Review Article

# The Evidence Based Medicine Approach to Diagnostic Testing: practicalities and limitations

**Robert C Hawkins**

Department of Pathology and Laboratory Medicine, Tan Tock Seng Hospital, Singapore 308433
For correspondence: Dr Robert Hawkins e-mail: Robert_Hawkins@ttsh.com.sg

**Abstract**

Evidence-Based Medicine (EBM) has become a popular approach to medical decision making and is increasingly part of undergraduate and postgraduate medical education. EBM follows four steps: 1. formulate a clear clinical question from a patient's problem; 2. search the literature for relevant clinical articles; 3. evaluate (critically appraise) the evidence for its validity and usefulness; 4. implement useful findings into clinical practice. This review describes the concepts, terminology and skills taught to attendees at EBM courses, focusing specifically on the approach taken to diagnostic questions. It covers how to ask an answerable clinical question, search for evidence, construct diagnostic critically appraised topics (CATs), and use sensitivity, specificity, likelihood ratios, kappa and phi statistics. It familiarises readers with the lexicon and techniques of EBM and allows better understanding of the needs of EBM practitioners.

## Introduction

Over the last decade, the term EBM has become increasingly popular in the medical literature.[1,2] There are almost 16,000 papers on EBM available through the PubMed portal of the National Library of Medicine up to the end of 2004. EBM has become well established as a component of both undergraduate and postgraduate medical education. Despite much enthusiasm for EBM, there has been some negative reaction.[3-5] Some nursing authors question the philosophical basis and applicability of EBM concepts to their discipline[6-9] while physician authors argue that it denigrates clinical expertise, promotes a cookbook approach to medicine, is a cost-cutting tool and leads to therapeutic nihilism in the absence of evidence from randomised controlled trials.[10] EBM courses are a popular medical education activity throughout the world and the users of the laboratory service are increasingly familiar with the terms, concepts and approaches of EBM.

Several excellent reviews have been published describing the importance of evidence based laboratory medicine.[11-15] This review instead examines the practical skills and concepts of EBM taught to clinicians attending EBM courses, focusing specifically on the approach to diagnostic tests. It should allow laboratory staff to better understand what information EBM-influenced clinicians seek and how they use it.

## What is EBM and What do They Teach at an EBM Course?

EBM has been defined as "the conscientious, explicit and judicious use of current best evidence in making decisions about the care of patients".[1] Although the term was first used in the 1980s at McMaster Medical School in Canada,[16] the philosophical basis of EBM has been suggested to stretch back much further to 18th century Europe or even to ancient China.[11] EBM asks questions, finds and appraises the relevant data, and harnesses that information for everyday clinical practice. EBM follows four steps: formulate a clear clinical question from a patient's problem; search the literature for relevant clinical articles; evaluate (critically appraise) the evidence for its validity and usefulness; implement useful findings in clinical practice.[17]

Courses on EBM are generally composed of three basic elements: teaching participants how to formulate answerable clinical questions, how to find the best current evidence and how to critically appraise the evidence. This last segment is often divided into different approaches for diagnostic, prognostic, therapy and harm questions. This review will focus on the approach to diagnostic questions.

## Formulating an Answerable Cinical Question

EBM courses are doctrinaire in their approach and require participants to follow a set method to EBM problems.

Questions are divided into two types: general medical questions (e.g. What causes hyperbilirubinaema?, When does cardiac rupture usually occur after acute myocardial infarction?) and specific patient-based questions (e.g. What is the expected benefit to a 60y old microalbuminuric normotensive diabetic of commencing angiotensin converting enzyme inhibitor treatment?) General questions are called "background" questions while specific questions are called "foreground" questions.[2]

EBM deals with extremely specific patient-based questions rather than general questions, and these specific questions should arise from a specific interaction with a specific patient in the clinic or ward. This very "patient-centred" approach can make it difficult for staff with no direct patient contact to participate in EBM courses or use the skills acquired. The four components of an answerable question can be described by the acronym PICO (patient, intervention, comparison, outcome). For diagnostic questions, P represents the patient/population, I represents the investigation, C represents the comparison investigation or gold standard while O represents the outcome of interest. For example, P = men over 60 y with symptoms of prostatism, I = PSA measurement, C = rectal examination, O = correct diagnosis of prostate cancer.

**Finding the Best Evidence**

Once the clinical question has been formulated in an answerable manner, the search for relevant evidence begins. EBM courses utilise the latest in on-line medical literature search strategies. As all participants may not have access to the commercial products such as Up-to-Date or Ovid, public access internet portals such as Pubmed (www.pubmed.com) are often used. Pubmed is a service of the National Library of Medicine (NLM) in the United States and allows access to over 15 million citations for biomedical articles back to the 1950s. It is free of charge and allows users to print or store the results of searches or alternatively download them to reference managers. The abstracts of the articles are readily available and links to the full text of articles are increasingly included.

Rather than typing in words or phrases of interest into the Pubmed search fields, EBM courses encourage participants to use a more rigorous approach to searching using the MeSH (Medical Subject Headings) subheading system. MeSH is NLM's controlled vocabulary used for indexing articles for MEDLINE/PubMed. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts. Under the MeSH database tab on the Pubmed website are a number of useful tutorials on use of MeSH. Particularly valuable are the subheadings such as "blood", "cerebrospinal fluid" and "diagnostic use", allowing for more specific searches. Different MeSH headings and subheadings can be combined using Boolean operators (such as AND and OR) to allow very specific searches.

