a group of people who receive an intervention.

Absolute risk of the primary endpoint event in the group receiving clopidogrel plus aspirin

$$= (a) / (a + b) = 582/6259 = 0.093 = 9.3\%$$

Absolute risk of the primary endpoint event in the group receiving placebo plus aspirin

$$= (c) / (c + d) = 719/6303 = 0.114 = 11.4\%$$

Relative risk (RR) estimates the magnitude of effect of an intervention of interest relative to the magnitude of effect of a comparator.

RR of the primary endpoint event in the group given clopidogrel plus aspirin compared to placebo plus aspirin

$$= \text{Absolute risk of the event with clopidogrel plus aspirin} / \text{Absolute risk of the event with placebo plus aspirin}$$

$$= [(a) / (a + b)] / [(c) / (c + d)] = 0.093/0.114 = 0.82$$

(NB. This is a crude RR. The CURE study correctly presents the relative risk over time as 0.80)

Absolute risk reduction (ARR) is the difference in event rates between two interventions. To be most useful, ARR must be set in the context of the underlying incidence of the event of interest. For example, without this information about the underlying incidence and risk of the event of interest, we would not know if an ARR of 1% represented a change in risk from 2% to 1% or from 21% to 20%.

ARR with clopidogrel plus aspirin compared with placebo plus aspirin

$$= \text{Absolute risk of event with placebo plus aspirin} - \text{Absolute risk of event in clopidogrel plus aspirin}$$

$$= [(c) / (c + d)] - [(a) / (a + b)] = 11.4\% - 9.3\% = 2.1\%$$

Relative risk reduction (RRR) is the reduction in risk of an event brought about by one intervention relative to the risk of the event in people using a comparator intervention. Without further information about the underlying incidence and risk of the event in the population, we would not know whether this relative reduction in risk represents a worthwhile

Worked example using crude incidence of efficacy and adverse events in the CURE study⁵ — for illustration of methods of calculating different expressions of study results only

| | Primary endpoint of cardiovascular death, non-fatal myocardial infarction or stroke over a mean of 9 months | | No. patients in each group |
|--------------------------|---|------|----------------------------|
| | Yes | No | |
| Clopidogrel plus aspirin | 582 | 5677 | 6259 (a+b) |
| Placebo plus aspirin | 719 | 5584 | 6303 (c+d) |
| | Major bleeding complications | | No. patients in each group |
| | Yes | No | |
| Clopidogrel plus aspirin | 231 | 6028 | 6259 |
| Placebo plus aspirin | 169 | 6134 | 6303 |

benefit. Results presented as RRR may appear more impressive than results presented as ARR. Therefore, RRR needs to be set in the context of the underlying incidence of the event to be meaningful.⁶

RRR with clopidogrel plus aspirin compared with placebo plus aspirin

= ARR with clopidogrel plus aspirin compared with placebo plus aspirin / Absolute risk of event with placebo plus aspirin

$$= \frac{[(c) / (c + d)] - [(a) / (a + b)]}{[(c) / (c + d)]} = 2.1\% / 11.4\% = 0.18 = 18\%$$

$$\text{or} = 1 - \text{RR} = 1 - 0.82 = 0.18 = 18\%$$

(NB. This is based on the crude RR. The CURE study correctly reports RRR as 20%)

Odds ratio (OR) expresses the odds of having an event compared with not having an event in two different groups. As long as the risk of the event of interest in both the intervention of interest group and the comparator group is low, OR and RR are approximately equal. This is the case in most RCTs so the use of one measure or the other is unlikely to have an important influence on treatment decisions. However, as the risk of the event of interest increases, estimates of RR and the OR diverge,⁷ so OR and RR should not be treated as the same in studies of subjects at high risk of events. For this reason, OR rather than RR should be used to express results in case-control studies.

OR for an event in the group given clopidogrel plus aspirin compared to placebo plus aspirin

= $\frac{\text{absolute risk of event compared with the absolute risk of no event in group given clopidogrel plus aspirin}}{\text{absolute risk of event compared with the absolute risk of no event in group given placebo plus aspirin}}$

$$= \frac{[(a) / (a + b)] / [(b) / (a + b)]}{[(c) / (c + d)] / [(d) / (c + d)]} = \frac{(a) / (b)}{(c) / (d)} = \frac{(a) \times (d)}{(c) \times (b)} = \frac{(582 \times 5584)}{(719 \times 5677)} = 0.80$$

(NB. The CURE study correctly reports RR)

Number needed to treat (NNT) expresses the number of people who would need to receive an intervention to prevent one event of interest. It is calculated by taking the reciprocal of ARR. The smaller the NNT, the greater the effectiveness of the intervention in the population studied. NNT is a useful, intuitive way of representing study results, but it must be remembered that, for making comparisons between different interventions by using their NNTs, like must be compared with like. An NNT for one intervention to prevent one type of event that is calculated on data collected over a certain number of months can only be compared directly with an NNT for another intervention if it relates to the same type of event calculated on data collected from a similar population over the same length of time.

