Diagnosis in primary care: probabilistic reasoning

Bruce Arroll MBChB, PhD;[1] **GM Allan** MD, CCFP;[2] **C Raina Elley** MBChB, PhD;[1] **Tim Kenealy** MBChB, PhD;[1] **James McCormack** Pharm D;[3] **Ben Hudson** MBBS, MRCGP;[4] **Karen Hoare** MSc[1]

[1] Department of General Practice and Primary Health Care, The University of Auckland, New Zealand

[2] Department of Family Medicine, University of Alberta, Edmonton AB, Canada

[3] Faculty of Pharmaceutical Sciences, University of British Columbia, Canada

[4] Department of General Practice, University of Otago, Christchurch, New Zealand

ABSTRACT

This article develops the concept of probabilistic reasoning as one of the techniques clinicians use in making a diagnosis. We develop the concept that every question and every examination is a diagnostic test ultimately leading to a rule in or rule out of a diagnosis. We also develop the concept of pre-test probability pointing out that false positive tests are an issue in low-prevalence settings and false negative tests are a problem. Investigative tests work best in medium-prevalence settings. The purpose of taking a history and conducting an examination is to increase the pre-test probability to a point where either treatment is commenced or more expensive/time-consuming/dangerous tests are indicated. Pre-test probabilities on their own can be used to rule out conditions. We also show how pre-test probabilities relate to the Fagan nomogram which enables visualisation of large changes in post-test probabilities which can lead to treatment/further investigation.

KEYWORDS: Likelihood ratio; pre-test and post-test probability; diagnostic accuracy; probabilistic reasoning

J PRIM HEALTH CARE 2012;4(2):166–173.

CORRESPONDENCE TO:
Bruce Arroll
Professor of General Practice and Primary Health Care, The University of Auckland, PB 92109 Auckland 1142, Auckland, New Zealand
b.arroll@auckland.ac.nz

Henegan and colleagues divide diagnostic reasoning into three stages—initiation, refinement and defining the final diagnosis. Probabilistic reasoning is one of five potential strategies to be used in the refinement stage (Figure 1).[1] They recognise that clinicians do not follow a rigid sequence of categories, and that different clinicians could appropriately use different combinations of strategies to diagnose the same condition.

The process of diagnosis in medicine may seem ‘intuitive’ and based on clinical experience or ‘pattern recognition’. However, there is also a ‘logic’ and a set of assumptions that underlie much of the diagnostic process. The mathematical principles of the diagnostic process can be made explicit (in our experience they rarely are), and many of the assumptions can be located in the evidence. This article discusses this ‘logic’ and use of assumptions, collectively known as ‘probabilistic reasoning’. We anticipate that the more medical students and physicians/nurses understand about these principles, the deeper will be their thinking. This may or may not produce better clinical outcomes, but we speculate this may increase clinician satisfaction.

