VI

Testing your skill
Now it’s your turn. See if you can apply what you have learnt about evaluating and applying the validity of information presented in the following examples from the media and research literature.

Do not be put off by jargon that you do not understand, particularly if you are reading the methods and results of studies published in medical journals. The abstracts of many studies are structured to make it easier to assess the quality of the study. In general you can skip the confidence intervals (CIs) if you don’t understand them, but, if you want more information on how to interpret confidence intervals, see Chapter 18. Relative risks and odds ratios are also explained more fully in that chapter. Do not be concerned if you do not understand statistical techniques and jargon. Skip over them. Most of the important study flaws have to do with study design rather than statistical errors, so rest assured that, with the help of the validity guides described in Chapters 9 and 10, you should be able to make good sense of the majority of health information, without having to understand the details of statistical methods.

Our appraisal of the evidence is included, but try not to look at this until you’ve made your judgement.

Please note that the examples in this chapter have been selected so that we can look at their method and validity; they do not necessarily represent the most up-to-date information on a particular topic, and should not be used to inform your own decisions.
CHAPTER 15

Remember, when assessing evidence on TREATMENTS, that you want to know:

1. Is there an evidence-based guideline or systematic review of randomised trials?
2. Is the evidence relevant to your needs?
   - Are the outcomes person centred rather than theoretical? (What should work in theory doesn’t always work in practice.)
   - Does the evidence describe both benefits and harms and say how likely they are to occur?
   - Does the evidence describe how treatments compare with other appropriate options?
3. Is the evidence reliable?
   - Is there a systematic review of randomised controlled trials or evidence-based guideline?
   - Are there randomised controlled trials?
   - Are there any non-randomised studies (in order of priority):
     - cohort and non-randomised trials?
     - population-based case–control studies?
     - hospital-based case–control studies?
     - other types of studies?
   - Are there any case reports or opinions?

When assessing evidence on CAUSES of disease or conditions you want to know:

1. Is the evidence about a causal relationship from a reliable source?
   - Is there a systematic review of randomised controlled trials?
   - Are there randomised controlled trials?
   - Are there any non-randomised studies (in order of priority):
     - cohort and non-randomised trials?
     - population-based case–control studies?
     - hospital-based case–control studies?
     - other types of studies?
   - Are there any case reports or opinions?
2. Did exposure to the supposed cause occur before the outcome?
3. Is there a strong relationship between the supposed cause and the outcome?
4. Does the risk of the outcome increase as the dose or length of time of the supposed cause increases?
5. Is there any other explanation for the relationship between the supposed cause and the outcome?
6. Are there other studies showing the same results?
7. Does the ‘cause-and-effect’ relationship make sense?

**Example 1: Making sense of a study**


**Abstract**

**OBJECTIVES:** To investigate whether breastfeeding is effective for pain relief during venepuncture in term neonates and compare any effect with that of oral glucose combined with a pacifier.

**DESIGN:** Randomised controlled trial.

**Participants:** 180 term newborn infants undergoing venepuncture; 45 in each group.

**Interventions:** during venepuncture infants were either breastfed (group 1), held in their mother’s arms without breastfeeding (group 2), given 1 ml of sterile water as placebo (group 3) or given 1 ml of 30 per cent glucose followed by pacifier (group 4). Video recordings of the procedure were assessed by two observers blinded to the purpose of the study.

**Main outcome measures:** pain-related behaviours evaluated with two acute pain rating scales: the Douleur Aiguë Nouveau-né scale (range 0–10) and the premature infant pain profile scale (range 0–18).

**RESULTS:** Median pain scores (interquartile range) for breastfeeding, held in mother’s arms, placebo and 30 per cent glucose plus pacifier groups were 1 (0–3), 10 (8.5–10), 10 (7.5–10) and 3
(0–5) with the Douleur Aiguë Nouveau–né scale, and 4.5 (2.25–8), 13 (10.5–15), 12 (9–13) and 4 (1–6) with the premature infant pain profile scale. Analysis of variance showed significantly different median pain scores ($p < 0.0001$) among the groups. There were significant reductions in both scores for the breastfeeding and glucose plus pacifier groups compared with the other two groups ($p < 0.0001$, two-tailed Mann–Whitney U tests between groups). The difference in Douleur Aiguë Nouveau–né scores between breastfeeding and glucose plus pacifier groups was not significant ($p = 0.16$).

**CONCLUSIONS:** Breastfeeding effectively reduces response to pain during minor invasive procedure in term neonates.

**Our appraisal**

This shows just how straightforward it can be to appraise a structured abstract. You don’t have to be a scientist!

