Workshop 7 — Appraising a study on diagnosis

In most clinical practice settings, the first task of the clinician is to arrive at a diagnosis for a complaint presented by a patient. In most circumstances, this process includes taking a history, doing a physical examination, and if necessary, ordering some laboratory tests. The purpose of the whole diagnostic process is to either increase the probability of being able to make a certain (usually more serious) diagnosis among a list of possibilities (differential diagnosis) up to a high enough level (rule in) that interventions, including further invasive investigations, are warranted. Alternatively it could be to reduce that probability to a low enough level (rule out), so that one could just let the patient go without further investigations or interventions. In theory, every question we ask in taking a medical history and every step in a physical examination can be regarded as a diagnostic test, but these steps in the diagnostic process are seldom studied and evaluated as such. More often, specific diagnostic tests or screening tests (laboratory, imaging, psychometric tools, etc) are evaluated on their performance in various studies.

Evaluation of test performance is usually done by conducting a cross-sectional study to compare the classification of subjects into those with or without the disease by the test result to that of a 'gold standard', which is usually regarded as reflecting the truth. Some studies also examine how good a certain test or combination of tests might be in predicting a certain health outcome that can be more objectively defined over time, eg clinically present or recurrence of malignancy, or death.[1] The discussion in this Workshop will focus on the former type of study for single diagnostic/screening tests.

The four major questions to be answered when appraising a study on diagnosis are shown below.

(1) Do the study results demonstrate an important ability of the diagnostic test to accurately discriminate between persons who have and do not have a specific disease (ie clinical importance)?

The property or accuracy of a diagnostic test is commonly measured by the following parameters:

- *Sensitivity*—the ability of the test to detect subjects with the disease; the higher the proportion (approaching 1 or 100%), the better the test;
- *Specificity*—the ability of the test to detect subjects without the disease; the higher the proportion, the better the test;
- *Likelihood ratio*—a ratio to compare the probability (likelihood) of a test being positive (or negative) in subjects with the disease to that among subjects without the disease; it is a composite parameter that incorporates information on sensitivity and specificity (sensitivity/[1 – specificity] for a positive test result and [1 – sensitivity]/specificity for a negative test result); the further it deviates from 1 (the non-discriminative value), the better the test; a value of 10 or 0.1 indicates very high discriminatory ability; and
- *Area under curve (AUC) of the receiver operating characteristic (ROC) curve*—an indication of the proportion of subjects that can be correctly classified by the test; the higher the proportion (approaching 1 or 100%), the better the test; a value of 0.5 indicates no discriminative value.

If the result of a diagnostic test is on a continuous scale, eg serum concentration of a certain biomarker, it would be necessary to identify the 'best cut-off point' for defining abnormality. Such a point is frequently identified from the ROC curve, visually or mathematically, as the point with the shortest distance from the left upper origin of the two axes (1 for sensitivity and 0 for 1-specificity or 1 for specificity) with the aim of maximising both sensitivity and specificity (optimal), under the dilemma that they run in opposite directions on a typical ROC curve. The clinical importance of such a mechanically defined optimal cut-off point needs to be examined in the light of the costs of missing a case (false negative) or over-diagnosing a case (false positive). This in turn depends on the prevalence (or pre-test probability) of the disease in the tested group of (or individual) patients, as well as other medical considerations.

(2) Are the study results about the accuracy of a diagnostic test valid?

As discussed in the general approach to critical appraisal,[2] valid study results mean that they are free from biases. The three major sources of bias should be examined systematically. The Box shows specific questions to be answered for ascertaining the validity of results in a study of diagnosis. One particular point to note is that, some studies only apply the 'gold standard' diagnostic procedure or confirmatory test to those with a positive result in the 'screening' test. In such circumstances, it would not be able to calculate the sensitivity or specificity, as the true and false negatives remain unknown. Only the predictive value of a positive test result can be obtained, but its application would be very limited (see below).

BOX. Validity of study results

Validity — selection bias
- Was the source of study subjects described and was a representative sample selected?
- Did the study sample include an appropriate spectrum of patients to whom the test is intended to be used in clinical practice?
- Was the response/participation rate for the sampled subjects reported, and was it reasonably high?

Validity — measurement/ information bias
- How objective was the test result? Was subjective interpretation involved in reporting the test result (eg colour changes)?
- What was the 'gold standard' used and how objective was it?
- Was the 'gold standard' applied to all subjects independent of the test result?
- Was the reporting of test results blinded to the reporting of 'gold standard' disease status and vice versa?

Validity — confounding
- Were additional factors that might modify the study results allowed for?
- Was the test validated in a second independent group of subjects?

Others might also include a subsample of those with a negative outcome in the 'screening' test (usually the majority) when applying the 'gold standard' confirmatory test, in order to calculate the sensitivity and specificity. However, if the numbers of true and false negatives were not weighted for (multiplied by) the sampling fraction, the sensitivity would be overestimated and the specificity underestimated. Valid results could only be achieved if a true representative sample of those screened negative were tested.

(3) Are the study results reasonably reliable or precise?

The precision of the test accuracy (sensitivity, specificity, likelihood ratios, AUC of ROC curve) should be reported using confidence intervals, though this gets missed out not infrequently in studies on diagnosis. It is possible for a test with a reported high sensitivity and specificity to have poor performance in identifying subjects with the disease from those without, because the lower confidence limits of sensitivity and specificity could be lower than 50%.

(4) Can the study results be applied to a specific patient in another setting?

It is easy to understand that the background (eg age, gender, possible stage of disease) of the specific patient must fall within the spectrum of the subjects involved in the appraised study before considering application of the study results.

In applying a test to a specific patient, the major concern is the ability of the test result to predict the disease status. The indicators are the positive predictive value (PPV)—the proportion (or probability) of subjects really having the disease among those tested positive, and the negative predictive value (NPV)—the proportion of subjects really not having the disease among those tested negative. Both depend heavily on the pre-test probability or prevalence of the disease in a group of subjects similar to the patient at hand. The PPV of a test tends to be higher in subjects with high pre-test probability of having the disease. Hence, one must have an idea on how likely is the disease present in the specific patient with a certain background before a test is performed. The PPV and NPV reported in one study cannot be loosely applied to another setting, due to probable differences in pre-test probability.

A diagnostic test will guide clinical decision and management only when the post-test probability of a positive or negative test result reaches over or below certain thresholds. The upper threshold (for further interventions) and lower threshold (ruling out a diagnosis) to be adopted vary in different scenarios, depending on the nature and seriousness of the disease, the effectiveness and side-effects of the available treatments, as well as other medical and non-medical considerations that are outside the scope of the current discussion. If the administration of a test is not likely to bring the post-test probability across these thresholds, one should rethink whether the test (with resource and other implications) should be carried out. The ability of a diagnostic test to bring the post-test probability across these thresholds depends on the test's properties (sensitivity and specificity), as well as the pre-test probability. The relationship can be expressed as:

$$\text{Post-test odds} = \text{Pre-test odds} \times \text{Likelihood ratio}$$

$$\left(\mathit{Odds} = \frac{\mathit{Probability}}{1 - \mathit{Probability}}\right)$$

Ignatius TS Yu
Shelly LA Tse
Clinical Epidemiology Group
Hong Kong Medical Journal

References

1. Yu IT, Tse SL. Clinical Epidemiology Workshop—Introduction. Hong Kong Med J 2011;17:315-6.
2. Yu IT, Tse SL. Clinical Epidemiology Workshop 2—General approach to critical appraisal of a medical journal paper. Hong Kong Med J 2011;17:405-6.