Another search strategy with Pubmed is the "Clinical Query". This is now very EBM-oriented and allows one to search for clinical studies, choosing from aetiology, diagnosis, therapy and prognosis and either a narrow specific search or a wide sensitive search. It should be noted that Pubmed frequently changes details of its search utilities in this area.

"Limits" are another useful option, allowing the search to choose article language, type (e.g. review) and publication date range, as well as details of study participants (age, gender).

**Critically Appraising the Evidence for a Diagnostic Question**

After searching the literature for the best evidence, the resulting literature is "critically appraised" to assess its validity. Assessment of validity involves four questions[2]:

1. Was there an independent, blind comparison with a reference standard of diagnosis? A clear definition of the requirements for the reference standard of diagnosis is important, ensuring that it does not rely on data from the test undergoing evaluation. For example, some troponin studies assess troponin results in predicting final discharge diagnosis/management/outcome, forgetting that if the troponin result was revealed to the clinician during admission, it will have inevitably affected diagnosis, management and outcome.
2. Was the diagnostic test evaluated in an appropriate population of patients comparable to the patient of interest? It should be clear if the patient group assessed was similar in age, sex, race, location (inpatient, outpatient, emergency department, primary care) and presenting symptoms to the patient of interest. Many studies of diagnostic tests use clearly healthy vs. clearly diseased patients for their assessments, giving over-optimistic assessments of the performance characteristics (including sensitivity, specificity etc.) of the test as no "grey zone" patients are included.[18] This "spectrum bias" overestimates the performance of the test[19] and is the cause of many of the differences in reported sensitivity and specificity figures found in the literature. One should choose performance characteristics from literature that describe the use of the test in a clinical setting comparable to that of the patient of interest.[20]
3. Was the reference standard applied regardless of the diagnostic test result? If the reference standard is expensive (e.g. hepatitis serology confirmation) or

invasive (e.g. bone marrow biopsy to assess iron stores), all patients may not undergo the reference procedure. This introduces a "verification bias" into the assessment, with classification of false negatives as true negatives, and can distort the resulting performance characteristics of the test.

4. Was the test validated in a second, independent group of patients? Have the initial performance parameters (sensitivity, specificity etc.) been confirmed by application to a second independent group of patients? This is particularly important for new tests where initial evaluations tend to use clear-cut healthy and diseased groups, resulting in performance parameters that may not be seen in clinical practice (as discussed above).

If a report fails one or more of these criteria, it may have fatal flaws that limit its usefulness to the EBM practitioner. However it may be difficult to find literature that fulfils all the requirements, especially use of a comparable patient population (question 2 above), so one must use what is available but be aware of its shortcomings. Of particular help in assessing diagnostic studies is the Standards for Reporting Diagnostic Accuracy (STARD) statement. This provides a checklist of 25 items and a flow diagram designed to improve the methodological and informational quality of studies of diagnostic accuracy.[21,22] The STARD statement was published in several leading journals in early 2003, including *Clinical Chemistry, Annals of Internal Medicine* and *BMJ*, and there is some evidence that the quality of published diagnostic studies has improved subsequently.[23]

**Sensitivity, Specificity and Likelihood Ratios**

Clinical sensitivity and specificity are not to be confused with the analytical sensitivity and specificity concepts familiar to laboratory staff. This approach is known as Bayesian analysis and involves calculation of parameters from a 2x2 matrix (see Table 1) that describe the ability of a test to identify a condition or disease. A useful phrase to remember the difference between them is "Sensitivity is positivity in disease, specificity is negativity in health". High values for sensitivity and specificity can be very useful in allowing one to "rule in" or "rule out" disease using the mnemonics "Snout" (Sensitivity – rule out") and "Spin" (Specificity – rule in). This means that 100% sensitivity corresponds to 100% negative predictive value (i.e. rule out) and conversely 100% specificity corresponds to 100% positive predictive value (i.e. rule in). This may appear counter-intuitive but is best understood by considering Figure 1. All cases above the 100% specificity cut-off have the disease while all cases below the 100% sensitivity cut-off do not. In situations where the sensitivity or specificity is not 100%, the positive and negative predictive values depend on the disease prevalence (= pre-test probability of disease) and must be calculated individually.

EBM goes further than the older concepts of sensitivity and specificity to the newer ideas of likelihood ratios and pre-test-odds and post-test odds. The mathematics behind these concepts are harder to grasp than sensitivity and specificity but their value is in allowing clinicians to apply the same sensitivity/specificity data to different patients with different pre-test probabilities of disease, arriving at different post-test