NNT with clopidogrel plus aspirin instead of placebo plus aspirin to prevent one primary endpoint event over the study period

= 1 / ARR with clopidogrel plus aspirin compared with placebo plus aspirin

$$= 1 / [(c) / (c + d)] - [(a) / (a + b)] = 1 / 2.1\% = 1 / 0.021 = 48 \text{ over 9 months}$$

Number needed to harm (NNH) expresses the number of people who would need to receive an intervention for one person to suffer a harmful event of interest. As with NNTs, like must be compared with like when considering NNHs for different interventions. The larger the NNH, the less harmful the intervention with respect to the harmful event considered.

NNH with clopidogrel plus aspirin instead of placebo plus aspirin to cause one additional major bleed

$$= 1 / [(231/6259) - (169/6303)] = 100 \text{ over 9 months}$$

Understanding confidence intervals

The confidence interval (CI) around a result obtained from a study sample indicates the range of values within which there is a specific level of certainty (usually 95%) that the true population value for that result lies. Interpretation of the CI depends upon the kind of statistic the CI is created around.

CI around the difference in mean effects of two interventions

If a study finds that the CI around the difference in mean effects of two interventions contains the value zero, then we cannot rule out the possibility that there is no difference in effect between the interventions; the difference between the interventions would not be statistically significant. However, if the CI excludes the value zero, we can be reasonably (95%) certain that there is a difference between the interventions; the difference would be statistically significant.

Example of no statistically significant difference: In the MATCH (Management of ATherothrombosis with Clopidogrel in High-risk patients) trial⁸ that compared 18-month treatment with clopidogrel plus aspirin (both 75mg daily) with

clopidogrel 75mg daily alone in 7,599 patients who had suffered a TIA or ischaemic stroke within the previous three months, the primary endpoint was a composite of ischaemic stroke, MI, vascular death and rehospitalisation for an acute ischaemic event. This primary endpoint occurred in 15.7% of the patients receiving clopidogrel plus aspirin and 16.7% of patients receiving aspirin alone. The RRR was 6.4% and the 95% CI was -4.6% to 16.3% . Graphically this would look like **Figure A** below. As the 95% CI for the RRR crosses the value zero (which indicates that the possibility of no RRR cannot be ruled out), this RRR is not statistically significant.

Example of a statistically significant difference: In the CURE study in patients with acute coronary syndromes,⁵ the RRR in the composite primary endpoint with clopidogrel plus aspirin compared with aspirin alone **in the first month** was 22% (95% CI 8.6% to 33.4%). Graphically this would look like **Figure B** below. As the 95% CI for the RRR does not cross the value zero, we can be 95% certain that there is a RRR. Therefore, the RRR is statistically significant.

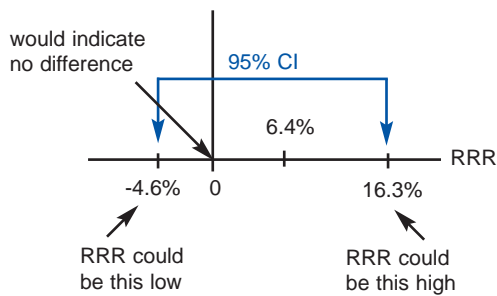


Figure A

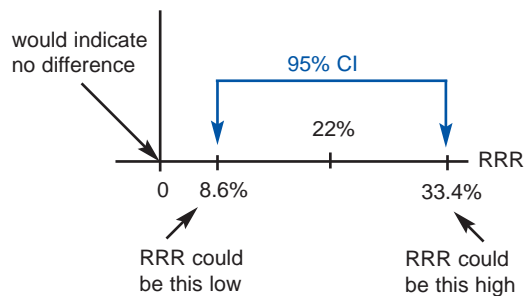


Figure B

CI around a ratio

When a CI is constructed around a statistic that is a ratio (e.g. RR, OR, hazard ratio — see MeReC Briefing Issue No. 30), if the CI does not contain the value 1.0, this would indicate that a statistically significant difference exists. However, if the CI contains the value 1.0, then this would indicate no statistically significant difference.