- It is controlled, and compares four groups – breastfeeding, being held in mother’s arms without feeding, sterile water (control), sugar followed by a pacifier.
- It is randomised.
- It is a double-blind trial. This tells you that the babies, not surprisingly, were unaware of which group they were in, and the investigators listening to the tapes were unaware of which babies were given which intervention.
- There was complete follow-up.

This is a randomised controlled trial, the best type of study for evaluating the effects of an intervention. It meets the validity criteria, and shows that breastfeeding reduces crying in babies when they have a painful procedure such as having blood taken. In addition, it shows that sugar followed by a pacifier was also an effective strategy to reduce crying but not as effective as breastfeeding. We should be careful, however, when applying these results more generally. As with most studies, not all the important issues will have been assessed. For example, this study does not report on the mothers’ stress levels if holding or breastfeeding their baby while a blood test is being taken.
Example 2: Should a woman in her 40s have screening mammography?

As a 40-year-old woman, you are concerned with maintaining your health. You see a sticker advertising Breast Screening for Women 50 and over . . . it’s FREE!

While taking your son to the GP, you pick up a brochure entitled Early Detection, the Best Protection, which tells you about the National Program for the Detection of Breast Cancer. It tells you that 1 in 14 women will develop breast cancer in her lifetime and that over 70 per cent of breast cancers occur in women over 50. Further on you read that the target population for mammographic screening is 50–69 years, and that screening in this age group reduces deaths from breast cancer. It goes on to say that the evidence of benefit is not strong enough to recommend routine screening for women in the 40- to 49-year age group and adds: (Women in this age group can be screened if they wish.)

Based on this information, how do you make the decision about whether you should be screened? You phone your local mammography clinic who send you a leaflet. You find a table showing the probability of developing breast cancer (incidence), and dying from breast cancer, at different age intervals. Here you see that your risk of developing breast cancer during the decade 40–49 years is about 14 cancers per 1000 women, and your risk of death from breast cancer is about 2.5 deaths per 1000 women. The risks increase with age.

You feel that there is still more you need to know. In particular, you want to know whether the risk of death is reduced by regular screening in your age group and what the possible harms are. You have read the chapter on finding the best evidence and using ‘Google’ you find an online decision aid about breast cancer screening in women aged 40–49 years (www.mammogram.med.usyd.edu.au). This shows you that the risk of dying from breast cancer between ages 40 and 49 changes from 2.5 per 1000 women without screening to 2.0 per 1000 women with screening. The decision aid also tells you about the accuracy of screening and the chance of
each possible outcome. It explains that over the same decade of 2-yearly screening:

- 21 women are diagnosed with breast cancer over the next 10 years
- 12 women have their cancer detected by screening
- 9 women develop symptoms and are diagnosed with breast cancer between screening mammograms
- 239 women have extra tests after an abnormal mammogram. The extra tests will show that these women don’t have breast cancer. Aside from the inconvenience of attending for these tests, some women will worry long after they have had them
- 740 women are correctly reassured that they do not have breast cancer.

Our appraisal

Weighing up the benefits and the harms of screening should be done on an individual basis. Most people believe that screening for cancer is something worthwhile but it’s important to understand that these tests are not perfect. Some, but not all, cancers will be detected by screening. Sometimes test results can be falsely positive and invasive follow-up tests might be performed on a completely healthy person. Different people will place different values on the benefits and risks of screening. For some the chance that one life might be saved over 10 years is highly regarded. For others, it might not seem a substantial benefit for them. Many women won’t mind having extra follow-up tests for an abnormal mammogram provided that they get the ‘all clear’ in the end. Others will be annoyed that this extra stress, expense and inconvenience have been caused. There’s no right or wrong answer. Everyone will be different.

Evidence-based decision aids to help you weigh up the benefits and risks of screening are available through the Sydney Health Decision Group (mammography in 40- to 49-year-old women at www.mammogram.med.usyd.edu.au and FOBT [faecal occult blood test] screening for bowel cancer at www.cancerscreeningdecision.org).
Example 3: Product information – Blackmores Hyperiforte 1800 (St John’s wort)

You are feeling anxious and depressed and have spoken to your practitioner about how you are feeling. He says you could take antidepressants, but your symptoms are relatively mild and may not warrant the risk of side effects. He suggests that you try St John’s wort (Hypericon) which is available from chemists or health food shops without a prescription and seems safe. When you are next at the shops, you find a bottle of St John’s wort, and are surprised to find a reference to an article in a medical journal on the product information on the side of the bottle. It reads:

Blackmores Hyperiforte 1800 helps relieve nervous tension, stress and mild anxiety. Hyperiforte 1800 is formulated to replicate the dose used in clinical trials, where it was demonstrated to be as effective as prescription drugs but with fewer side effects.