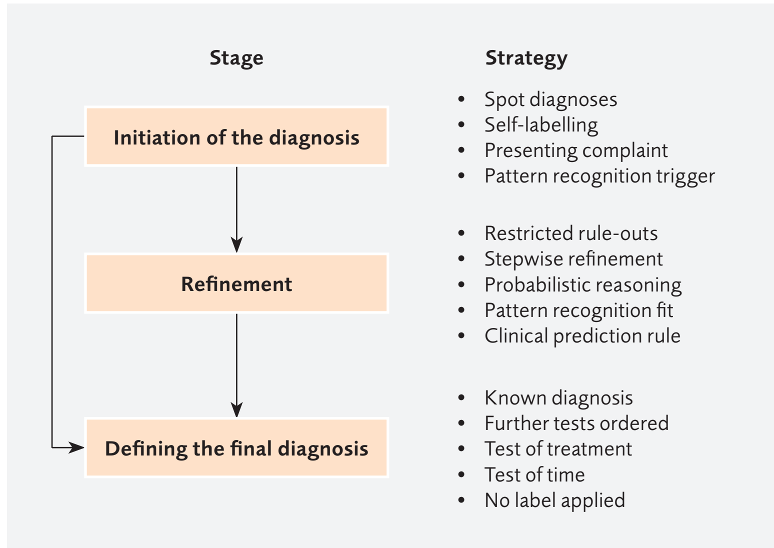
Probabilistic reasoning is the use of the diagnostic value of specific symptoms, signs, or tests to rule in or rule out a diagnosis. A diagnosis is generally not completely ruled in or ruled out at the time of the first management decision in clinical care. Probabilistic reasoning requires knowing, or estimating, two key pieces of information: the prevalence or pre-test probability of the diagnosis being considered, and the degree to which a positive or negative result from a specific test adjusts the probability of that diagnosis.[1] Knowledge of the pre-test probability is followed by early hypothesis generation and the hypothetic-deductive process of attempting to defend or refute a particular hypothesis.[2] This is carried out by asking a series of condition-specific questions and examining physical signs. Each question and physical examination is a diagnostic test in itself with a sensitivity and specificity (sensitivity and specificities can be converted to likelihood ratios which are very useful and will be discussed later—Box 1). This may be followed by laboratory or other diagnostic tests, each with their own sensitivity and specificity.

The diagnostic questioning process continues until a diagnosis is judged sufficiently probable or improbable to make a management decision. The probability that is 'sufficient' to reach a 'diagnosis' varies with the seriousness, treatability or novelty of the diagnosis being considered. For example, a patient with a 1% chance of meningitis may get immediate antibiotics in primary care while a patient with definitive otitis media may not (less serious condition).[3] When probabilistic reasoning is 'dissected' in this manner, it may sound difficult, cumbersome and lengthy. In practice, it is often undertaken quickly and intuitively.

When asked if they use pre-test probabilities in their clinical work, most clinicians will answer 'no'. This is true to the extent that they do not recognise and use them explicitly, but most clinicians use them implicitly. For example, when seeing a child aged three with a cough, most clinicians would not include emphysema in the differential diagnosis list as it is most unlikely (much less than 1%) in that age group. Another way of expressing that is to say the pre-test probability (prevalence) of emphysema in a three-year-old is very low.

A common question about pre-test probabilities is—where do they come from? Clinicians most commonly rely on their experience (never seeing a three-year-old with emphysema) and knowledge of the population and environment in which they work. Clinicians also check their judgments with colleagues (not necessarily using the word 'probability' but perhaps saying 'emphysema unlikely'), and probabilities can be found in the literature (also known as prevalences). Importantly, an approximate estimate of prevalence is often adequate to guide the use and interpretation of appropriate questions, examinations and diagnostic tests to increase or decrease the estimated probability of the condition. For example, a 50-year-old man presenting with crushing central chest pain would warrant a cluster of questions around the cardiovascular system and, if that was unfruitful, then on other adjacent organs.

Figure 1. Stages and strategies in arriving at a diagnosis



Why do we need numbers? Can't we stick with adjectives?

A study conducted in 1980 asked 16 physicians to ascribe a probability to words commonly used on radiology and laboratory reports.[4] For terms such as 'moderate risk' the range was 20–75%, for 'pathognomonic' (absolutely certain) it was 55–100% and for 'high probability' it was 55–95%. The authors concluded that physicians should use numerical probabilities in communication.