Example of a statistically significant difference: In the CURE study in patients with acute coronary syndromes,⁵ the incidence of major bleeding was 3.7% in patients taking clopidogrel plus aspirin and 2.7% in patients taking aspirin alone. The RR of a major bleed was 1.38 (95% CI 1.13 to 1.67). Graphically this would look like **Figure C** below. As the 95% CI for RR excludes the value 1.0 (which indicates that the possibility of no difference in major bleeding rates can be ruled out with 95% certainty) and its range is greater than 1.0, this indicates that the risk of a major bleed with clopidogrel plus aspirin is statistically significantly greater than the risk with aspirin alone.

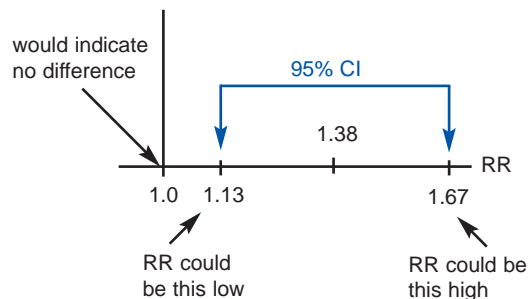


Figure C

CIs can also be constructed around the measure of effect of each separate intervention. If the CIs around each measure of effect do not overlap, this would indicate that the effects of each separate intervention are significantly different. However, if they do overlap it is less certain whether there is or is not a significant difference between the effects of each intervention.

The width of the CI (i.e. the range of values of the CI) can also provide useful information. A CI that is tight around the point estimate indicates that the study has sufficient power to be relatively precise. If the CI is very wide, this indicates that the study could be underpowered and the point estimate is imprecise.

Further sources of information on these concepts can be found in the Resources section.

Understanding diagnostic tests

It may be thought that diagnostic tests provide conclusive information about the presence or absence of a particular disease state (or condition). However, this is not so.⁹ Although test results may be reported simply as 'positive' or 'negative', the reliability of these results can be influenced by how common the disease state is in the population being tested and a number of properties of the tests used. It is therefore essential to understand how a particular test result predicts the risk of a disease in an individual in clinical practice.¹⁰

The reliability of a diagnostic test can be described by a number of measures⁹⁻¹² as outlined below and exemplified using a hypothetical test which produces the results in the table below. This demonstrates that out of 1000 people tested, 100 people truly have the disease (a + c) and 900 people are truly disease-free (b + d), for a disease prevalence of 10%.

Accuracy

Accuracy is a crude summary measure that describes the proportion of all test results that are correct. Presented alone, accuracy may appear impressive but when considering the other measures below it can be seen that accuracy alone is not necessarily a good indicator of a test's reliability.

Accuracy of the test below

$$= (a + d) / (a + b + c + d) = (95 + 855) / 1000 = 950 / 1000 = 0.95 = 95\%$$

Sensitivity and specificity

Sensitivity is a measure of how good the test is at picking up those people who have the disease being tested for. It is calculated as the proportion of people who actually have the disease that are identified as positive by the test. If a test has a high sensitivity, a negative result rules out the disease.^{9,11} This can be remembered using the mnemonic SnNout.¹³

Sensitivity of the test below

$$= a / (a + c) = 95 / 100 = 0.95 = 95\%$$

Specificity is a measure of how good the test is at identifying those people who do not have the disease. It is calculated as the proportion of people who are free of the disease that are identified as negative by the test.^{8,10} If a test has a high specificity, a positive test result rules in the disease. This can be remembered using the mnemonic SpPin.¹³

Specificity of the test below

$$= d / (b + d) = 855 / 900 = 0.95 = 95\%$$

Predictive values

Positive predictive value (PPV) is a measure of the probability that a person has the disease, given that their test result is positive. It is calculated as the proportion of people who test positive and actually have the disease.^{9,12} The closer the PPV is to 1.0 (100%), the greater the chance that a person who tests positive will actually have the disease.

PPV of the test below

$$= a / (a + b) = 95 / 140 = 0.68$$

(i.e. a 68% chance that a person who tests positive will actually have the disease, or around one in three results will be false-positive results).