Vorbach et al.1

You could find the abstract of the article on PubMed (www.ncbi.nlm.nih.gov/PubMed), which is free to use, but it might be more sensible to see whether there is any more recent and complete information. You look at the Cochrane Library home page (www.thecochranelibrary.com) and a search for ‘wort’ finds the following:

St John’s Wort for Depression Linde K, Mulrow CD Date of most recent substantive amendment: 25/02/2005

Plain language summary

Available evidence suggests that several specific extracts of St John’s wort may be effective for treating mild to moderate depression, although the data are not fully convincing.

Extracts of St John’s wort (botanical name Hypericum perforatum L.) are prescribed widely for the treatment of depression. They seem more effective than placebo and similarly effective as standard antidepressants for treating mild to moderate depressive symptoms.
Beneficial effects for treating major depression appear minimal. Side effects are usually minor and uncommon. However, as extracts of St John’s wort can influence adverse effects of other drugs, patients should consult their physicians before using St John’s wort. The results of this review apply only to the preparations tested in trials; the content of marketed preparations might vary considerably from those tested in trials.

Linde et al.4

Our appraisal

This is a good systematic review. There are many trials, showing St John’s wort to be beneficial when compared with placebo, and that it may be as effective as standard antidepressants, but with fewer side effects. You may decide that it is worth taking.

Example 4: Keyhole or open surgery for a hernia? Which is best for me?

You are a middle-aged man with a swelling in your groin that your doctor tells you is an inguinal hernia. Although there is no urgency, you decide to have it operated on because it is uncomfortable and interferes with your work and leisure activities. While chatting to a friend, he tells you he also had a similar operation a while ago and noticed that there was some press coverage about inguinal hernias at that time. After some searching among some old papers in his filing cabinet, he finds an article from a newspaper which reads:

It was business as usual for Dr Michael Aroney as he performed keyhole surgery to repair a hernia on a 62-year-old man injured at work.

The operation, which Dr Aroney performed yesterday at the Holroyd Private Hospital, Guildford, is a typical example of the problems facing the health funds.

The treatment – the finest available – means the man will be able to return to work within two weeks. But this comes at a price. The
operation will cost nearly $1,000 more than conventional surgery, which would have kept the patient out of work for at least six weeks.

‘Medicine is not cheap,’ said Dr Aroney. ‘It comes at a price.

‘People have high expectations, and those high expectations require high-tech medicine, and that does not come cheaply.’

Health funds are keen to introduce a system where decisions to operate in the most modern manner possible will come under more rigorous scrutiny. For example, should Dr Aroney have used cheaper, more conventional surgical techniques?

Was surgery even necessary? It would have been possible for the man to have been prescribed a truss at virtually no cost. Such a system was not mentioned in yesterday’s report, but it may not be far off.

The Government has not ruled out further inquiries into how to control costs in hospitals, even to the extent of deciding which is the most appropriate and most effective treatment.

The minister has guaranteed that the final decision will be with doctors, but not necessarily the surgeon performing the surgery.

Dr Aroney, like many doctors, believes it would be the wrong way to go, saying:

‘Doctors jealously guard the fact that they have a patient–doctor relationship and they are directly responsible to their patients, rather than to a third party.’


You want to have some say in the operation you have, particularly as it is important to you to return to work soon as you work on your own and have difficulty in taking time off. You chat to your doctor about surgeons and ask which surgeons are best and whether any of the ones whom she recommends has experience with ‘laparoscopic’ (keyhole) surgery for hernia. (As discussed in Chapter 4,
technical expertise hangs on three broad criteria: qualification to perform the procedure, experience in performing the procedure, and being part of a quality assurance scheme or some similar credentialing programme.)

You also ask what the evidence is that keyhole surgery does as well as open surgery. Your doctor checks for systematic reviews on the Cochrane Library online while you are there and finds one: McCormack K, Scott N, G. P, Ross, S and Grant A, Laporoscopic techniques versus open techniques for inguinal hernia repair. There was no plain language summary at the time this book was published.