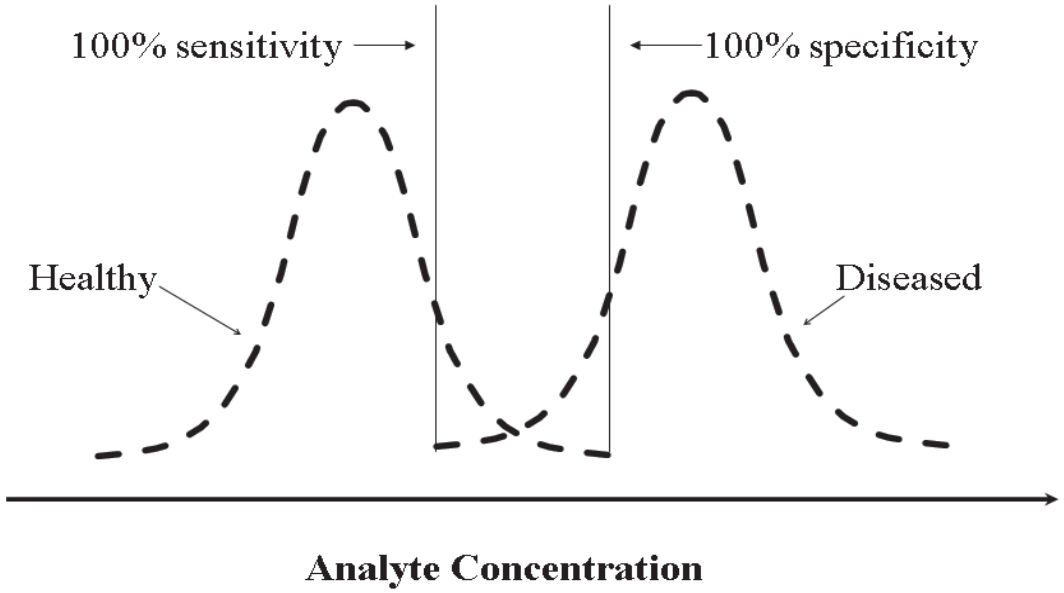


**Figure 1**. "**Sp***in*" (a positive result for a 100% **sp**ecific test rules *in* disease) and "**Sn***out*" (a negative result for a 100% **sen**sitive test rules *out* disease) rules.

**Table 1.** Test performance parameters (adapted from reference 24).

| | | Disease | |
|---|---|---|---|
| | | *Positive* | *Negative* |
| **Test Result** | ***Positive*** | True Positive (TP) | False Positive (FP) |
| | ***Negative*** | False Negative (FN) | True Negative (TN) |

Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

Positive Predictive Value (PPV) = TP / (TP + FP)

Negative Predictive Value (NPV) = TN / (TN + FN)

Likelihood Ratio for Positive Test (LR+) = (TP / (TP +FN)) / (FP /(FP + TN))

Likelihood Ratio for Negative Test (LR-) = (FN / (TP + FN)) / (TN / (FP + TN))

Diagnostic Accuracy = (TP + TN) / (TP+TN+FP+FN)

Pre-test probability (prevalence) = (TP + FN) / (TP+FP+TN+FN)

Pre-test Odds = prevalence / (1-prevalence) = (TP + FN) / (FP + TN)

Post-test Odds = pre-test odds x likelihood ratio

Post-test Probability = post-test odds / (1+ post-test odds)

probabilities. For example, for a test with a given sensitivity of 95% and specificity of 98%, if the patient's pre-test probability of disease is 12%, a positive result gives a post-test probability of disease of 86.6% and a negative result gives a post-test probability of disease of 0.7%. For the same test, if the pre-test probability is 50%, the respective values are 97.9% and 4.9%. Generally the pre-test probability is the same as the disease prevalence in the population of interest; however clinicians may choose to customise it for the individual patient. Readers should appreciate the mathematical difference between odds and probability and note that a final step to convert from odds to probability is needed (see Table 1). These mathematical manipulations are difficult to perform manually and so in practice this approach is best suited to computerised applications or printed nomograms allowing clinicians to read off the post-test probability for a given pre-test probability and likelihood ratio.[2,25]

Likelihood ratios can be useful in assessing the potential utility of a test. Likelihood ratios of >10 or <0.1 can generate large changes in post-test disease probability while conversely likelihood ratios of 0.5-2 have little effect.[24] Likelihood ratios can also be used when considering a sequence of independent tests (e.g. ECG findings followed by TnI testing) since likelihood ratios can be multiplied in series. However many laboratory tests are not independent of each other (e.g. AST/ALT, urea/creatinine) and one should bear this in mind when using this approach.

A further extension of this approach is the use of multilevel likelihood ratios. Rather than choosing a single cut-off for a test (results are either positive or negative relative to the cut-off), a series of different bands of results are considered, each with its own likelihood ratio. If a patient result falls in the band, the appropriate likelihood ratio is chosen for the calculation. This approach makes use of the quantitative nature of many laboratory results rather than reducing them to qualitative "positive/negatives" with the use of a single cut-off. This is intuitively better as clearly the interpretation of a result only just above the cut-off should be different from one many magnitudes above the cut-off. Unfortunately this approach is less easy to computerise and generally requires manual calculation steps.

Examples of tests used to demonstrate multilevel likelihood ratios include serum ferritin to identify iron deficiency[26,27] and free PSA to identity prostate cancer.[28] Although many papers offer only sensitivity and specificity summaries in the abstracts and text, the data in the full text of the paper may allow construction of multilevel likelihood ratios. Laboratories can expect clinicians to increasingly ask that likelihood data be available for their patients, either on the

reports or via a computerised solution. Such an approach is only possible for tests where there is a specific disease-test link and is not applicable to laboratory tests such as electrolyte measurement. However, for the minority of tests with such associations, reporting 95% reference intervals or a single cut-off is no longer satisfactory and laboratories need to develop innovative solutions to meet this new reporting demand.