Negative predictive value (NPV) is a measure of the probability that a person is free of the disease, given that their test result is negative. It is calculated as the proportion of people who test negative and are actually free of the disease.^{9,12}

Hypothetical example of results obtained with a diagnostic test for a disease with a prevalence of 10%

| Test result | Disease | | | Total tested |
|-------------|----------------|---|----------------|-------------------------|
| | Truly present | | Truly absent | |
| Positive | 95 a | b | 45 | 140 (a + b) |
| Negative | 5 c | d | 855 | 860 (c + d) |
| Totals | 100 (a + c) | | 900 (b + d) | 1000 (a + b + c + d) |

NPV of the above test

$$= d / (c + d) = 855 / 860 = 0.99$$

(i.e. a 99% chance that a person who tests negative will actually be disease-free).

PPV and NPV are highly dependent on the prevalence of the disease or condition in the patients being tested. This means that the PPV and NPV calculated in studies of a test may be very different from those that could be obtained in clinical practice if the prevalence of the disease or condition differs significantly. If the prevalence of the disease or condition is very low (as is often the case in the broad primary care population), the PPV will not be close to 1.0 (100%) even if both the sensitivity and specificity of the test are high.¹² Therefore, **it is not always appropriate to routinely use diagnostic tests on individuals at very low risk of the having the disease or condition being tested for.** This can lead to many false-positive results, which in turn can lead to unnecessary further investigations and anxiety for the individual.

Likelihood ratios

Likelihood ratios (LRs) summarise how many times more (or less) likely people with disease are to have a positive or negative result compared with people who are free of disease.¹⁰ At first sight they may seem complicated, but they provide more useful information than sensitivity and specificity and predictive values can. Unlike predictive values, LRs are not dependent on the prevalence of the disease or condition. However, they can be used to calculate the probability of having a disease or condition, based on the prevalence of the disease or condition.

The known or observed population prevalence of a disease or condition represents the crude probability of any individual person in that population having that disease or condition. In this example the prevalence of the disease is 10%, so the probability of an individual having the disease before a test is performed is 100 in 1000, or 0.10. This is called the 'pre-test probability' of having the disease or condition.

To use LRs, the pre-test probability of having the disease needs to be converted into the pre-test odds of having compared with not having the disease. In this example, the pre-test odds would be calculated as $0.10 / (1 - 0.10) = 0.11$.

The LR for a positive result (LR+) = sensitivity / (1 – specificity)

For the test above, $LR+ = 0.95 / (1 - 0.95) = 19$

This means that a person with the disease is 19 times more likely to have a positive test than a person without the disease. The pre-test odds multiplied by the LR+ provides the post-test odds of the individual having the disease ($0.11 \times 19 = 2.09$).

The post-test odds then needs converting to the post-test probability (i.e. the probability of having the disease according to our test). This is calculated as $2.09 / (1 + 2.09) = 0.676$, which is 67.6%. So in this example, in an individual who has tested positive for the disease, the probability of that person having the disease has increased from 10% to 67.6%. Note this still means there is approximately a one in three chance of a person with a positive result from this test actually not having the disease.

The LR for a negative result (LR–) = (1 – sensitivity) / specificity

For the test above, $LR- = (1 - 0.95) / 0.95 = 0.053$

The pre-test odds of having the disease was calculated as 0.11. The post-test odds of having the disease given a negative test result is calculated as $0.11 \times 0.053 = 0.0058$. From this, the post-test probability of having the disease given a negative test result is $0.0058 / (1+0.0058) = 0.0058$, which is 0.58%. So in this example, in an individual who has tested negative for the disease, the probability of having the disease has decreased from 10% to 0.58%.

A LR greater than 1.0 indicates that the test result is associated with the presence of disease and a LR less than 1.0 indicates that the test result is associated with the absence of disease. A LR equal to 1.0 is completely uninformative. LRs above 10 and below 0.1 are considered to provide strong evidence to rule in or rule out diagnoses in most circumstances.¹⁰

So what does all this mean?

The above test has an accuracy of 95%, a sensitivity of 95% and a specificity of 95%. For a population with a 10% prevalence of the disease being tested for, the probability of people being truly free of disease if they have a negative test result is over 99%. However, the probability of people actually having the disease if they have a positive test result is only around 68%. Therefore, around a third of all positive test results will actually be false positive results. As the prevalence of the disease changes in the population being tested, the number of false positive and false negative results will change. When used in populations with different disease prevalences this test generates the following:

Example of the effect of the prevalence of disease on the reliability of a diagnostic test

| Prevalence (pre-test probability of disease) | Probability of having the disease given a positive test result | Probability of having the disease given a negative test result |
|---|---|---|
| 1% | 16% (84% false positive results) | 0.053% (99% true negative result) |
| 10% | 68% (32% false positive results) | 0.58% (99% true negative result) |
| 25% | 86% (14% false positive result) | 1.74% (98% true negative result) |

In primary care, the probability of many people having a specific disease or condition is generally quite low. Using diagnostic tests on people at low risk of the disease or condition being tested for can lead to significant numbers of false positive results, which can lead to further unnecessary investigations to determine whether they truly have the disease or condition, and anxiety for the individual patient. Diagnostic tests are more usefully used in patients who are thought to be at high risk for the disease or condition being tested for. It is important to understand the limitations of diagnostic tests and appreciate what the measures of reliability actually mean.