Abstract

OBJECTIVES: To compare minimal access laparoscopic mesh techniques with open techniques.
SEARCH STRATEGY: We searched Medline, Embase and the Cochrane Central Controlled Trials Registry for relevant randomised controlled trials. The reference list of identified trials, journal supplements, relevant book chapters and conference proceedings were searched for further relevant trials. Through the EU Hernia Trialists Collaboration (EUHTC) communication took place with authors of identified randomised controlled trials to ask for information on any other recent and ongoing trials known to them.
SELECTION CRITERIA: All published and unpublished randomised controlled trials and quasi-randomised controlled trials comparing laparoscopic groin hernia repair with open groin hernia repair were eligible for inclusion.
DATA COLLECTION AND ANALYSIS: Individual patient data (IPID) were obtained, where possible, from the responsible trialist for all eligible studies. Where IPD was unavailable additional aggregate data were sought from trialists and published aggregate data checked and verified by the trialists. Where possible, time to event analysis for hernia recurrence and return to usual activities were performed on an intention to treat principle. The main analyses were based on all trials. Sensitivity analyses based on the data source and trial quality were also performed. Predefined subgroup analyses based on recurrent hernias, bilateral hernias and femoral hernias were also carried out.
MAIN RESULT: Forty-one eligible trials of laparoscopic versus open groin hernia repair were identified involving 7161 participants (with individual patient data available for 4165). Meta-analysis was performed, using IDP where possible. Operation times for laparoscopic repair were longer and there was a higher risk of rare serious complications. Return to usual activities was faster, and there was less persisting pain and numbness. Hernia recurrence was less common than after open non-mesh repair but not different to open mesh methods.

AUTHORS’ CONCLUSIONS: The review showed that laparoscopic repair takes longer and has a more serious complication rate in respect of visceral (especially bladder) and vascular injuries, but recovery is quicker with less persisting pain and numbness. Reduced hernia recurrence rates of around 30–50 per cent were related to the use of mesh rather than the method of mesh placement.

Our appraisal
This systematic review combines the results of 41 randomised trials that involved over 7000 people in total. It tells you that laparoscopic repair allows you to get back to work quicker and that persistent pain and numbness are less likely. However, the operation takes longer to perform and rare serious complications of the bladder and blood vessels are more likely. Hernia recurrence rates are about the same as open-mesh surgery. We had to look into the main part of the review to find the complication rates for laparoscopic repair. Although the abstract doesn’t mention it, the laparoscopic complication rate is about 87 per 1000 operations with blood clots, 15 per 1000 with wound infections and 3 in 1000 with bladder damage. Compare this to open-mesh complications rates and you have 107 blood clots per 1000, 31 wound infections per 1000 and fewer than 1 per 1000 cases of bladder damage.

Your doctor explains that you need to weigh up the slightly higher chance of more serious complications against the better short-term outcomes including earlier return to work. As early return to work is critical to you at the moment, you decide to have a laparoscopic repair if there is a surgeon in your town who has experience with the operation.
Example 5: Do mobile phones cause brain cancer?


Mobile phone use does not raise the risk of cancer, at least in the first 10 years of use, the largest investigation to date shows.

Some past studies had suggested an increased risk of acoustic neuroma – a tumour of the nerve connecting the ear and the brain – but others did not.

The latest Institute of Cancer Research work includes data from five European countries and more than 4000 people.

Expert advice is still to limit mobile phone use as a precautionary measure.

There are more than one billion mobile phone users worldwide.

Longer follow-up is needed to check that health problems do not arise with many more years of use, the researchers say in the British Journal of Cancer.

An independent group for the UK government, led by Sir William Stewart, that looked into the safety of mobile phones in the late 1990s also concluded that mobile phones did not appear to harm health.

However, the group said that there was evidence that radiation from mobile phones could potentially cause adverse health effects, and therefore a ‘precautionary approach’ to their use should be adopted.

Precautions

The government currently advises mobile phone users to keep their call times short.

And children under the age of 16 should use mobile phones for essential calls only, because their head and nervous systems may still be developing.
MAKING SENSE OF HEALTH ADVICE

The latest data from the UK, Denmark, Finland, Norway and Sweden included 678 people with acoustic neuroma and 3553 without this form of tumour.

This revealed no relationship between the risk of acoustic neuroma and the number of years for which the mobile phones had been used, the time since first use, total hours of use or total number of calls.

Nor was there any link with analogue or digital phones or whether or not a hands-free kit was used.

On balance, the evidence suggests that there is no substantial risk of acoustic neuroma in the first decade of use – but the possibility of some effect after longer periods remains open, the researchers concluded.

Senior investigator Professor Anthony Swerdlow said: ‘Whether there are longer-term risks remains unknown, reflecting the fact that this is a relatively recent technology.’

Dr Michael Clark from the Health Protection Agency said: ‘This is good news but we still need to be a bit cautious.’

Dr Julie Sharp, senior science information officer at Cancer Research UK, said: ‘This study provides further evidence that using mobile phones does not increase the risk of brain tumours.

