Evidence-based medicine: how to practice and teach EBM

3. Diagnosis and screening

In this chapter we will help you answer three questions about diagnostic tests:

1. Is this evidence about the accuracy of a diagnostic test valid?

2. Does this (valid) evidence show that this test can accurately distinguish patients who do and do not have a specific disorder?

3. How can I apply this valid, accurate diagnostic test to a specific patient?

After retrieving evidence about a test's accuracy, questions 1 and 2 suggest we need to decide if it's valid and important before we can apply the evidence to our individual patients. As with therapy, the order in which we consider validity and importance is not crucial, and depends on individual preference, but both should be done before applying the study results. Because the screening and early diagnosis of symptomless individuals have some similarities with, but also some crucial differences from, the diagnosis of sick ones, we'll close with a special section devoted to these acts at the interface of clinical medicine and public health.

Figure 3.1 shows the post-test probabilities after positive (upper curve) and negative (lower curve) test results for the range of possible pre-test probabilities of disease. The first question, on validity, asks if we can believe the information in the graph. The second question, on importance, asks if the results show clinically worthwhile shifts in uncertainty (the further apart the post-test curves, the larger this shift). The third question means we need to understand how the test results might change our diagnostic uncertainty on application, not only for the patients in the study but, more importantly, for a particular patient.

To illustrate our discussion, we'll consider the following patient.

**CLINICAL SCENARIO**

Suppose that we’re working up a patient with anemia and think that the probability that she has iron-deficiency anemia is 50% (i.e. the odds are about 50:50 that the anemia is due to iron deficiency). When we present the patient to our boss, we ask for an educational prescription to determine the usefulness of performing a serum ferritin on our patient as a means of detecting iron-deficiency anemia. By the time we’ve tracked down and studied the external evidence, our patient’s serum ferritin comes back at 60 mmol/L. How should we put all this together?
Before looking at the scenario and our three questions, we should take a short detour through the land of “abnormality” so that what we might understand what could be meant by a normal or an abnormal ferritin.

**WHAT IS (AB)NORMAL?**

Most test reports will wind up calling some results “normal” and others “abnormal”. There are at least six definitions of “normal” in common use (listed in Table 3.1). This chapter will focus on definition 5 (“diagnostic” normal), because we think that the first four definitions have important flaws. The first two (the gaussian and percentile definitions) focus just on the diagnostic test results in either a normal (the left-hand cluster in Figure 3.2) or an undifferentiated group of people (an unknown mix of the left and right clusters in Figure 3.2), with no reference standard. They define the “normal range” on the basis of statistical properties (standard deviations or percentiles). They not only imply that all “abnormalities” occur at the same frequency, but suggest that, if we perform more and more diagnostic tests on our patient, we are increasingly likely to find something “abnormal”, thus leading to all sorts of inappropriate further testing.

**Table 3.1 Six definitions of normal**

The third definition of “normal” (culturally desirable) represents the sorts of value judgment seen in fashion advertisements, and at the fringes of the “lifestyle” movement where medicine becomes confused with morality. The fourth (risk factor) definition has the drawback that it “labels” or stigmatizes some patients regardless of whether we can intervene to lower their risk—a big problem with neonatal genetic testing and other screening maneuvers, as you’ll learn in the concluding section of this chapter. The fifth (diagnostic) definition is the one that we will focus on here, and we will show you how to work with it in the next bit of this chapter. The final (therapeutic) definition (does treating at and beyond this level do more good than harm?) is in part an outgrowth of the fourth (risk factor) definition, but has the great clinical advantage that it changes with our knowledge of efficacy. Thus, the definition of normal blood pressure has changed radically over the past few decades as we have learned that treatment of progressively less pronounced elevations of blood pressure does more good than harm.

**IS THIS EVIDENCE ABOUT THE ACCURACY OF A DIAGNOSTIC TEST VALID?**

Having found a possibly useful article about a diagnostic test, how can we quickly critically appraise it for its proximity to the truth? The questions listed in Table 3.2 are for individual reports, but we
can also apply them to the interpretation of a systematic review (overview) of several different studies of the same diagnostic test for the same target disorder.*

Table 3.2 Is this evidence about a diagnostic test valid?

1. Measurement: Was there an independent, blind comparison with a reference ("gold") standard?†

The patients in the study should have undergone both the diagnostic test in question (say, an item of the history or physical examination, a blood test, etc.) and the reference (or "gold") standard (an autopsy or biopsy or other confirmatory "proof" that they do or do not have the target disorder). Sometimes investigators have a difficult time coming up with clear-cut reference standards (e.g. for psychiatric disorders), and we’ll want to give careful consideration to their arguments justifying the selection of their reference standard. Moreover, we caution you against the uncritical acceptance of reference standards, even when they are based on "expert" interpretations of biopsies; for example, in an Evidence Based Medicine note, Kenneth Fleming1 reported that the degree of agreement over and above chance in reading breast, skin and liver biopsies is less than 50%! The results of one test should not be known to the folk who are applying and interpreting the other. (For example, the decision to complete the biopsy should not be dependent on the result of the diagnostic test under study, and the pathologist interpreting the biopsy that comprises the reference standard for the target disorder should be "blind" to the result of the blood test that comprises the diagnostic test under study.) In this way, investigators avoid the conscious and unconscious bias that might otherwise cause the reference standard to be "over-interpreted" when the diagnostic test is positive and "under-interpreted" when it is negative.

2. Representative: Was the diagnostic test evaluated in an appropriate spectrum of patients (like those in whom we would use it in practice)?

Did the report include patients with all the common presentations of the target disorder (including those with its early manifestations), and patients with other commonly confused diagnoses? Studies that confine themselves to florid cases vs. asymptomatic volunteers (a diagnostic "case–control" study) are useful only as a first crude check of the test, because when the diagnosis is obvious to the eye we don’t need any diagnostic test. The really useful articles will be set in the diagnostic dilemmas we face, and include patients with mild as well as severe symptoms, early as well as late cases of the target disorder, and among both treated and untreated individuals.