Box 1. Terminology and definitions

Pre-test probability = prevalence

Post-test probability of a positive test = positive predictive value

Post-test probability of a negative test = 1 – negative predictive value

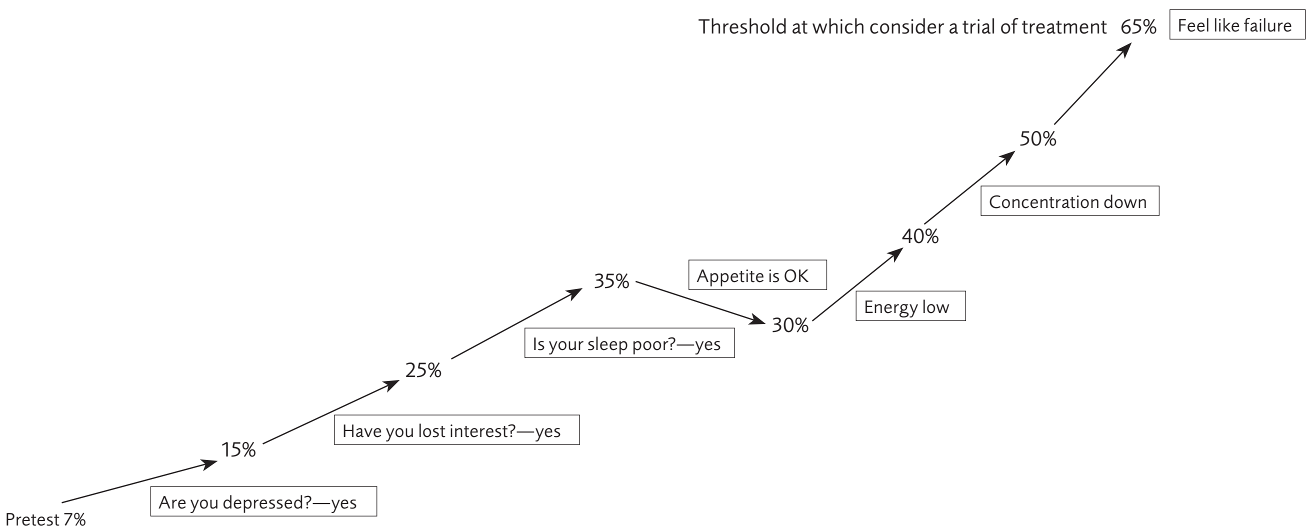
Likelihood ratio (LR) = $\frac{\text{probability of test result in patients with disease}}{\text{probability of test result in patients without disease}}$

Likelihood ratio positive sensitivity/(1-specificity) (raises probabilities)

Likelihood ratio negative (1-sensitivity)/specificity (lowers probabilities)

It is also possible to create multilevel likelihood ratios, but that is beyond the scope of this article.

Figure 2. Schematic view of asking sequential questions to increase or decrease probability of depression and at 'some' point exceed the clinician's threshold for treatment



Using pre-test probability, history and examination to diagnose Eustachian tube dysfunction

A 54-year-old lawyer phoned her GP because she still had bilateral ear pain one week after being seen by a doctor in an emergency clinic. She had been afebrile at the time of diagnosis and recently had had a viral upper respiratory tract infection. The emergency clinic doctor who saw her had said that he could not get a good view of her eardrums, but on the basis of her pain assumed she had bilateral otitis media. He prescribed oral amoxicillin for one week. On the phone, her GP said that it was unlikely that she had had otitis media, knowing that in adults otitis media is very rare, that the most likely cause of her pain was Eustachian tube dysfunction and that it would resolve over the next one to two weeks (which it did). After the phone call she came into the office where examination confirmed that both drums were clear and she was unable to insufflate the eardrums with a Valsalva manoeuvre—further confirming the GP diagnosis.

It is likely that the emergency room doctor had seen ear pain only in the context of acute otitis media, probably in children and probably in secondary care. He may never have seen Eustachian tube dysfunction, which is usually seen in primary care. The learning point for this case is that clinicians need to know the epidemiology of signs and symptoms in the clinical context in which they are making their diagnosis.