Customising the pre-test probability (or prevalence) of the disease to the particular patient is not only important in allowing one to calculate the post-test probability but in deciding whether to perform the test at all. The aim of the diagnostic process is not to seek 100% certainty but rather to reduce the level of uncertainty to a level to allow optimal therapeutic decisions.[29] Figure 2 shows a spectrum of pre-test probabilities for a given disorder, together with two treatment decision thresholds. The position of each threshold depends on factors such as the characteristics of the disease and the cost/efficacy/toxicity/availability of treatment. The threshold for initiating chemotherapy for malignancy is much higher than for giving aspirin for the common cold. Testing should only be performed when the result of the test (either positive or negative) will cause the post-test probability to cross one of the thresholds. Such formal thresholds are available for few, if any, diseases but this theoretical approach illustrates the practical maxim: "Only request a laboratory test if the result will change the management of the patient". One can see from the cartoon that testing is most useful in patients with intermediate pre-test probabilities and is of little benefit when the pre-test probability is very high or very low.

**Diagnostic CATs**

The concept of the CAT was developed by internal medicine fellows at McMaster University[30] as a vehicle to simplify the task of making the results of EBM available for patient care, teaching and learning.[31] It is an instrument by which the clinician can maintain and retrieve relevant evidence quickly and easily for application in patient care. In form, it is a 1-2 page summary that condenses the process by which a well formulated question leads to a literature search, the choice of a study, and critical appraisal of its validity, results and applicability using published criteria.[24] The exact form and content of a given CAT depends on its practice context but all share the same basic structure of a title, search, summary and appraisal.[31] Different varieties of CAT include therapy, diagnosis, harm and prognosis CATs. An example of a diagnostic CAT for a non-laboratory test is shown in the Appendix to illustrate the generic approach (details of the evidence search strategy have been removed for brevity).

The CAT process is patient-focused and both starts and finishes with a patient interaction. A specific clinical scenario that evokes a clinical question is described. The description should include sufficient demographic and patient history details to allow potential customisation of the literature results to the specific patient who is the focus of the CAT. These details should include patient preferences, comments or any clues as the patient's concerns and desire for information. Since the final step in the CAT process will be responding to the patient's concern or question, it is essential to understand this at the outset so that the appropriate question can be framed.

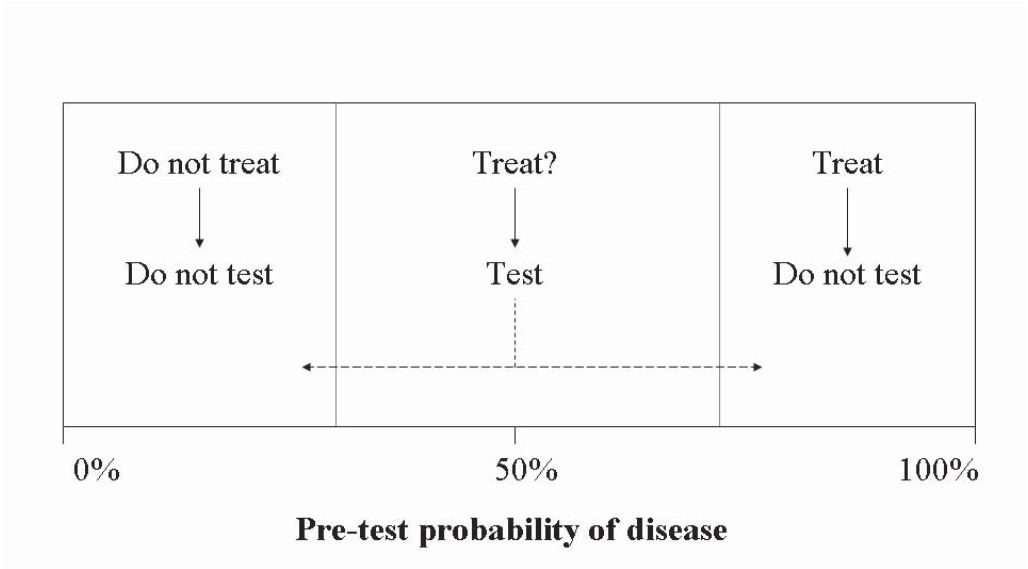


**Figure 2.** Test-treatment thresholds. Testing is most useful in cases with intermediate probability of disease (Adapted from reference 2).

Participants at EBM courses are encouraged to follow the CAT model when approaching an EBM problem. All CATs follow the same PICO framework and a variety of different templates are available for participants to use to aid in CAT development. The example shown here uses a shortened version of the template developed at Auckland University by Rod Jackson (www.epiq.co.nz) and is very useful in prompting the participant to answer a series of questions for each step. Another software product available includes CATmaker available from the Centre for Evidence-Based Medicine (CEBM) in Oxford (www.cebm.net/catmaker.asp). Such tools simplify the task of constructing a CAT but the basic elements can be easily performed without such aids.