3. Ascertainment: Was the reference standard ascertained regardless of the diagnostic test result?

When patients have a negative diagnostic test result, investigators are tempted to forego the reference standard, and when the latter is invasive or risky (e.g. angiography) it may be wrong to carry it out on patients with negative test results. To overcome this, many investigators now employ a reference standard for proving that a patient does not have the target disorder; this requires that the patient doesn’t suffer any adverse health outcome during a long follow-up despite the absence of any definitive treatment (for example, convincing evidence that a patient with clinically suspected deep vein thrombosis did not have this disorder would include no ill-effects during a prolonged follow-up despite the absence of antithrombotic therapy).

If the report we’re reading fails one or more of these three tests, we’ll need to consider whether it has a fatal flaw that renders its conclusions invalid. If so, it’s back to more searching (either now or later; if we’ve already used up our time for this week, perhaps we can interest a colleague or trainee in taking this on as an "educational prescription"—see p. 26 if this term is new to you). On the other hand, if the report passes this initial scrutiny and we decide that we can believe its results, and we haven’t already carried out the second critical appraisal step of deciding whether these
results are important, then we can proceed to the next section.

DOES THIS VALID EVIDENCE DEMONSTRATE AN IMPORTANT ABILITY OF THIS TEST TO ACCURATELY DISTINGUISH PATIENTS WHO DO AND DON’T HAVE A SPECIFIC DISORDER?

Sensitivity, specificity, and likelihood ratios

In deciding whether the evidence about a diagnostic test is important, we will focus on the accuracy of the test in distinguishing patients with and without the target disorder. We’ll consider the ability of a valid test to change our minds from what we thought before the test (we’ll call that the “pre-test” probability of some target disorder) to what we think afterwards (we’ll call that the “post-test” probability of the target disorder). Diagnostic tests that produce big changes from pre-test to post-test probabilities are important and likely to be useful to us in our practice.

Returning to our clinical scenario, suppose further that, in filling our prescription, we find a systematic review of several studies of this diagnostic test (evaluated against the reference standard of a bone marrow stain for iron), decide that it is valid (based on the guides in Table 3.2), and find their results as shown in Table 3.3. The prevalence (or study pre-test probability) overall is 809/2579 = 31%. For low ferritin (<65 mmol/L), the post-test probability of iron deficiency anemia among patients in the studies is \( \frac{a}{a+b} = \frac{731}{1001} = 73\% \). This study post-test probability is known as the “positive predictive value”. For high ferritin (>65 mmol/L), the post-test probability of iron deficiency anemia among patients in the studies is \( \frac{c}{c+d} = \frac{78}{1578} = 5\% \). This study post-test probability of 5% means that the study probability of not having iron deficiency anemia after a negative result is 95%; this is known as the “negative predictive value”. So, within the study, the uncertainty regarding iron deficiency has been shifted from the initial 31% to probabilities of either 73% or 5%, which appear to be clinically important shifts.

But we thought our patient’s pre-test probability of iron deficiency anemia was greater than that in the study; in fact, we estimated it to be 50%. We could do a direct adjustment of the predictive values for the patient’s different pre-test probability using the following equation:

\[
\text{Patient post-test odds} = (\text{study post-test odds}) \times \left(\frac{\text{patient pre-test odds}}{\text{study pre-test odds}}\right)
\]

which is analogous to the adjustment of a therapy trial’s NNT (number needed to treat) for the patient’s PEER (patient’s expected event rate; see Ch. 5). This is fine if you have the study in hand, but generally it is easier to derive some test accuracy measures (sensitivity, specificity, and likelihood ratios) and apply these directly to the patient’s individual pre-test probability. So let’s look at these measures.

As you can see from Table 3.3, our patient’s result (60mmol/L) places her in the top row of the table, either in cell “a” or in cell “b”. You might note from Table 3.3 that 90% of patients with iron deficiency have serum ferritin in the same range as our patient \( \frac{a}{a+c} \); that property, the proportion of patients with the target disorder who have positive test results, is called “sensitivity”. The sensitivity of a test is defined as the probability of a positive test given the presence of the target disorder. In diagnostic studies, we sometimes see a description of the proportion of patients who don’t have the target disorder and who have negative (or normal) test results; this is called the “specificity”. Returning to our patient, you might also note that only 15% of patients with other
causes of their anemia have results in the same range as our patient.*

This means that our patient’s result would be about six times as likely (90%/15% = 6) to be seen in someone with iron deficiency anemia as in someone without the condition; that ratio is called the “likelihood ratio” for a positive test result (LR+). The likelihood ratio for a positive test result is:

$$LR+ = \frac{\text{probability of positive test in presence of target disorder}}{\text{probability of positive test in absence of target disorder}}$$

Since we thought ahead of time (before we had the result of the serum ferritin) that our patient’s odds of iron deficiency were 50:50, that’s called a “pre-test odds” of 1:1. As you can see from the formulae towards the bottom of Table 3.3, we can multiply the pre-test odds of 1 by the likelihood ratio of 6 to get the “post-test odds” of iron deficiency anemia after the test (1 × 6 = 6); that’s a post-test odds of 6:1 in favor of iron deficiency anemia. Since, like most clinicians, you may be more comfortable thinking in terms of probabilities than odds, this post-test odds of 6:1 converts (as you can see at the bottom of Table 3.3) to a post-test probability of 6/(6 + 1) = 6/7 = 86%. (To check yourself out on these calculations, try calculating the post-test probability for the same ferritin result for a patient who, like those in Table 3.3, has a pre-test odds of 0.45.† You’ll know you did it right if you wind up with an answer for post-test probability that is identical to its equivalent, the positive predictive value.)