Using probabilistic reasoning through history taking to diagnose major depression

About 7% of adults attending primary care who are not already on antidepressants will have depression, i.e. the pre-test probability for such people is 7%.[5] If the patient presents with a history of, for example, fatigue or sleep difficulty, the clinician may decide to start asking about depression. If the patient answers yes to one of the first two questions of the DSM-IV, the pre-test probability rises from 7% to 18% (the post-test probability after the two questions becomes the new pre-test probability before the next set of questions).[5] Asking subsequent questions from the DSM-IV increases (when answer is positive) or decreases the probability (when answer is negative) until the threshold of five positive questions is reached and/or the clinician judges that there is sufficient probability of a diagnosis to take the next step in management—either starting therapy or opting for watchful waiting, which uses time as a further diagnostic test.[1]

This is explained in Figure 2. The patient starts off at 7% and answering yes to feeling depressed and loss of interest and pleasure in all or most activities will raise the probability of depression to 15%.[3] A positive answer to having problems with sleep will raise the probability to 35%. If the appetite is normal (a negative response in this context) this decreases the probability of depression to say 30%. The next question about lethargy is positive and increases the probability to 40%. This will continue until all nine DSM-IV questions have been asked and the clinician will make a diagnosis and/or initiate a management option. On the schematic Figure 2 it is assumed that a clinician may wish to initiate some form of treatment at 65% but could be higher or lower depending on their and the patient's preference.

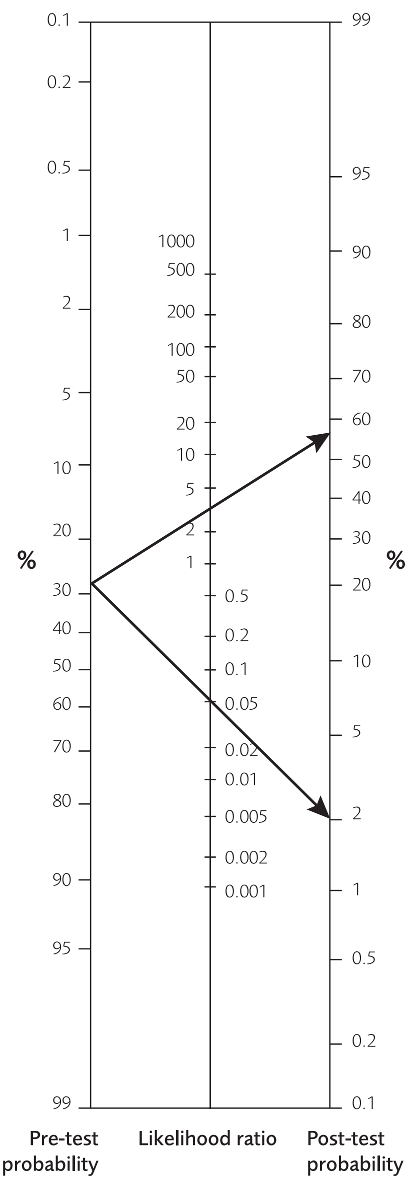
Using a likelihood ratio and nomograms to obtain post-test probabilities for streptococcal sore throat

For a child presenting with a sore throat it is important to know if it is Group A streptococcal (GAS) pharyngitis/tonsillitis or another cause (usually viral). Treatment of 'streptococcal sore throat' with an antibiotic in a developed country may shorten symptoms by about 16 hours. However, it may also prevent rheumatic fever in communities with a high prevalence of this disease.[6]

For a 10-year-old patient in primary care we can use the table from the paper by McIsaac et al. (1998).[7] Table 1 shows the scoring system that involves taking a history and undertaking a physical examination of the head and neck. Each positive item (e.g. a high temperature) increases the probability of streptococcal pharyngitis or tonsillitis. An example is shown in Table 2, given a prevalence of GAS pharyngitis among people with current 'sore throat' according to their scores.

The key issue in probabilistic reasoning is how to get from the pre-test probability to the post-test probability. One tool for this is the likelihood ratio and can be visually simple when presented inside a nomogram (Figure 3). The left-hand vertical on the nomogram is the pre-test probability, in the middle is the likelihood ratio and

Figure 3. Fagan nomogram depicting pre-test probability and post-test probability of Group A streptococcal pharyngitis or tonsillitis following a positive or negative throat swab in a medium-prevalence setting (28%).



the right-hand vertical is the post-test probability. The likelihood ratio is a number which, when greater than one, can increase the probability of a condition (more likely to rule in) and when less than one decrease the probability of a condition (more likely to rule out). In our sore throat example the likelihood ratio of a positive test is 2.9, i.e. if the throat swab comes back positive, there