Resources on evidence-based medicine and critical appraisal
Some trusted sources of evidence and information

- The National Institute for Health and Clinical Excellence — www.nice.org.uk
- The Department of Health — www.dh.gov.uk
- The Cochrane Library — www.nelh.nhs.uk/cochrane.asp
- Clinical Evidence— www.nelh.nhs.uk/clinical_evidence.asp
- InfoPOEMS — www.infopoems.com
- MeReC Publications — www.npc.co.uk/merec.htm
- Drug and Therapeutics Bulletin — www.nelh.nhs.uk/idth/default.asp
- NHS Centre for Reviews and Dissemination — www.york.ac.uk/inst/crd
- PRODIGY — www.prodigy.nhs.uk
- Bandolier — an evidence-based healthcare journal — www.jr2.ox.ac.uk/bandolier

Information on statistics and critical appraisal

- Using Evidence to Guide Practice. NPC Plus online training resource — www.npc.co.uk/npc_plus.htm
- Information Mastery materials — www.npc.co.uk/information_mastery.htm
- Greenhalgh T. How to read a paper. London: BMJ Publishing Group 1997
- Guyatt G, Rennie D, eds. Users' guides to the medical literature. Chicago: American Medical Association Press 2002
- Sackett DL, Richardson WS, Rosenberg W, et al. Evidence-based medicine: how to practice and teach EBM. Edinburgh: Churchill Livingstone 1997
- Public Health Resource Unit critical appraisal skills programme— www.phru.nhs.uk/casp
- British Medical Journal — topic collections on statistics and research methods — <http://bmj.bmjournals.com/collections/>

Communication tools

- Visual Rx — allows generation of visual tools for demonstrating NNTs — www.nntonline.net
- BestTreatments website — clinical evidence website for patients with decision support tools for explaining risk, etc. — www.besttreatments.co.uk

Supplement references

- 1 National Prescribing Centre. An introduction to assessing medical literature. MeReC Briefing 1995;9:1–8.
- 2 National Prescribing Centre. Sources of evaluated information on clinical effectiveness. MeReC Briefing 1997;11:1–8.
- 3 Davis HTO, Crombie IK. What is meta-analysis? Newmarket: Hayward Medical Communications, May 2003. Available from: www.evidence-based-medicine.co.uk. Accessed 22/09/05.
- 4 Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998;317:1185–90.
- 5 The clopidogrel in unstable angina to prevent recurrent events trial investigators. Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST-segment elevation. N Engl J Med 2001;345:494–502.
- 6 National Prescribing Centre. Evidence based medicine. MeReC Bulletin 1995;6:45–8.
- 7 Guyatt G, Rennie D, eds. Users' guides to the medical literature. Chicago: American Medical Association Press 2002.
- 8 Diener HC, Bogousslavsky J, Brass LM, et al. Aspirin and clopidogrel compared with clopidogrel alone after recent ischaemic stroke or transient ischaemic attack in high-risk patients (MATCH): randomised, double-blind, placebo-controlled trial. Lancet 2004;364:331–7.
- 9 Greenhalgh T. Papers that report diagnostic or screening tests. BMJ 1997;317:540–3.
- 10 Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ 2004;329:168–9.
- 11 Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. BMJ 1994;308:1552.
- 12 Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ 1994;309:102.
- 13 Sackett DL, Richardson WS, Rosenberg W, et al. Evidence-based medicine: how to practice and teach EBM. Edinburgh: Churchill Livingstone 1997.

The National Institute for Clinical Excellence (NICE) is associated with MeReC Publications published by the NPC through a funding contract. This arrangement provides NICE with the ability to secure value for money in the use of NHS funds invested in its work and enables it to influence topic selection, methodology and dissemination practice. NICE considers the work of this organisation to be of value to the NHS in England and Wales and recommends that it be used to inform decisions on service organisation and delivery. This publication represents the views of the authors and not necessarily those of the Institute.