Note that we could use the graph in Figure 3.1 (created with a program available on the accompanying CD) which allows us to determine the post-test probability by drawing a line from the 50% pre-test probability to the post-test positive (+ve) line and across to the 86% post-test probability.

Extremely high values of sensitivity and specificity are useful, but not for the reasons you may think. When a test has a very high sensitivity (such as the loss of retinal vein pulsation in increased intracranial pressure), a negative result (the presence of pulsation) effectively rules out the diagnosis (of raised intracranial pressure). One of our clinical clerks suggested that we apply the mnemonic “SnNout” to such findings: when a sign has a high sensitivity (Sn), a negative (N) result rules out (out) the diagnosis. Similarly, when a sign has a very high specificity (Sp), such as the face of a child with Down’s syndrome, a positive (P) result effectively rules in (in) the diagnosis (of Down’s); not surprisingly, our clinical clerks call such a finding a “SpPin”. We’ve listed some SpPins and SnNouts in Table 3.4, and have generated a longer list on our website (www.cebm.utoronto.ca). A perfect test would be both a SpPin and a SnNout (please let us know if you find one!). A useless test is one that leaves the probabilities unchanged (and the post-test lines both lie on the diagonal of the pre-test/post-test graph), which occurs when the sensitivity (%) + specificity (%)−100% (known as the Youden index) is zero.

We can generate likelihood ratios directly, or by reference to the sensitivity and specificity using formulae in Table 3.3:

The likelihood ratio for a positive test result is

$$LR+ = \frac{\text{sensitivity}}{\text{1 - specificity}}$$

The likelihood ratio for a negative test result is

$$LR- = \frac{\text{1 - sensitivity}}{\text{specificity}}$$

**CAN I APPLY THIS VALID, IMPORTANT DIAGNOSTIC TEST TO A SPECIFIC PATIENT?**

Having found a valid systematic review or individual report about a diagnostic test, and having decided that its accuracy is sufficiently high to be useful, how do we apply it to our patient? To transfer the study results, adapt them to our patient’s unique pre-test probability, and decide this would be clinically useful, there are three questions we should ask; these are summarized in Table 3.5 and discussed below.
Table 3.4 Some SpPins and SnNouts

Table 3.5 Questions to answer in applying a valid diagnostic test to an individual patient

1. **Is the diagnostic test available, affordable, accurate, and precise in our setting?**

We obviously can’t order a test that is not available. Even if it is available, we may want to check around to be sure that it’s performed and interpreted in a competent, reproducible fashion, and that its potential consequences (see below) justify its cost. For example, some of us work on medical units at more than one hospital and have found that the labs at these different hospitals use different assays for assessing ferritin D-dimer values, making the interpretation for clinicians more challenging! Moreover, diagnostic tests often behave differently among different subsets of patients, generating higher likelihood ratios in later stages of florid disease, and lower likelihood ratios in early, mild stages.

At least some diagnostic tests based on symptoms or signs lose power as patients move from primary care to secondary and tertiary care. Reference back to Table 3.3 can show you why. If patients are referred onward, in part because of symptoms, their primary care clinicians will be sending along patients in both cells “a” and “b”, and subsequent evaluations of the accuracy of their symptoms will tend to show falling specificity due to the referral of patients with false-positive findings. If we think that any of these factors may be operating, we can try out what we judge to be clinically sensible variations in the likelihood ratios for the test result and see whether the results alter our post-test probabilities in a way that changes our diagnosis (the short-hand term for this sort of exploration is “sensitivity analysis”).

2. **Can we generate a clinically sensible estimate of our patient’s pre-test probability?**

This is a key topic, and deserves its own “section-within-a-section”. How can we estimate our patient’s pre-test probability? We’ve used five different sources for this vital information: clinical experience, regional or national prevalence statistics, practice databases, the original report we used for deciding on the accuracy and importance of the test, and studies devoted specifically to determining pre-test probabilities. Although the last is ideal, we’ll take each in turn.

First, we can recall our clinical experience with prior patients who presented with the same clinical problem, and backtrack from their final diagnoses to their pre-test probabilities. While easily and quickly accessed, our memories are often distorted by our last patient, our most dramatic (or embarrassing) patient, our fear of missing a rare but treatable cause, and the like, so we use this source with caution.

And if we’re early in our careers, we may not have enough clinical experience to draw upon. Thus, while we always use our remembered cases, we need to learn to supplement them with other sources, unless we have the time and energy to document all of our diagnoses and generate our
own database.

Second, we could turn to regional or national prevalence statistics on the frequencies of the target disorders in the general population or some subset of it. Estimates from these sources are only as good as the accuracy of their diagnoses, and, although they can provide some guidance for "baseline" pre-test probabilities before taking symptoms into account (useful, say, for patients walking into a general practice), we may be more interested in pre-test probabilities in just those persons with a particular symptom.

Third, we could overcome the foregoing problems by tracking down local, regional or national practice databases that collect patients with the same clinical problem and report the frequency of disorders diagnosed in these patients. Although some examples exist, such databases are mostly things of the future. As before, their usefulness will depend on the extent to which they use sensible diagnostic criteria and clear definitions of presenting symptoms.

Fourth, we could simply use the pre-test probabilities observed in the study we critically appraised for the accuracy and importance of the diagnostic test. If they really did sample the full spectrum of patients with the symptom or clinical problem (the second of our accuracy guides), we can extrapolate the pre-test probability from their study patients (or some subgroup of it) to our patient.