Table 1. McIsaac score for streptococcal sore throat[7]

Criteria	Points
Temp >38°C	1
No cough	1
Tender anterior cervical adenopathy	1
Tonsil swelling or exudates	1
Age 3–14 years	1
Age 15–44 years	0
Age ≥45 years	-1
Total score	

Table 2. Post-test probabilities of streptococcal infection as cause of sore throat after applying McIsaac score, presuming a pre-test probability (usual community prevalence) of 2–3%[7]

Total score	Post-test probability of streptococcal infection	Suggested management
0	2–3%	No culture or antibiotic
1	4–6%	
2	10–12%	Culture all, treat if +ve
3	27–28%	
4	38–63%	Start treatment

Likelihood ratio of a positive throat swab 2.9
Likelihood ratio of a negative throat swab 0.04

is roughly a 2.9 increase in the chance of the patient having streptococcal tonsillitis (it could still be a false positive but we will never be sure on this). If the patient had a score of 3 on the McIsaac Table 1, the pre-test probability would be 28% so a positive test would increase that to 53% (the post test probability) thereby making the use of antibiotics more convincing. If the swab came back negative the likelihood ratio would be 0.04.

Using the nomogram (Figure 3) gives a post-test probability of 2% and supports the decision not to give antibiotics unless the patient was at extremely high risk of rheumatic fever. Using a likelihood ratio with a nomogram enables a simple visual presentation of how the post-test probability of one test can become the pre-test for the next test (question, physical exam or laboratory investigation). The results and clinical interpretation for low, medium and high probability for streptococcal pharyngitis are summarised in Table 3.

How do you get a likelihood ratio?

There are three ways of getting likelihood ratios. They can be calculated using the formulae in Box 1 (for simple 2x2 tables); they can be obtained from the literature; and, finally, it is possible to

Table 3. Post-test probabilities of sore throat being due to Group A *Streptococcus* following a throat swab

	Low prevalence	Medium prevalence	High prevalence
Pre-test probability of Streptococcus	2%	28%	63%
Post-test probability after a positive test	6%	53%	83%
Post-test probability after a negative test	0.2%	2%	7%
Comment	A negative swab result in this setting confidently rules out streptococcal tonsillitis. Most positive results will be false positives.	The test works well in this setting. It is reasonable to assume *Streptococcus* is present if the swab is positive and not *Streptococcus* if the swab is negative*	A positive swab result in this setting confidently rules in streptococcal tonsillitis. However, an increasing proportion of negative results will be false negatives, which is risky—when there is risk of rheumatic fever following untreated streptococcal infection.

*The gold standard for streptococcal infections is repeated serum antibodies which are not available soon enough to help clinically

Table 4. Likelihood ratios and their approximate post test probabilities (modified by McGee 2002)[8]

Likelihood ratios	Approximate changes in post-test probabilities
Values between 0 and 1 decrease the probability of disease	
0.1	-45%
0.2	-30%
0.3	-25%
0.4	-20%
0.5	-15%
Likelihood ratio = 1	0%
Values greater than 1 increase the probability of disease	
2	+15%
3	+20%
4	+25%
5	+30%
6	+35%
8	+40%
10	+45%

calculate multilevel 3x2 or higher—however, this latter category is beyond the scope of this paper.