Some EBM courses emphasize the importance of calculating the uncertainty associated with the various mathematical parameters within EBM. This is very familiar to laboratory staff and complements the recent interest in Measurement of Uncertainty in laboratory medicine.[32] The use of uncertainty calculations within EBM should familiarise clinicians with similar concepts in analytical reporting and pave the way for laboratories to introduce these concepts to users. The mathematical formulae for these calculations are not straightforward (see Table 2) and are best approached by use of computer spreadsheets. A variety of online and downloadable tools are available at websites including the Oxford CEBM site mentioned above and the CEBM at the University of Toronto website (www.cebm.utoronto.ca).

**Qualitative Comparisons**

Although much laboratory data is quantitative, the final interpretation is usually qualitative (e.g. disease present or absent, treat or do not treat, result is low or medium or high). It is thus useful to be familiar with statistical tools for comparison of categorical data. One of the most familiar of these is the kappa co-efficient, which is a measure of chance – corrected agreement between "observers" (= assays in the laboratory context) (see Table 3). This can be used for 2 or more categories and produces a result between 0 (poor agreement) and 1.0 (perfect agreement). There are a variety of approaches in assessing the kappa value, but one approach is to consider values of 0 = poor agreement, 0-0.2 = slight agreement, 0.2-0.4 = fair agreement, 0.4- 0.6 = moderate agreement, 0.6-0.8 = substantial agreement and 0.8-1.0 = almost perfect agreement.[24,33]

One disadvantage to kappa calculation is that it is affected by the proportion of cases in each category e.g. as the proportion of positive cases increases, kappa decreases despite no change in "observer" performance. An alternative is the phi statistic, which is chance-independent.[24] In a 2x2 cell comparison with 50% proportion of positive and negative cases, phi and kappa will be identical but will diverge with more extreme distributions. The interpretation of phi is similar to that of kappa.

**Practical Issues When Applying the EBM Approach to Diagnostic Questions**

Arguably the most important step in CAT construction is asking an answerable question and learning this is a basic EBM skill.[34] Describing the P component without being too exclusive can be hard. In the diagnosis CAT for Tinel's sign in the Appendix, choosing which of the many patient characteristics (Type 2 middle aged female Chinese diabetic subject complaining of hand weakness) to include in the search can be difficult. If the P component becomes too detailed, one risks not being able to find any appropriate studies. On the other hand, if it is too vague, too many studies will be found with no guarantee that they are applicable to the individual patient. Only with experience can one learn how specific to be at this stage. Even if not entered into the formal question, the patient-specific characteristics may be applied later when considering the subgroups or partitioning that has occurred within studies.

The I component is generally straightforward but for diagnostic CATs, clinicians are often unaware of the poor standardisation of diagnostic tests. Laboratory staff should remind clinicians that published assay performance data are not necessarily applicable to the locally available assay. Clinicians should have ready access to the local assay details such as manufacturer, instrument, methodology and assay details to minimise such problems.

Once a clinical query has been framed in the PICO manner, computerised searching of the literature is relatively straightforward. If the patient is a child, paediatric age limits may be of value. However in general, the use of various adult age and gender limits can be counterproductive and lead to the exclusion of relevant studies. Ethnicity is unfortunately not an available limit, creating problems when searching for articles on non-Caucasian populations. Although it is possible to use prefiltered information sources, such as the Cochrane library or *Clinical Evidence*, these options are costly, may not always be available to the EBM practitioner and remove the user from the nuts-and-bolts of the search. It is thus less easy to appreciate the success or failure of the question construction and potentially slows down the acquisition of question framing skills. The specific search strategy should be stated in the CAT to allow the reader to appreciate the inclusions and exclusions of the particular search and possible replication of the same search at a later date.

**Table 2.** Confidence Intervals (adapted from reference 24).

Using the general formulae below, customise for each parameter based on equations in Table 1.

| | Column 1 | Column 2 | Total |
|---|---|---|---|
| **Row 1** | a | b | n |
| **Row 2** | c | d | m |

*For sensitivity, specificity, PPV, NPV:*

If point estimate is a/n, the standard error is $\sqrt{a\,(n-a)\,/\,n^3}$

*For LR+ and LR-:*

If point estimate is (a/n) / (c/m), the standard error is $\sqrt{(1/a) - (1/n) + (1/c) - (1/m)}$

A particular problem for diagnostic CATs is the accurate assessment of pre-test probability. There is a general paucity of differential diagnosis studies in the literature, and those that are available lack the sophistication of combining multiple signs in the presence or absence of disease. For example, there are no data available in the literature on the prevalence of carpal tunnel syndrome (CTS) in diabetics complaining of hand weakness to use in the Appendix CAT. A guess must often be made which rather undermines the complicated mathematics with its attendant confidence intervals.

There are some limitations to the CAT process that should be recognised. It is dependent on the search strategy and the specific article chosen. This emphasises the importance of appropriate question framing and use of a good search strategy. Article quality is not necessarily the over-riding factor affecting the choice of article. Factors such as journal accessibility, abstract availability on line and the position of the article in the on-line search result listing can bias the choice of article. CATs lack the rigour of publishable systemic reviews and are statistically weaker than the methods used in formal meta-analyses.[31] The focus on a specific patient may also limit or slow the transferability of the results to another patient – calculations and even conclusions may need to be reworked for a different patient. There is also a continuing difficulty in applying the results of population studies to individual patients. Finally CATs can become stale and it is recommended that all should have an expiry date depending on the authors' sense of how rapidly evidence is working in the area. This is particularly relevant for diagnostic CATs given rapid changes in laboratory assay formulations and technical characteristics.