Fifth, and finally, we could track down a research report of a study expressly devoted to documenting pre-test probabilities for the array of diagnoses that present with a specific set of symptoms and signs similar to those of our patient. When done well, among patients closely similar to our patient, these studies provide the least biased source of pre-test probabilities for our use. Such studies are challenging to carry out, and one of us led the group who generated guides for their critical appraisal.² We've summarized these guides in Table 3.6. You'll see that most of them are already familiar to you, for they apply equally to reports of the accuracy and importance of diagnostic tests. We've provided examples of pre-test probabilities in Table 3.7, and will add to this list on our website (www.cebm.utoronto.ca).

### Table 3.6 Guides for critically appraising a report about pre-test probabilities of disease

### Table 3.7 Examples of pre-test probabilities

---

3. **Will the resulting post-test probabilities affect our management and help our patient?**

There are three elements of the answer to this final question and we begin with the bottom line: Could the test results move us across some threshold that would cause us to stop all further testing? Two thresholds should be borne in mind, as shown in Figure 3.3. First, if the diagnostic test
was negative or generated a likelihood ratio down near 0.1, the post-test probability might become so low that we would abandon the diagnosis we were pursuing, and turn to other diagnostic possibilities. Put in terms of thresholds, this negative test result has moved us from above to below the “test threshold” in Figure 3.3 and we won’t do any more tests for that diagnostic possibility.

Second, if the diagnostic test came back positive or generated a high likelihood ratio, the post-test probability might become so high that we would also abandon further testing because we’ve made our diagnosis and would now move to choosing the most appropriate therapy; in these terms, we’ve now crossed from below to above the “treatment threshold” in Figure 3.3.

It is only if our diagnostic test result leaves us stranded between the test and treatment thresholds that we would continue to pursue that initial diagnosis by performing other tests. Although there are some very fancy ways of calculating test–treatment thresholds from test accuracy and the risks and benefits of correct and incorrect diagnostic conclusions,* intuitive test–treatment thresholds are commonly used by experienced clinicians and are another example of individual clinical expertise.

We suggest you look at several pre-test scenarios using a post-test probability graph (Figure 3.1) to get a feel of when the test results in clinically useful shifts in decisions.

We may not cross a test–treatment threshold until we’ve performed several different diagnostic tests, and here is where another nice property of the likelihood ratio comes into play: provided the tests are independent we can “chain” the likelihood ratios. The post-test odds that result after the first diagnostic test we apply becomes the pre-test odds for our second diagnostic test. Hence we can simply keep multiplying the running product by the likelihood ratio generated from the next test. For example, when a 45-year-old man walks into our office, his pre-test probability of >75% stenosis of one or more of his coronary arteries is about 6%. Suppose that he gives us a history of atypical chest pain (only two of the three symptoms of substernal chest discomfort, brought on by exertion, and relieved in less than 10 minutes by rest are present, generating a likelihood ratio of about 13), and that his exercise ECG reveals 2.2 mm of non-sloping ST-segment depression (generating a likelihood ratio of about 11). Then his post-test probability for coronary stenosis is his pre-test probability (converted into odds) times the product of the likelihood ratios generated from his history\(^{13}\) and exercise ECG,\(^{11}\) with the resulting post-test odds converted back to probabilities (through dividing by its value +1), i.e.:

\[
(0.06/0.94) \times 13 \times 11 = 9.13, \text{ and then } 9.13/10.13 = 90\%
\]

The final result of these calculations is strictly accurate as long as the diagnostic tests being combined are “independent”; that is, given the “true” condition, the accuracy of one test does not further depend on another test. However, some dependence is common, and means we tend to over-estimate the informativeness of the multiple tests. Accordingly, we would want the calculated post-test probability at the end of this sequence to be comfortably above our treatment threshold before we would act upon it. This additional example of how likelihood ratios make lots of implicit diagnostic reasoning explicit is another argument in favor of seeking reports of overall likelihood ratios for sequences or clusters of diagnostic tests (see section on multiple tests below).

We should have kept our patient informed as we worked our way through all the foregoing considerations, especially if we’ve concluded that the diagnostic test is worth considering. If we haven’t yet done so, we certainly need to do so now. Every diagnostic test involves some invasion of privacy, and some are embarrassing, painful, or dangerous. We’ll have to be sure that the patient is an informed, willing partner in the undertaking. In particular, they should be aware of the possibility of false-positive or false-negative outcomes so that this is not a surprise when they return to
discuss the results. The ultimate question to ask about using any diagnostic test is whether its consequences (reassurance when negative, labeling and possibly generating awful diagnostic and prognostic news if positive, leading to further diagnostic tests and treatments, etc.) will help our patient achieve his or her goals of therapy. Included here are considerations of how subsequent interventions match clinical guidelines or restrictions on access to therapy designed to optimize the use of finite resources for all members of our society.

**MULTILEVEL LIKELIHOOD RATIOS**

The more extreme a test result is, the more persuasive it is. Although the dichotomized serum ferritin sensitivity (90%) and specificity (85%) look impressive, expressing a test's accuracy with level-specific likelihood ratios reveals its even greater power and, in this particular example, shows how we can be misled by the restriction to just two levels (positive and negative) of the test result. Many test results, like serum ferritin, can be divided into several levels, and in Table 3.8 we show you a particularly useful way of dividing test results for ferritin into five levels.

<table>
<thead>
<tr>
<th>Table 3.8 The usefulness of five levels of a diagnostic test result</th>
</tr>
</thead>
</table>

When this is done, we see how much more informative extreme ferritin results are. The likelihood ratio for the “very positive” result is a huge 52, so that one extreme level of the test result can be shown to rule in the diagnosis, and in this case we can SpPin 59% (474/809) of the patients with iron deficiency anemia, despite the unimpressive sensitivity (59%) that would have been achieved if the ferritin results had been split just below this level. Likelihood ratios of 10 or more, when applied to pre-test probabilities of 33% or more (0.33/0.67 = pre-test odds of 0.5) will generate post-test probabilities of 5/6 = 83% or more.