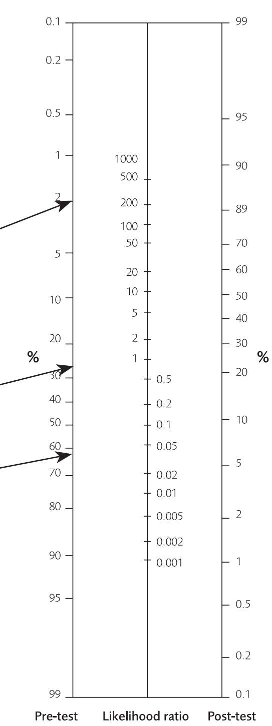
The mathematics behind likelihood ratios requires manipulations that are beyond the daily interest of most clinicians. However, the principles of interpreting tests results in low, medium and high prevalence settings are important to understand. The actual increase or decrease in probability will depend on the likelihood ratios. McGee has estimated the approximate changes in probability with different likelihood ratios (Table 4), which is another way of thinking about likelihood ratios.[8]

Clinical use of tests depends on pre-test probabilities

The fact that clinical interpretation of the same test for GAS varies markedly with the prevalence, as summarised in Table 3, leads to the

Figure 4. Combining the prevalence table with the nomogram

	Low prevalence	Medium prevalence	High prevalence
Pre-test of Streptococcus	2%	28%	63%
Issue	False +ve	Tests work well	False –ve
Strategy	**+ve test:** do another test, or gold standard test* **-ve test:** don't treat	**+ve test:** treat **-ve test:** don't treat	**+ve test:** treat **-ve test:** perform gold standard test*
What test can do	rule out not rule in	rule out rule in	rule in not rule out



*need to consider the risk/cost benefit

Box 2. Key learning points

1. Probabilistic reasoning requires knowing the degree to which a positive or negative result of a test adjusts the probability of a given disease.
2. It is important to know the pre-test probabilities (epidemiology/prevalence) of the symptoms and signs of patients in the setting in which you are working.
3. There are three ways you can get a 'feel' for the pre-test probability:
   i. By experience
   ii. By looking at the literature
   iii. By discussing with a colleague.
4. The purpose of taking a history and doing a physical examination is to take a presenting pre-test probability and increase it to a probability where standard tests will rule in or rule out a diagnosis, or to a probability where expensive and/or invasive and/or time-consuming investigations can be justified.
5. Each question and physical examination is a diagnostic test in itself with a sensitivity and specificity.
6. Standard everyday tests (e.g. x-ray, blood or urine) work best in medium prevalence settings in that they both reasonably rule in and rule out conditions.
7. False positives are the problem in low-prevalence settings.
8. False negatives are the problem in high-prevalence settings.
9. Numbers may facilitate clinician communication more accurately than words.

important general pattern shown in Figure 4. In conditions of low prevalence, most positive tests are false positives. In conditions of high prevalence, a significant proportion of negative tests are false negatives. Only in conditions of medium prevalence, from 10% to 50% at the upper end, do tests perform as we (and the public) would intuitively expect. In this range it may be reasonable to regard a positive test as a true positive and a negative test as a true negative, although it also depends on the degree of uncertainty a clinician is likely to accept given the seriousness of the condition and the drawbacks of using a treatment when it is not needed (e.g. side effects and resistance development from antibiotic use). It is instructive to note that our routine primary care screening tests perform relatively poorly (e.g. cervical smears, mammography, colorectal screening, and prostate screening) and this raises the issues of false positives.

Why do we take a history, perform examinations and do investigations on patients?

The goal of taking a history and doing a physical examination is to increase the pre-test probability to a medium probability where routine testing can be undertaken (e.g. x-rays or blood tests) or to a high probability point where empirical treatment can be initiated (e.g. migraine prophylaxis) or else expensive (e.g. magnetic resonance imaging), time-consuming or invasive (e.g. colonoscopy) investigations can be justified. If the answers to the clinician's questions are mainly/all negative then ruling out that condition may be reasonable given appropriate safety netting. Each question, physical examination and investigation is a diagnostic test in itself which can raise or lower probability of the target condition accordingly.

Therefore, the diagnostic process using probabilistic reasoning involves the clinician choosing and interpreting questions, examinations and investigations according to the pre-test probability (prevalence) and validity of the test (sensitivity, specificity or likelihood ratios), and making a treatment decision on the basis of the post-test probability, severity of disease and consequence of treating when it is not needed.