‘However, it is important that researchers continue to monitor phone users over the coming years as mobiles are still a relatively new invention.’

The research is part of a bigger study that will be published next year.

A Swedish study identified an increased risk of acoustic neuromas among people who had used mobile phones for 10 years or more.

People have been concerned that the radiofrequency from phones might cause cancers, despite the absence of a known biological mechanism for this.
Our appraisal

The evidence about the relationship should be from a reliable source and the best study type

This news article doesn’t tell us what type of study the claim is based on, only that it has been conducted across five European countries. If you go to PubMed and type in (acoustic neuroma) AND (mobile phone) AND 2005, you will find that this is describing a population-based case–control study in which the prior use of mobile phones is measured in people with acoustic neuroma and in a sample of the general population. Randomised trials are impractical because one could not randomise people to use mobile phones. Cohort studies, the next best design, would be very difficult to do because acoustic neuromas are quite rare, only occurring in about 6 out of every 100,000 people each year. This means that a cohort study would require follow-up of millions of mobile phone users and non-users over a decade or more. Population-based case–control studies are therefore likely to be about as good as it gets. You can read more about cohort and case–control studies in Chapter 10.

The exposure to the supposed cause should occur before the outcome

From reading this news story, it seems that they asked all participants about mobile phone use over the preceding 10 years. Although this does record mobile phone exposure before developing an acoustic neuroma, it is prone to bias because people’s recollections about use over such a long period of time might not be very accurate. It might be that people who have a brain tumour tend to over-recall higher phone usage in hindsight compared with those who don’t have a brain tumour.

There should be a strong relationship between the supposed cause and the outcome

The study didn’t find any overall association between mobile phone use and acoustic neuroma at least within 10 years.
There should be a dose–response or exposure–response relationship between the supposed cause and outcome, that is, the greater the exposure, the more likely someone is to get a disease. The study looked at the number of hours of mobile phone use and the number of years of mobile phone use and did not find any relationship to cancer risk.

There should not be any other factors that could explain the relationship. Actually as we don’t know much about the cause of brain tumours, it is difficult to know whether other factors could play a part in the development of acoustic neuroma.

The same results should be shown in several studies. Looking at PubMed by typing in (mobile phone) AND (brain cancer) we can see that the interphone study involving the five countries that this BBC item refers to is the main study that has been conducted. It has shown that there is also no link between mobile phone use for less than 10 years and gliomas (another type of brain tumour).7

The relationship should make sense. This has been addressed at the end of the news item where it states that people’s concerns have not been based on a known biological cause.

The bottom line on this question is that there is reasonable evidence that mobile phone use does not increase your risk of brain cancer within 10 years. The type of study design is a population-based case–control study, which is the best study type that is feasible for a rare condition. As the BBC has suggested, mobile phone technology is new and we don’t know about potential risks after 10 years of use.

Applying the principles we have set out in this book, you will hopefully be weighing up the benefits and convenience of mobile
phone use in your own situation against the evidence to date. Researchers have not shown any link so far and it’s important to bear in mind that brain cancers are less common than many others.

References


16

Is this a useful diagnostic test?

The next three chapters have been provided for those readers who really want to understand and learn some basic epidemiological skills. You may be a health consumer who has really found this book interesting and wants to go a bit further. You may be a health practitioner or practitioner in training and want to brush up on some skills in evidence-based practice. Whoever you are, if you are the sort of person who does not like numbers, you might want to skip over this part.

**Sensitivity and specificity of a diagnostic test**

- **Sensitivity** indicates the probability that the test will accurately pick up disease when there truly is disease.
- **Specificity** indicates the probability that the test will accurately detect ‘NO disease’ when the disease is truly absent.

To illustrate these, imagine I have a bag of toffees, some of which are liquorice flavoured (L) and some of which are not (NOT-L). L and NOT-L toffees have a slightly different shape so it’s easy for me (or so I believe) to feel which is which without looking. To see how accurate I am at detecting which are which, I try it out.

This is the result: of 100 L toffees, my hand correctly calls 80 of them L toffees. Of 100 NOT-L toffees, my hand correctly calls 90 of them NOT-L toffees.
CHAPETR 16

In technical jargon, if I consider my hand as a diagnostic test, it has a sensitivity of 80 per cent (the proportion of L toffees that I correctly identified) and a specificity of 90 per cent (the proportion of NOT-L toffees that I correctly identified).