Similarly, the other extreme level (>95) is a SnNout 75% (1332/1770) for those who do not have iron deficiency anemia (again despite a not-very-impressive specificity of 75%). Likelihood ratios of 0.1 or less, when applied to pre-test probabilities of 33% or less (0.33/0.67 = pre-test odds of 0.5), will generate post-test probabilities of 0.05/1.05 = 5% or less. The two intermediate levels (moderately positive and moderately negative) can move a 50% prior probability (pre-test odds of 1:1) to the useful but not necessarily diagnostic post-test probabilities of 4.8/5.8 = 83% and 0.39/1.39 = 28%. The indeterminate level (“neutral”) in the middle (containing about 10% of both sorts of patients) can be seen to be uninformative, with a likelihood ratio of 1. When diagnostic test results are around 1.0, we’ve learned nothing by ordering them. To give you a better “feel” for this, the impact of different likelihood ratios on different pre-test probabilities are shown in Figure 3.4.

We've provided additional examples of likelihood ratios on this book’s website (www.cebm.utoronto.ca).

<table>
<thead>
<tr>
<th>Figure 3.4 Probability revision graph using the likelihood ratios for four levels of ferritin.</th>
</tr>
</thead>
</table>

Figure 3.4 can be used to do some approximate calculations. An easier way of manipulating all
these calculations is the nomogram of Figure 3.5 (also provided in the pocket cards that come with this book). You can check out your understanding of this nomogram by using it to replicate the results of Tables 3.3 and 3.8.

Now return to our patient with a pre-test probability for iron deficiency of 50% and a ferritin result of 60mmol/L. To your surprise (we reckon!), our patient’s test result generates an indeterminate likelihood ratio of only 1, and the test which we thought might be very useful, based on the old sensitivity and specificity way of looking at things, really hasn’t been helpful in moving us toward the diagnosis. We’ll have to think about other tests (including perhaps the reference standard of a bone marrow examination) to sort out her diagnosis.

More and more reports of diagnostic tests are providing multilevel likelihood ratios as measures of their accuracy. When their abstracts report only sensitivity and specificity, we can sometimes find a table with more levels and generate our own set of likelihood ratios; at other times, we can find a scatterplot (of test results vs. diagnoses) that is good enough for us to be able to split them into levels.

MULTIPLE TESTS

Some reports of diagnostic tests go beyond even likelihood ratios, and one of their extensions deserves mention here. This extension considers multiple diagnostic tests as a cluster or sequence of tests for a given target disorder. These multiple results can be presented in different ways, either as clusters of positive/negative results or as multivariate scores, and in either case they can be ranked and handled just like other multilevel likelihood ratios. When they perform (nearly) as well in a second, independent (“test”) set of patients, we often refer to them as “clinical prediction guides” (CPGs). In appraising the validity of a study of a CPG, we need to consider a 4th question in addition to those above:

Was the cluster of tests validated in a second, independent, group of patients?

Diagnostic tests are predictors, not explainers, of diagnoses. As a result, their initial evaluation cannot distinguish between real diagnostic accuracy for the target disorder and chance associations due to idiosyncrasies in the initial (“training” or “derivation”) set of patients. This problem is compounded for clusters of diagnostic features (often called “clinical prediction guides”), where the large numbers of possible tests considered mean we may over-estimate the value of the few chosen in the CPG. The best indicator of accuracy in these situations is the demonstration of similar levels of accuracy when the test or cluster is evaluated in a second, independent (or “test”), set of patients. If it performs well in this “test” set, we are reassured about its accuracy. If it performs poorly, we should look elsewhere. And if no “test set” study has been carried out, we’d be wise to reserve judgment. CPGs are also used to help establish prognosis.

PRACTICING EBM IN REAL-TIME

CPGs often include several variables that we have to remember when trying to apply them to our patients. Several colleagues have attempted to make this easier and have provided interactive versions of clinical prediction guides available on websites; we’ve provided some of these on the accompanying website.

LEARNING AND TEACHING WITH CATS
Now that we have invested precious time and energy into finding and critically appraising an article, it would be a shame not to summarize and keep track of it so that we (and others) can use it again in the future. The means that Stephane Sauve, Hui Lee, and Mike Farkouh, residents on Dave Sackett’s clinical service a few years ago, invented to accomplish this was to create a standardized one-page summary of the evidence organized as a “critically appraised topic”, which they called a “CAT”. We’ll discuss these in more detail in Chapter 7, but note here that they provide a brief summary of the evidence that we can store for later retrieval. To help generate CATs, we’ve provided a copy of CATMaker on the CD, or it can be downloaded from the book’s website (www.cebm.utoronto.ca) or from www.cebm.net. This software takes learners step by step through the creation of a CAT, calculates some of the clinically useful measures of therapy (NNTs, likelihood ratios), and automatically generates their confidence intervals. The CATMaker allows CATs to be saved (even in a draft “kitten” form that can be retrieved for later revision) or outputted in “.txt” files or “.html” formats. This means that you can create your own database to store your CATs in an easily retrievable format, make copies available to your students and colleagues, or even place them on your local intranet. Now take a look at the CAT we generated for ferritin.

SCREENING AND CASE-FINDING

So far, this chapter has focused on making a diagnosis for sick patients who have come to us for help. They are asking us to diagnose their ills and to help them as best we can, and only charlatans guarantee them longer life at the initial encounter. This final section of the chapter focuses on making early diagnoses of pre-symptomatic disease among well individuals in the general public (we’ll call that “screening”), or among patients who have come to us for some other unrelated disorder (we’ll call that “case-finding”). Individuals whom we might consider for screening and case-finding are not ill from the target disorders, so we are soliciting them with the promise (overt or covert) that they will live longer, or at least better, if they let us test them. Accordingly, the evidence we need about the validity of screening and case-finding goes beyond the accuracy of the test for early diagnosis; we need hard evidence that patients are better off, in the long run, when such early diagnosis is achieved.