Case: The false negative error in congestive heart failure (CHF)—a worked example

Probabilistic reasoning may have saved an admission (using Table 4 rather than the nomogram). An 85-year-old woman presented with shortness of breath which had come on over the past few days, but she had no chest pain. She had a past history of hypertension, was overweight (BMI 30) and had never smoked. She was afebrile, had slightly swollen ankles, a JVP that was not measurable and a clear chest. Her pre-test probability was 'guessed' at 50%.[9] A chest x-ray (CXR) was arranged that day, which showed cardiomegaly but reported no evidence of left ventricular failure. A CXR without interstitial or alveolar oedema has a likelihood ratio negative of 0.61 and 0.72. These likelihood ratios would only reduce the probability of CHF to about 40%, which is still a pre-test probability warranting attention. From the probabilistic reasoning point of view the cardiomegaly would increase the post-test probability by 25% (likelihood ratio 4.0). At this point the probability of CHF was 65%. However, the clinician at the time believed the CXR (i.e. patient did not have CHF) and sent the patient home, which turned out to be a mistake. A BNP was ordered but in that setting took over one week to come back to the clinician. One week later the patient was admitted to the local hospital with a CXR reporting left ventricular failure. What should the clinician have done faced with a probability of CHF of 65%? The clearest action would have been to give some intravenous frusemide; she almost certainly would have lost a few kilograms of weight (thereby reinforcing the CHF diagnosis by increasing her post-test probability and making her feel better) and potentially it would have kept her out of hospital. A subsequent echocardiogram confirmed CHF with preserved left ventricular ejection fraction (formerly diastolic dysfunction). The CXRs were reviewed by the radiologists who found the first one to be under-penetrated thereby missing the signs of CHF. Of course, there are many reasons for errors in reporting investigations, e.g. quality of production through to errors in interpretation. Our clinician could have learned that CXRs are not good at ruling out CHF but are better at ruling in, or been told that by someone with a lot of experience, or else he could have worked it out as shown in this example.

Pulling it all together

The purpose of this paper has been to demonstrate the role of pre-test probabilities in clinical decision-making. They are the starting point for our clinical questions and physical examination, increasing and decreasing the probability of the target question. The aim is to get to a probability where we can start using standard tests, or to high probabilities where we can start using invasive, time-consuming and expensive tests, and/or start treatment empirically. There are tools that can help us with that. The likelihood ratio combined with the Fagan nomogram can give us a good visual presentation of what is going on. Alternatively we can use the approximations of McGee to increase or decrease probabilities. Finally, Figure 4 shows how the pre-test probabilities link with the Fagan nomogram/likelihood ratios to increase and decrease the post-test probabilities.

References

1. Heneghan C, Glasziou P, Thompson M, et al. Diagnosis in general practice. Diagnostic strategies used in primary care. BMJ 2009;338:1003–6.
2. Elstein AS, Kagan N, Shulman LS, Jason H, Loupe MJ. Methods and theory in the study of medical inquiry. J Med Educ 1972;47:83–92.
3. Little P, Gould C, Williamson I, Moore M, Warner G, Dunleavey J. Pragmatic randomised controlled trial of two prescribing strategies for childhood acute otitis media. BMJ 2001;322:336–42.
4. Bryant GD, Norman GR. Expressions of probability: words and numbers. N Engl J Med 1980;302(7):411.
5. Arroll B, Khin N, Kerse N. Two verbally asked questions are simple and valid. BMJ 2003;327:1144–6.
6. Del Mar CB, Glasziou PP, Spinks AB. Antibiotics for sore throat. Cochrane Database Syst Rev (4):CD000023.
7. McIssac WJ, White D, Tannebaum D, Low DE. A clinical score to reduce unnecessary antibiotic use in patients with sore throat. Can Med Assoc J 1998;158:75–83.
8. McGee S. Simplifying likelihood ratios. J Gen Intern Med 2002;17:647–50.
9. Polmear A. Evidence-based diagnosis in primary care. Philadelphia: Elseveir Ltd; 2008.

COMPETING INTERESTS
None declared.