**Pre-test and post-test probability**

Now if I put my hand in a bag of toffees and say ‘This is a liquorice toffee’, what are my chances of being correct? Well, I cannot tell what my chances are of being right unless I know something about the existing probability of finding a liquorice toffee. This is referred to as the pre-test probability of an L toffee. For example, if there are no L toffees in that bag, all of those that I call L would be wrong calls. On the other hand, if the bag contains only L toffees, all those that I call L will be correct calls (and, of course, any ‘NOT-L’ calls would be wrong!). So even though I may know the sensitivity and specificity of my hand as a test, I need more information to interpret the test result.

Clearly the interpretation of the test depends on what percentage of the toffees in the bag were L or NOT-L before I put my hand in it. Put another way, it depends on the pre-test probability of a

<table>
<thead>
<tr>
<th></th>
<th>Truly L toffees</th>
<th>Truly NOT-L toffees</th>
<th>Total</th>
<th>Probability of a toffee being L</th>
</tr>
</thead>
<tbody>
<tr>
<td>I think that they are L toffees</td>
<td>80</td>
<td>30</td>
<td>110</td>
<td>Post-test prob. for a positive test = 80/110 = 73%</td>
</tr>
<tr>
<td>I think that they are NOT-L toffees</td>
<td>20</td>
<td>270</td>
<td>290</td>
<td>Post-test probability for a negative test = 20/290 = 7%</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>300</td>
<td>400</td>
<td>Pre-test probability = 100/400 = 25%</td>
</tr>
</tbody>
</table>

Table 16.1 My probability of correctly detecting L toffees: pre-test probability of 25 per cent

206
tess being L. Now if I knew that, working out the result of my test would be easy. Here are a few numerical examples of the toffee test:

Suppose I have a bag of 400 toffees, of which 25 per cent (i.e. 100 toffees) are L. If I have been told that this is the case, I can apply the sensitivity and specificity of my hand to this set of information as shown in Table 16.1.

In Table 16.1, the sensitivity is 80 per cent (80/100) and the specificity 90 per cent (270/300). I know this from applying my known sensitivity and specificity in detecting L and NOT-L toffees as described earlier. Now, if I put my hand in and detect a toffee as L the probability of being correct is 73 per cent (80/110). If I think that it is NOT-L, of course, there is still a chance that it actually is L – a 7 per cent (20/290) chance to be precise.

Now, let’s imagine that I am given another bag of 400 toffees and, this time, 75 per cent of them are L toffees instead of 25 per cent. Needless to say, the sensitivity and specificity of my hand (remember my hand is the diagnostic test) remain the same, so this time the table would be as shown in Table 16.2.

Now, if I say I think that a toffee is L, I will be correct 96 per cent of the time and, if I identify it as NOT-L, there is a 40 per cent

<table>
<thead>
<tr>
<th></th>
<th>Truly L toffees</th>
<th>Truly NOT-L toffees</th>
<th>Total</th>
<th>Probability of a toffee being L</th>
</tr>
</thead>
<tbody>
<tr>
<td>I think that they are L toffees</td>
<td>240</td>
<td>10</td>
<td>250</td>
<td>Post-test probability for a positive test = 240/250 = 96%</td>
</tr>
<tr>
<td>I think that they are NOT-L toffees</td>
<td>60</td>
<td>90</td>
<td>150</td>
<td>Post-test probability for a negative test = 60/150 = 40%</td>
</tr>
<tr>
<td>Total</td>
<td>300</td>
<td>100</td>
<td>400</td>
<td>Pre-test probability = 300/400 = 75%</td>
</tr>
</tbody>
</table>
chance that it turns out to be L, which translates to a 60 per cent chance that I will be right in my call.

In medical jargon, then, the PRE-TEST PROBABILITY is the probability that L toffees are in the bag before I put my hand in it or, more appropriately, the probability that there really is disease before a diagnostic test is carried out. The POST-TEST PROBABILITY if the test turns out to be POSITIVE is the probability that I will detect an L toffee when it truly is one, whereas the POST-TEST PROBABILITY if the test is NEGATIVE is the probability of it really being L when I judge it not to be. In terms of disease, it is the probability of the existence of disease when the test detects no disease.

Table 16.3 Summary of all the above information

<table>
<thead>
<tr>
<th>Pre-test probability of L (%)</th>
<th>Post-test probability of a positive test (%)</th>
<th>Post-test probability of a negative test (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>73</td>
<td>7</td>
</tr>
<tr>
<td>75</td>
<td>96</td>
<td>40</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

In summary, the post-test probability of disease given a diagnostic test result depends on the sensitivity and specificity of the test AND on the pre-test probability. There is no such thing as being absolutely certain of what a test result means; it varies from one patient to another depending on his or her pre-test probability. For instance, a positive HIV test in an intravenous drug user means something different to a positive HIV test on a blood donation from, say, a nun. For the drug user the pre-test probability may be appreciable and a positive test is likely to indicate HIV. In the nun, on the other hand, the pre-test probability is close to zero and any positive test is likely to be a false positive.