**Table 3.** Agreement statistics (adapted from reference 24).

| | | Method B | |
|---|---|---|---|
| | | Grade 1 | Grade 2 |
| **Method A** | Grade 1 | a | b |
| | Grade 2 | c | d |

Raw agreement = (a+d)/(a+b+c+d)

Kappa = (observed agreement – expected agreement)/(1 – expected agreement)

where observed agreement = (a+d) / (a+b+c+d)

and expected agreement = ((a+b)(a+c))/(a+b+c+d) + ((c+d)(b+d))/(a+b+c+d)

Odds Ratio (OR) = ad/bc

Phi = $((\sqrt{OR} - 1)/(\sqrt{OR} + 1)) + ((\sqrt{ab} - \sqrt{bc})/(\sqrt{ad} + \sqrt{bc}))$

The ability to perform critical appraisals, as exemplified by the construction of CATs, was initially considered to be essential for practitioners of EBM. However it has been acknowledged that "not all clinicians need to appraise evidence from scratch but all need some skills".[35] While many British general practitioners use evidence-based summaries generated by others (72%) and evidence-based practice guidelines or protocols (84%), the majority (95%) believe that "learning the skills of EBM is not the most appropriate method of moving …to EBM".[36] Other countries[35] and other groups of health care workers[37] share this lack of enthusiasm. Many clinicians lack the interest in learning the skills required for sophisticated appraisal of the original literature and those who do will often be short of time. EBM training in undergraduate studies is not necessarily successful in ensuring students can apply critical appraisal skills to management of an individual patient.[38] Even for postgraduate participants, standalone EBM teaching performs poorly compared to clinically integrated EBM teaching, improving knowledge but not skills, attitudes or behaviour.[39] Proponents of EBM now accept that the majority of clinicians will be "evidence users" who will use secondary sources of preappraised evidence to provide immediately applicable information[35] but argue that medical trainees should achieve the highest possible skill level in EBM. Most EBM courses thus continue to teach the elements of CAT construction to participants.

**Conclusions**

EBM is here to stay[40] and CAT construction continues to be a key component of EBM courses. The CAT is a useful tool for EBM practitioners assessing the primary literature as it results in a concise, retrievable and practical product. It is however unlikely to become the most popular source of evidence based medical information for the busy clinician. Instead, evidence-based clinical guidelines[41] and other secondary sources seem destined to be the preferred options. Nevertheless, the laboratory needs to be conversant with the terminology and concepts of EBM and CAT construction to better answer requests from clinicians. EBM promises a more scientific approach to medical decision making. It offers opportunities for laboratory staff to work with clinicians to improve use of laboratory resources and demonstrate the value of laboratory testing to patient health.[15]

Competing interests: none declared.

**References**


1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312:71-2.
2. Sackett DL, Straus SE, Richardson WS, Rosenberg WM, Haynes B. Evidence based medicine: how to practice and teach EBM. (2nd ed.) Churchill Livingstone, Toronto, 2000.
3. Correspondence. Evidence based medicine. Lancet 1995;346:1171-2.
4. Editorial. Evidence based medicine: in its place. Lancet 1995;346:785.
5. Graham-Smith D. Evidence based medicine: Socratic dissent. BMJ 1995;310:1126-7.
6. Clarke JB. Evidence-based practice: a retrograde step? The importance of pluralism in evidence generation for the practice of health care. J Clin Nurs 1999;8:89-94.
7. Rolfe G. Insufficient evidence: the problems of evidence-based nursing. Nurse Education Today 1999;19:433-42.
8. Regan JA. Will current clinical effectiveness initiatives encourage and facilitate practitioners to use evidence-based practice for the benefit of their clients? J Clin Nurs 1998;7:244-50.
9. White SJ. Evidence-based practice and nursing: the new panacea? Br J Nurs 1997;6:175-8.
10. Straus SE, McAlister FA. Evidence-based medicine: a commentary on common criticisms. CMAJ 2000;163:837-40.
11. McQueen MJ. Overview of evidence-based medicine: challenges for evidence-based laboratory medicine. Clin Chem 2001;47:1536-46.
12. McQueen MJ. Evidence-based medicine: its application to laboratory medicine. Ther Drug Monit 2000;22:1-9.
13. Price CP. Evidence-based laboratory medicine: supporting decision-making. Clin Chem 2000;46:1041-50.
14. Price CP. Application of the principles of evidence-based medicine to laboratory medicine. Clin Chim Acta 2003;333:147-54.
15. Panteghini M. The future of laboratory medicine: understanding the new pressures. Clin Biochem Rev 2004;25:207-15.
16. Evidence-based medicine. A new approach to teaching the practice of medicine. Evidence-Based Medicine Working Group. JAMA 1992;268:2420-5.
17. Rosenberg W, Donald A. Evidence based medicine: an approach to clinical problem-solving. BMJ 1995;310:1122-6.
18. Kazmierczak SC. Statistical techniques for evaluating the diagnostic utility of laboratory tests. Clin Chem Lab Med 1999;37:1001-9.
19. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.
20. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ 2002;324:669-71.

21. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. Clin Chem 2003;49:1-6.
22. McQueen M. Evidence-based laboratory medicine: addressing bias, generalisability and applicability in studies on diagnostic accuracy: the STARD initiative. Clin Biochem 2003;36:1-2.
23. Bossuyt PMM. The Quality of Reporting in Diagnostic Test Research: Getting Better, Still Not Optimal. Clin Chem 2004;50:465-6.
24. Guyatt G, Rennie D (eds). Users' guide to the medical literature: a manual for evidence-based clinical practice. AMA Press, Chicago, 2002.
25. Fagan TJ. Letter: Nomogram for Bayes theorem. N Engl J Med 1975;293:257.
26. Gordon LG, Lowry WS, Pedlow PJ, Patterson CC. Poor prognosis for malignant melanoma in Northern Ireland: a multivariate analysis. Br J Cancer 1991;63:283-6.
27. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C. Laboratory diagnosis of iron-deficiency anemia: an overview. J Gen Intern Med 1992;7:145-53.
28. Chen YT, Luderer AA, Thiel RP, Carlson G, Cuny CL, Soriano TF. Using proportions of free to total prostate-specific antigen, age, and total prostate-specific antigen to predict the probability of prostate cancer. Urology 1996;47:518-24.
29. Kassirer JP. Our stubborn quest for diagnostic certainty. A cause of excessive testing. N Engl J Med 1989;320:1489-91.
30. Sauve S, Lee HN, Meade MD, et al. The critically appraised topic: a practical approach to learning critical appraisal. Ann Roy Coll Phys Surg Canada 1995;28.
31. Wyer PC. The critically appraised topic: closing the evidence-transfer gap. Ann Emerg Med 1997;30:639-40.
32. White G, Farrance I. Uncertainty of measurement in quantitative medical testing: a laboratory implementation guide. Clin Biochem Rev 2004;25:S1-S24.
33. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126:161-9.
34. Richardson WS, Wilson MC, Nishikawa J, Hayward RSA. The well-built clinical question: a key to evidence-based decisions. ACP Journal Club 1995;123:A-12.
35. Guyatt GH, Meade MD, Jaeschke RZ, Cook DJ, Haynes B. Practitioners of evidence based care. BMJ 2000;320:955-6.
36. McColl A, Smith H, White P, Field J. General practitioners' perceptions of the route to evidence based medicine: a questionnaire survey. BMJ 1998;316:361-5.
37. White R, Taylor S. Nursing practice should be informed by the best available evidence, but should all first-level nurses be competent at research appraisal and utilization? Nurse Education Today 2002;22:220-4.
38. Bergus G, Vogelgesang S, Tansey J, Franklin E, Feld R. Appraising and applying evidence about a diagnostic test during a performance-based assessment. BMC Med Educ 2004;4:20.
39. Coomarasamy A, Khan KS. What is the evidence that postgraduate teaching in evidence based medicine changes anything? A systematic review. BMJ 2004;329:1017.
40. Del Mar C, Glasziou P, Mayer D. Teaching evidence based medicine. BMJ 2004;329:989-90.
41. Oosterhuis WP, Bruns DE, Watine J, Sandberg S, Horvath AR. Evidence-based guidelines in laboratory medicine: principles and methods. Clin Chem 2004;50:806-18.

**Appendix :** Example of Diagnostic CAT

**Title:** Tinel's Sign – Useless in the Diagnosis of CTS.

**Case Description:** A 45 year old Chinese woman was seen at Diabetic Clinic for a routine 3 monthly follow-up appointment. She has a 10 year history of Type 2 Diabetes with no complications to date. Her diabetes was well controlled on metformin and glipizide and her latest HbA1c was 6.8%. She mentioned that she has noticed some right hand weakness when holding objects over the last few months. On examination, there were no signs of peripheral neuropathy with normal vibration and light touch in both hands and feet. Hand strength was normal with no muscle wasting. Tinel's sign was negative bilaterally. I wondered whether the woman has carpal tunnel syndrome and how useful the negative Tinel's sign is in ruling this out.

**EBM Clinical Question:** In patients with diabetes, what is the negative predictive value of Tinel's sign in the diagnosis of CTS?

**Search Findings:**
Perkins BA, Olaleye D, Bril V. Carpal tunnel syndrome in patients with diabetic polyneuropathy. Diabetes Care 2002;25:565-9.