CAT Ferritin can diagnose iron deficiency in the elderly

Clinical bottom line Serum ferritin can be very useful in diagnosing iron deficiency anemia in the elderly.

Clinical scenario. 75-year-old retired school-teacher (in for a check-up) found to have an Hb of 100, with an MCV of 80, a negative history and physical, and on no medications that are likely to suppress her marrow or cause a bleed. I think her probability of iron deficiency is 1 out of 2 or 50%.

Three-part question. In an elderly symptomless woman with mild anemia, would a serum ferritin help determine whether her bone marrow iron stores were depleted?

Search terms. Searching ACP Journal Club using the terms “iron deficiency anemia” and “ferritin”, we find a study that appears to be of interest and that provides a link to an overview of this topic.

The study

Independent…? Yes

Blind…? Yes

Standard applied regardless of test result…? Yes

Appropriate spectrum…? Can’t tell

Target disorder and gold standard. Bone marrow, stained for iron.
Patients. Consecutive anemic patients in several inpatient and outpatient settings. Transfused patients excluded.

Diagnostic test. Serum ferritin by radioimmunoassay.

The evidence

<table>
<thead>
<tr>
<th>Test result</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>Prop.</td>
</tr>
<tr>
<td>&lt;15</td>
<td>474</td>
<td>0.59</td>
</tr>
<tr>
<td>15–34</td>
<td>175</td>
<td>0.22</td>
</tr>
<tr>
<td>35–64</td>
<td>82</td>
<td>0.10</td>
</tr>
<tr>
<td>65–94</td>
<td>30</td>
<td>0.04</td>
</tr>
<tr>
<td>≥95</td>
<td>48</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Comments

1. For elderly patients with symptomless anemia, go to the CT on anemia in the elderly to determine the yields from upper and lower gastrointestinal investigations.

2. Lots of labs are very slow in returning ferritin requests.

Expiry date: 2005.

REFERENCES


All screening and case-finding, at least in the short-run, harms some people. Early diagnosis is just that: people are "labeled" as having, or as being at high risk for developing, some pretty awful diseases (breast cancer, stroke, heart attack, and the like). And this labeling takes place months, years, or even decades before the awful diseases will become manifest as symptomatic illness (often in only a small portion of those who screen positive). Labeling hurts. For example, a cohort of working men studied both before and after they were labeled hypertensive displayed increased absenteeism, decreased psychological well-being, and progressive loss of income in comparison to their normotensive workmates (and these bad effects could not be blamed on drug side-effects, for they occurred even among men who were never treated!). What's even worse is that those with false-positive screening tests will experience only harm (regardless of the efficacy of early treatment). But even individuals with true-positive tests who receive efficacious treatment have had "healthy time" taken away from them; early diagnosis may not make folks live longer, but it surely makes all of them "sick" longer!

We've placed this discussion at the end of the chapter on diagnosis, with the chapter on therapy the next but one, on purpose. In order to decide whether screening and case-finding do more good than harm, we'll have to consider the validity of claims about both the accuracy of the early diagnostic test and the efficacy of the therapy that follows it. We've summarized the guides for doing this in Table 3.9. Its elements are discussed in greater detail elsewhere (consult the "Further reading" at the end of this chapter).
Table 3.9 Guides for deciding whether a screening or early diagnostic maneuver does more good than harm

1. Is there RCT evidence that early diagnosis really leads to improved survival, or quality of life, or both?

Earlier detection will always appear to improve survival. The "lead time" (between screen and usual detection; Figure 3.6) is always added to your apparent survival whether or not there is any real change. Follow-up studies of placebo groups in randomized controlled trials (RCTs) have taught us that patients who faithfully follow health advice (by volunteering for screening or by taking their medicine) are destined for better outcomes before they begin. And, early diagnostic maneuvers preferentially identify patients with slower progressing, more benign disease. As a result, the only evidence we can trust in determining whether early diagnosis does more good than harm is a true experiment in which individuals were randomly assigned. As shown in Figure 3.7, this may be: (1) randomization to either undergo the early detection test (and, if truly positive, treated for the target disorder), or to be left alone (and only treated if and when symptomatic disease develops); or (2) be screened, and positives randomized to early treatment or usual care. The latter sort of evidence has been used for showing the benefits (and harms) of detecting raised blood pressure and cholesterol. The former sort of evidence showed the benefit of mammography for reducing deaths from breast cancer,* and showed the uselessness (indeed, harm) of chest X-rays for lung cancer. Ideally, their follow-up will consider functional and quality-of-life outcomes as well as mortality and discrete clinical events, and we should not be satisfied when the only favorable changes are confined to "risk factors".

2. Are the early diagnosed patients willing partners in the treatment strategy?

Even when therapy is efficacious, patients who refuse or forget to take it cannot benefit from it and are left with only the damage produced by labeling. Early diagnosis will do more harm than good to
these patients, and we forget the magnitude of this problem at their peril (even by self-report, only half of patients describe themselves as "compliant"). There are quick ways of diagnosing low compliance and we'll show them to you in Chapter 5 (they comprise looking for non-attendance and non-responsiveness, and by non-confrontational questioning), but this is a diagnosis that you need to establish before, not after, you carry out any screening or case-finding.

3. How do benefits and harms compare in different people and with different screening strategies?

4. Do the frequency and severity of the target disorder warrant the degree of effort and expenditure?

This question raises, at the levels of both our individual practice and our community, the unavoidable question of rationing. Is going after the early diagnosis of this condition worth sacrificing the other good we could accomplish by devoting our own or our town’s resources to some other purpose?

We don’t want to sound too gloomy here, and won’t leave this topic without pointing you to places where you can find some of the triumphs of screening and case-finding. A good place to start is the Canadian Task Force on the Periodic Health Examination, where there are some rigorously evaluated ones.