Note that the post-test probabilities for negative and positive tests straddle the pre-test probability – that is, a positive test increases the probability of disease above the pre-test level, whereas
a negative test decreases it to below the pre-test level. When you do choose to have a further diagnostic test, use of pre-test and post-test probabilities will tell you how the test results affect your chances of having the disease.

To give you some examples, screening mammography has a sensitivity of about 85 per cent in women over 50 and about 70 per cent in women aged between 40 and 49. Specificity is about 95 per cent – that is, about 5 per cent of women without cancer will require some further investigation. Ultrasound has a sensitivity of about 85 per cent and a specificity of 90 per cent in detecting blockage in the arteries to the brain. However, ultrasound has a sensitivity of only about 60 per cent and a specificity of 97 per cent for detecting clots in the veins in the legs after operations. Of course, one can also assess the accuracy of symptoms and the signs. For example, if you are admitted to hospital with possible appendicitis, the pain in the bottom right part of your abdomen has a sensitivity of 81 per cent and a specificity of 53 per cent.

Like you, many practitioners find this complex. This type of information may not have been part of their medical training. Consequently, you are likely to find the answers about your pre-test and post-test probabilities less satisfactory than answers about the effects of treatments.

References

Suppose that you have a cough that has persisted for several days. Your doctor tells you that it should be treated with antibiotics if it is a bacterial infection. You could:

- wait to see if it clears up without treatment
- treat with antibiotics in case there is a bacterial chest infection
- have a test to establish whether there is a bacterial chest infection.

How you decide goes something like this: if the probability that the cough is caused by a bacterial chest infection is 0 per cent, you clearly would not have antibiotics. On the other hand, if the probability is 100 per cent, you would take antibiotics. Anywhere between these extremes a diagnostic test would help to determine the chance that the infection is bacterial. So why not just go ahead and have a test? For two reasons: first, there are almost always harms associated with tests and, second, a test does not guarantee absolute certainty about the diagnosis. One way of dealing with this uncertainty is to use the decision threshold, which is illustrated by the following example.

Say your practitioner tells you that the probability of your having a bacterial infection is less than 10 per cent. You might decide to wait a few days to see if you feel better without any
treatment. Ten per cent may be your threshold below which the harms of a test or treatment outweigh its potential benefit. If, on the other hand, your practitioner thinks there is a very high probability of a bacterial infection, say above 90 per cent, a course of antibiotics may be the best way to go. Ninety per cent, in this case, may be your threshold above which treatment is advisable without undergoing a test. This is because the benefit of treatment outweighs the potential harms of not treating, of the test itself and of the delay while waiting for a test result.

But what if your practitioner tells you that the probability of bacterial infection is about 50 per cent? This is between your thresholds for doing nothing (below 10 per cent) or taking antibiotics (above 90 per cent). In this case, having a diagnostic test to establish the real cause of the cough may be a good choice. These concepts are summarised in the Figure below.

To summarise, using the decision threshold method of deciding the best course of action requires that you and your practitioner have an estimate of your chance of having a particular disease, and comparing that to thresholds below which it would be best not to treat and above which it would be best to treat. Only between these thresholds can a test be of any help.
How do you or your practitioner decide exactly where the thresholds are? It really depends on the benefits and harms of treating or not treating. Suppose, for example, a safe and effective treatment exists for a disease that you may have. This would lower the threshold above which to treat because there would be less reason to avoid treatment. By the same token, if the only diagnostic test available is invasive (an aggressive procedure) and not very accurate, you may decide either to "wait and watch" or to treat (just in case) and avoid the test.
Relative and absolute risks

How do you interpret the results of a randomised controlled trial? A common measure of a treatment is to look at the frequency of bad outcomes of a disease in the group being treated compared with those who were not treated. For instance, supposing that a well-designed randomised controlled trial in children with a particular disease found that 20 per cent of the control group developed bad outcomes, compared with only 12 per cent of those receiving treatment. Should you agree to give this treatment to your child? Without knowing more about the adverse effects of the therapy, it appears to reduce some of the bad outcomes of the disease. But is its effect meaningful?

This is where you need to consider the risk of treatment versus no treatment. In healthcare, risk refers to the probability of a bad outcome in people with the disease.