| **Section 1: Study Validity** | |
|---|---|
| *Evaluation criterion* | *How well was this criterion addressed?* |
| What were the key selection (inclusion & exclusion) criteria? Were they well defined? Were they replicable? | 180 consecutive patients referred for electrodiagnostic evaluation for suspected CTS with at least 1 symptom indicative of possible CTS (numbness /tingling /nocturnal aggravation / pain that can be "shaken out"/ pain after frequent wrist or hand use / dropping items / difficulty holding items / hand pain). Exclusions: generalized peripheral neuropathy, previous CTS surgery, cervical radiculopathy, other neuromuscular disorders. There were bilateral symptoms in 85, thus a total of 225 hands in the study. |
| Were selection criteria appropriate given study question? | The study did not specifically examine a diabetic population although there were diabetic patients in the study (number not specified in paper). A higher prevalence of CTS would be anticipated in this population (patients referred for testing re ?CTS) and the severity of disease would be higher than in the population of Type 2 diabetics with ?CTS that our patient comes from. |
| Did selection lead to an appropriate spectrum of participants (like those assessed in practice) | Yes but the number of diabetics is not stated and the study uses a pre-selected population as discussed above. |
| What was the reference standard of diagnosis? Was it clearly defined, independent & valid? | Nerve conduction studies (NCS) were used as the gold standard. The details of the testing and the electrodiagnostic criteria needed for CTS diagnosis were clearly described. |
| Was the reference standard applied regardless of test result? | Yes, all patients in the study underwent NCS. |
| Was the reference standard assessed blind to test result? | No. Reference testing was performed after the Tinel (and other) signs were tested and recorded. It appears that the same investigator performed both the Tinel's test and the NCS. The order in which the 6 different signs were performed is not stated and there was no blinding, meaning that the results of one test could well affect another. However objective criteria (as above) were used for the classification of the NCS results. |

| | |
|---|---|
| What tests were used? Were they well defined? Replicable? | 6 different CTS signs were used: Phalen sign, Hoffman-Tinel sign, hypesthesia, abductor pollicis brevis (APB) weakness, median nerve compression and square-shaped wrist. The tests are clearly described, including photographs of the less well-known signs. |
| Was the test applied regardless of the reference standard result? | Yes. |
| Was test assessment blind to reference standard result? | Yes, eliciting of signs occurred prior to NCS testing. |
| Was the test validated in a second, independent group? | No. |

**Section 2: Study Results: Accuracy & Precision**

| | |
|---|---|
| What measures of test accuracy were reported (sensitivity, specificity, LRs)? | Sensitivity, specificity, positive and negative predictive value. |
| What measures of precision were reported (CIs, p-values)? | None are given but it is possible to calculate them from the data given. |

| TEST | Sensitivity (95% CI) | Specificity (95% CI) | LR + / LR - | PV (+) / PV (-)* |
|---|---|---|---|---|
| Phalen | 51.4% (43.2–59.6) | 75.6% (66.5 – 84.7) | 2.11 (1.40 – 3.16) / 0.64 (0.52 – 0.79) | 26 / 91, *78 / 48* |
| Hoffman-Tinel | 23.2% (16.3 – 30.2) | 87.2% (80.2 – 94.3) | 1.82 (0.97 – 3.4) / 0.88 (0.78 – 0.99) | 23 / 87, *76 / 40* |
| Hypesthesia | 51.4% (43.2 – 59.6) | 84.9% (77.3 – 92.5) | 3.40 (2.01 – 5.75) / 0.57 (0.47 – 0.69) | 36 / 91, *85 / 51* |
| APB weakness | 66.2% (58.4 – 74.0) | 66.3% (56.3 – 76.3) | 1.96 (1.43 – 2.70) / 0.51 (0.39 – 0.67) | 24 / 92, *77 / 54* |
| Median nerve compression | 28.2% (20.8 – 35.6) | 74.4% (65.2 – 83.6) | 1.10 (0.70 – 1.72) / 0.97 (0.82 – 1.13) | 15 / 86, *65 / 38* |
| Square-shaped wrist | 69.0% (61.4 – 76.6) | 73.3% (63.9 – 82.6) | 2.58 (1.79 – 3.72) / 0.42 (0.32 – 0.56) | 30 / 94, *81 / 58* |

* PPV and NPV calculated for 2 different pre-test probabilities: 14% from CTS prevalence in DM study, *63% from CTS in patients referred for NCS studies for possible CTS.*

| | |
|---|---|
| Could useful measures of test accuracy (i.e. likelihood ratios [LR]) be calculated? | Yes |
| What was the magnitude of the LR estimates? | All less than 5 so not very high |
| Was the precision of the LR estimates sufficient? | Some of the confidence limits for the LRs, including that for Tinel's test, include zero. The CI is relatively narrow. |
| If no statistically significant associations detected, was there sufficient power? | No power calculation done. However 225 hands were tested which sounds reasonable. |

| **Section 3: Study Applicability** | |
|---|---|
| Will resulting post-test probabilities affect management and help patients? For which target group(s)? | The value of the Tinel sign varies as expected with the pre-test probability but is never a good test (LR- and LR+ include or are very close to 1). In a low prevalence population, its PV- of 87% is only a 1% improvement on the pre-test prob of not having CTS of 86%. In a high prevalence population with pre-test prob of 63%, PV+ is only 76%, again not really helpful. Because each of the 6 signs evaluated was not evaluated separately, there is the possibility of cross-contamination of one sign result with another, which weakens this study. Each sign is clearly not independent of the other signs, so it is not possible to combine the LRs of the tests in a sequence. |

**Take Home Message:**
I would say, "As a diabetic, you have a 14% chance of having carpal tunnel syndrome. The fact that you are complaining of hand weakness makes the likelihood higher than that but I don't know how high. The examination findings don't really help me decide whether you are more or less likely to have CTS, so I suggest that we organise nerve conduction studies to settle it one way or other."