TIPS FOR TEACHING AROUND DIAGNOSTIC TESTS

We usually begin by asking learners why we perform diagnostic tests, because they often respond: “To find out what’s wrong with the patient [dummy!”]. This provides an opening for helping them to recognize that diagnosis is not about finding absolute truth but about limiting uncertainty, and establishes both the necessity and the logical base for introducing probabilities, pragmatic test–treatment thresholds, and the like. It’s also a time to get them to start thinking about what they’re going to do with the results of the diagnostic test and about whether doing the test will really help their patient (maybe they’ll conclude that the test isn’t necessary!). A useful sequence is to elicit some disagreement between students (e.g. about a measurement or sign), but don’t step in and suggest a “right” answer. The elicited disagreement can be used as an opening to unreliability and uncertainty. Comparison to a “gold standard” can introduce the issues of validity. While the formal calculations can be difficult, the qualitative ideas of SpPin and SnNout can be used, then be introduced to start students thinking about the accuracy and utility of a test.

When teaching about early diagnosis, we often challenge our learners with the statement “Even when therapy is worthless, early diagnosis always improves survival!”, and then help them recognize the distortions that arise from drawing conclusions about volunteers, from starting survival measurements unfairly early in screened patients, and from failing to recognize that early detection tests preferentially identify slowly—rather than rapidly—progressive disease. Once they’ve grasped those ideas, we think they’re safe from the evangelists of early diagnosis.

REFERENCES


Further reading


© Elsevier Ltd 2005. Straus et al.: Evidence-based medicine
Table 3.3: Results of a systematic review of serum ferritin as a diagnostic test for iron deficiency anemia

<table>
<thead>
<tr>
<th>Target disorder (iron deficiency anemia)</th>
<th>Present</th>
<th>Absent</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive (&lt;65 mmol/L)</td>
<td>731 a</td>
<td>270 b</td>
<td>1001 a+b</td>
</tr>
<tr>
<td>Negative (≥65 mmol/L)</td>
<td>c 78</td>
<td>d 1500</td>
<td>c+d 1578</td>
</tr>
<tr>
<td>Totals</td>
<td>a+c 809</td>
<td>b+d 1770</td>
<td>a+b+c+d 2579</td>
</tr>
</tbody>
</table>

Prevalence = (a + c)/(a + b + c + d) = 809/2579 = 31%.
Positive predictive value = a/(a + b) = 731/1001 = 73%.
Negative predictive value = d/(c + d) = 1500/1578 = 95%.
Sensitivity = a/(a + c) = 731/809 = 90%.
Specificity = d/(b + d) = 1500/1770 = 85%.
LR+ = sensitivity/(1 – specificity) = 90%/15% = 6.
LR – = (1 – sensitivity)/specificity = 10%/85% = 0.12.
Study pre-test odds = prevalence/(1 – prevalence) = 31%/69% = 0.45.
Post-test odds = pre-test odds x likelihood ratio.
Post-test probability = post-test odds/post-test odds + 1.

© Elsevier Ltd 2005. Straus et al.: Evidence-based medicine
<table>
<thead>
<tr>
<th>Target disorder</th>
<th>SpPin (and specificity) [presence rules in the target disorder]</th>
<th>SnNout (and sensitivity) [absence rules out the target disorder]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascites (by imaging or tap)</td>
<td>Fluid wave (92%)</td>
<td>History of ankle swelling (93%)</td>
</tr>
<tr>
<td>Pleural effusion</td>
<td>Auscultatory percussion note loud and sharp (100%)</td>
<td>Auscultatory percussion note soft and/or dull (96%)</td>
</tr>
<tr>
<td>Increased intracranial pressure (by CT scan or direct measurement)</td>
<td></td>
<td>Loss of spontaneous retinal vein pulsation (100%)</td>
</tr>
<tr>
<td>Cancer as a cause of lower back pain (by further investigation)</td>
<td></td>
<td>Age &gt;50 or cancer history or unexplained weight loss or failure of conservative therapy (100%)</td>
</tr>
<tr>
<td>Sinusitis (by further investigation)</td>
<td>Maxillary toothache or purulent nasal secretion or poor response to nasal decongestants or abnormal transtilumination or history of colored nasal discharge (LR = 0.1)</td>
<td></td>
</tr>
<tr>
<td>Alcohol abuse or dependency</td>
<td>Yes to &gt;3 of the CAGE questions (99.8%)</td>
<td></td>
</tr>
<tr>
<td>Splenomegaly (by imaging)</td>
<td>Positive percussion (Nixon method) and palpation</td>
<td></td>
</tr>
<tr>
<td>Non-urgent cause for dizziness</td>
<td>Positive head-hanging test and either vertigo or vomiting (94%)</td>
<td></td>
</tr>
</tbody>
</table>

*To find more examples, and to nominate additions to the databank of SpPins and SnNouts, refer to this textbook's website at: <http://www.library.utoronto.ca/medicine/ebn/>.

7 JAMA 1993; 270; 2218–21.

© Elsevier Ltd 2005. Straus et al.: Evidence-based medicine
Table 3.5 Questions to answer in applying a valid diagnostic test to an individual patient

1. Is the diagnostic test available, affordable, accurate, and precise in our setting?

2. Can we generate a clinically sensible estimate of our patient's pre-test probability?
   - From personal experience, prevalence statistics, practice databases, or primary studies?
   - Are the study patients similar to our own?
   - Is it unlikely that the disease possibilities or probabilities have changed since this evidence was gathered?

3. Will the resulting post-test probabilities affect our management and help our patient?
   - Could it move us across a test-treatment threshold?
   - Would our patient be a willing partner in carrying it out?
   - Would the consequences of the test help our patient reach his or her goals in all this?