Absolute risk reduction (ARR) – also called risk difference (RD) – is the most useful way of presenting research results to help your decision-making. In this example, the ARR is 8 per cent (20 per cent - 12 per cent = 8 per cent). This means that, if 100 children were treated, 8 would be prevented from developing bad outcomes. Another way of expressing this is the number needed to treat (NNT).
If 8 children out of 100 benefit from treatment, the NNT for one child to benefit is about 13 (100 ÷ 8 = 12.5).

For technical reasons, some other measures are often used. The relative risk (RR) of a bad outcome in a group given intervention is a proportional measure estimating the size of the effect of a treatment compared with other interventions or no treatment at all. It is the proportion of bad outcomes in the intervention group divided by the proportion of bad outcomes in the control group. In this hypothetical case, the RR is 0.6 (12 per cent ÷ 20 per cent = 0.6).

When a treatment has an RR greater than 1, the risk of a bad outcome is increased by the treatment; when the RR is less than 1, the risk of a bad outcome is decreased, meaning that the treatment is likely to do good. For example, when the RR is 2.0 the chance of a bad outcome is twice as likely to occur with the treatment as without it, whereas an RR of 0.5 means that the chance of a bad outcome is twice as likely to occur without the intervention. When the RR is exactly 1, the risk is unchanged. For example, a report may state ‘The relative risk of blindness in people given drug T was 1.5’. This shows that the drug increased the risk of blindness. Another measure that is used is the odds ratio. For practical purposes, assume that the odds ratio is the same as the relative risk. Sometimes the outcome is a good one and the interpretation of relative risk is the opposite of what we have just outlined.

Relative risk reduction (RRR) tells you by how much the treatment reduced the risk of bad outcomes relative to the control group who did not have the treatment. In the previous example, the relative risk reduction of fever and rash in the group of the children on the intervention was 40 per cent (1 – 0.6 = 0.4 or 40 per cent).

<table>
<thead>
<tr>
<th>% Control with poor outcomes</th>
<th>% intervention with poor outcomes</th>
<th>RR</th>
<th>RRR</th>
<th>ARR (%)</th>
<th>NNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>40</td>
<td>0.67</td>
<td>0.33</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
<td>0.6</td>
<td>0.33</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>0.67</td>
<td>0.33</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.67</td>
<td>0.33</td>
<td>2</td>
<td>50</td>
</tr>
</tbody>
</table>
Relative risk

The RR (and therefore the RRR) is often the same in people irrespective of their level of risk, which means that the ARR will be greatest in those at greatest risk, as shown in Table 18.1. The greater your risk, the more you stand to gain from the intervention.

Confidence intervals

Confidence intervals (CIs) aim to give you an idea of how confident you can be about a study’s estimate of a treatment’s effects. Even when a study is of impeccable quality, the results may have happened by chance. Statisticians deal with this uncertainty by doing some nifty calculations to determine how confident one can be about the results, which give us the confidence interval. The narrower the range, the more precise the study’s estimates, and the more confident you can be that it is a ‘real’ finding and not due to chance.

This is usually expressed in terms of a 95 per cent confidence interval (95%CI), which represents the range of results within which we can be 95 per cent certain that the true answer lies.

As an illustration of how confidence intervals can help, imagine that you are doing a study investigating whether there is gender bias in the method used by a university to choose its students. If there were no such bias you would expect 50 per cent of its students to be men and 50 per cent to be women. Supposing that you check a small sample, say 10 students, and found 4 of them were men. How sure can you be that this is a true reflection of the student population? Statistical calculations show that you can be 95 per cent certain that the true quota of men in the entire university population is somewhere between 12 and 74 per cent. This is an unhelpfully wide range.

But supposing you randomly sample 100 students and find that 40 are men. Statistical calculations show that you can be 95 per cent certain that the true quota of men in the entire university population is somewhere between 30 and 50 per cent – a narrower range.

Imagine also that you randomly examined a large sample of 1000 students, of whom 400 were men. The 95%CI would be from 37 per cent to 43 per cent – a much narrower range showing a very
high level of confidence that this represents a true reflection of the gender ratio in the university.

In the sample of 10 students, finding four men is compatible with our expected value for society at large – 50 per cent males and 50 per cent females. In the group of 1000 students finding that only 40 per cent are men is not expected. The result from this large sample is statistically significant, which means that the disparity between the observed 40 per cent and the expected 50 per cent is real – that is, it is very unlikely to have arisen by chance. In the sample of 100 students, the upper end of the confidence interval is just on the expected value of 50 per cent and therefore just statistically significant.

The same principle applies to studies investigating treatments, except that we might be looking at the relative risks of a poor outcome in the group receiving the intervention compared with the control group.