© Elsevier Ltd 2005. Straus et al.: Evidence-based medicine
<table>
<thead>
<tr>
<th>Table 3.6 Guides for critically appraising a report about pre-test probabilities of disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is this evidence about pre-test probability valid?</td>
</tr>
<tr>
<td>• Did the study patients represent the full spectrum of those who present with this clinical problem?</td>
</tr>
<tr>
<td>• Were the criteria for each final diagnosis explicit and credible?</td>
</tr>
<tr>
<td>• Was the diagnostic work-up comprehensive and consistently applied?</td>
</tr>
<tr>
<td>• For initially undiagnosed patients, was follow-up sufficiently long and complete?</td>
</tr>
<tr>
<td>2. Is this evidence about pre-test probability important?</td>
</tr>
<tr>
<td>• What were the diagnoses and their probabilities?</td>
</tr>
<tr>
<td>• How precise were these estimates of disease probability?</td>
</tr>
</tbody>
</table>

© Elsevier Ltd 2005. Straus et al.: Evidence-based medicine
### Table 3.7 Examples of pre-test probabilities

<table>
<thead>
<tr>
<th>Symptom or clinical problem</th>
<th>Source</th>
<th>Work-up</th>
<th>Disease probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anemia of chronic disease</td>
<td>90 adults admitted to a general medical ward of a county hospital in North America&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Clinical exam, blood testing, selected other testing</td>
<td>Infection, 36%; Inflammatory, 6%; Malignant, 19%; Renal, 15%; Other, 24%</td>
</tr>
<tr>
<td>Dizziness &gt;2 weeks</td>
<td>100 adult patients seen in primary care sites in one North American city&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Clinical exam, neurological, ophthalmologic, and psychological testing, selected other tests</td>
<td>Vertigo, 54%; Psychiatric, 16%; Multicausal, 13%; Other, 19%; Unknown, 8%</td>
</tr>
<tr>
<td>Dyspnea &gt;4 weeks, unexplaned by exam, radiograph and spirometry</td>
<td>72 adults referred to outpatient pulmonary clinic in North America&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Standardized exam, testing and treatment</td>
<td>Respiratory, 36%; Cardiac, 14%; Hyperventilation, 19%; Other, 12%; Unexplained, 19%</td>
</tr>
<tr>
<td>Epilepsy, new onset in adults</td>
<td>333 adults presenting to a major urban emergency department in North America excluded alcohol, head trauma, hypoglycemia&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Standardized exam, testing (including head CT), and treatment</td>
<td>Unknown, 44%; Stroke, 11%; Tumor, 7%; Infection, 17%; Metabolic, 5%; Other, 19%</td>
</tr>
<tr>
<td>Palpitations</td>
<td>190 patients from acute care sites in one North American city&lt;sup&gt;e&lt;/sup&gt;</td>
<td>Clinical exam, cardiac and psychological testing, selected other tests</td>
<td>Cardiac, 43%; Psychiatric, 31%; Miscellaneous, 10%; Unknown, 16%</td>
</tr>
</tbody>
</table>

<sup>a</sup> Am J Med 1989; 87: 538–44.
<sup>c</sup> Chest 1991; 100: 1293–9.

© Elsevier Ltd 2005. Straus et al.: Evidence-based medicine
© Elsevier Ltd 2005. Straus et al.: Evidence-based medicine
<table>
<thead>
<tr>
<th>Diagnostic test result</th>
<th>Serum ferritin (mmol/L)</th>
<th>Target disorder (Iron deficiency) present</th>
<th>Target disorder absent</th>
<th>Likelihood ratio</th>
<th>Diagnostic impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very positive</td>
<td>&lt; 15</td>
<td>474 59 (474/809)</td>
<td>20 1.1 (20/1770)</td>
<td>5.2</td>
<td>Rule-in “SpPin”</td>
</tr>
<tr>
<td>Moderately positive</td>
<td>15–34</td>
<td>175 22 (175/809)</td>
<td>79 4.5 (79/1770)</td>
<td>4.8</td>
<td>Intermediate high</td>
</tr>
<tr>
<td>Neutral</td>
<td>35–64</td>
<td>82 10 (82/809)</td>
<td>171 10 (171/1770)</td>
<td>1</td>
<td>Indeterminate</td>
</tr>
<tr>
<td>Moderately negative</td>
<td>65–94</td>
<td>30 3.7 (30/809)</td>
<td>168 9.5 (168/1770)</td>
<td>0.39</td>
<td>Intermediate low</td>
</tr>
<tr>
<td>Extremely negative</td>
<td>≥ 95</td>
<td>48 5.9 (48/809)</td>
<td>1332 75 (1332/1770)</td>
<td>0.08</td>
<td>Rule-out “SnNout”</td>
</tr>
</tbody>
</table>

© Elsevier Ltd 2005. Straus et al.: Evidence-based medicine
<table>
<thead>
<tr>
<th></th>
<th>Guides for deciding whether a screening or early diagnostic maneuver does more good than harm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Is there RCT evidence that early diagnosis really leads to improved survival, or quality of life, or both?</td>
</tr>
<tr>
<td>2.</td>
<td>Are the early diagnosed patients willing partners in the treatment strategy?</td>
</tr>
<tr>
<td>3.</td>
<td>How do benefits and harms compare in different people and with different screening strategies?</td>
</tr>
<tr>
<td>4.</td>
<td>Do the frequency and severity of the target disorder warrant the degree of effort and expenditure?</td>
</tr>
</tbody>
</table>

© Elsevier Ltd 2005. Straus et al.: Evidence-based medicine
Screen detection

Usual detection

No screening → Death

Lead time → Ineffective screening → Death

Apparent increase in survival

Lead time → Effective screening → Death

Real increase in survival

© Elsevier Ltd 2005. Straus et al.: Evidence-based medicine
© Elsevier Ltd 2005. Straus et al.: Evidence-based medicine