Malaysian Family Physician 2009; Volume 4, Number 2&3
ISSN: 1985-207X (print), 1985-2274 (electronic)

Online version: http://www.e-mfp.org

Review Article

Evidence Based Medicine Series
Part 3. APPRAISING THE EVIDENCE
Are the results valid and clinically Important?

NM Lai *MRCP, MRCPCH* School of Medicine & Health Sciences, Monash University Malaysia.

Address for correspondence: Dr Lai Nai Ming, Senior Lecturer in Paediatrics, School of Medicine and Health Sciences, Monash University Malaysia, JKR 1235, Bukit Azah, 80100 Johor Bahru, Johor, Malaysia. Tel: +607-2190600, Fax : +607-2190601, Email: lainm123@yahoo.co.uk

Lai NM. Evidence-based Medicine Series. Part 3. Appraising the evidence. Are the results valid and clinically important? Malaysian Family Physician. 2009;4(2&3):57-62

INTRODUCTION

Many of us incorporate new research findings uncritically into our knowledge bank. We may be inclined to accept whatever the authors claim, especially on the subjects that arouse interest or conclusions that conform to our beliefs. Such practice, although easy, is at best risky, as bad research is rampant, even in reputable journals[1]. Our clinical practice might change according to these misleading findings, and so might others' who believe us, at the perils of our patients.

In critical appraisal, we examine a piece of clinical evidence systematically to determine its validity, clinical importance and applicability to our patient[2]. We look beyond the surface of the evidence, identify possible biases and decide for ourselves how much we should believe its messages, and whether and how we should use them in our clinical practice.

There are many online guides to critical appraisal, many of which are freely accessible[3,4]. The guides contain a very similar set of rules in critical appraisal, differing only in presentation to aid memory for different groups of users. Essentially, we ask three major questions in critical appraisal[2]:

1. Is this piece of research VALID?
2. Are the results CLINICALLY IMPORTANT?
3. Are the recommendations APPLICABLE to my patient?

When we assess validity, we focus on the methods. When we assess clinical importance, we focus on the results. When we assess applicability, we look through both methods and results. In this article, we cover issues on validity. The next two articles in this series deal with clinical importance and applicability respectively.

To appraise an article, we first identify the type of study that the article belongs to, i.e. whether it is an article about therapy, diagnosis, prognosis or a systematic review. This is covered elsewhere in the series. The appraisal differs between different types of study, although the same three areas of validity, clinical importance and applicability are assessed.

METHODS: GOOD ENOUGH FOR THE RESULTS TO BE VALID?

A valid piece of research gives the readers confidence that its findings are close to the truth. Various problems in the conduct of a research-called biases in EBM terms, may affect its validity. In critical appraisal, we essentially determine the "risk of bias" of a study. The lower is the risk of bias, the more likely are its findings to be valid. Following is a summary of the key criteria to note in assessing validity for different types of study. These criteria, adapted from the appraisal guides of the Centre for Evidence Based Medicine Oxford[5], are organised under the mnemonic headings of "RAM", which we hope some readers will find useful. The words represented by the letters "RAM" differ slightly for different types of study.

Key criteria on assessing validity or risk of bias

Studies on diagnosis (studies that assess the performance of a diagnostic test, i.e. the index test)

- **R**ecruitment: Was a representative sample of diseased population included?
- **R**eference standard: Was there an independent and blind comparison between the index test and a reference (gold) standard test?
- **A**ccountability: Did everyone have both the index test and the reference standard test (regardless of the index test result)?
- **M**easurement: Was the index test an objective test?

Studies on therapy (studies that assess the effects of an intervention, e.g. medicine or other treatment methods)

- **R**andomisation:
  - Was the assignment of patients randomized? Was the randomisation method well-described (including how random sequence was generated, whether there was sufficient safeguards to prevent the tampering of the sequence allocation (allocation concealment))

Malaysian Family Physician 2009; Volume 4, Number 2&3
ISSN: 1985-207X (print), 1985-2274 (electronic)

Online version: http://www.e-mfp.org

- Was there evidence that both groups were prognostically comparable at the start of the study
- Was there evidence that both groups would have been treated equally throughout the study, other than one receiving the intervention and the other comparison?

- **A**ccountability:
  - Were all subjects accounted for throughout the study?
  - Was each subject analysed in the group to which they were randomized, regardless of whether they received the intended allocation (i.e. was Intention-to-treat analysis followed)?
- **M**easurement: Was there blinding of investigators, care givers and data collector on the allocation of the subjects?

Studies on prognosis (studies that assess the association between a prognostic factor, like a disease or patient characteristics, and an outcome, like mortality)

- **R**ecruitment: Was there a well-defined and representative sample?
- **A**ccountability:
  - Was the follow-up period appropriate for the outcomes assessed?
  - Was the drop-out rate small enough?
- **M**easurement:
  - Was the outcome objectively measured?
  - Was there adjustments for other confounders that could have influenced the results?

Systematic review (a study that summarises other studies on the same research question, using reproducible methods of identifying, evaluating, selecting and analyzing the studies)

- **R**ecruitment and **A**ccountability:
  - Was there a description on how the studies were identified?
  - Was there a description on how the studies were included and excluded?
- **M**easurement:
  - Was there a meta-analysis pooling the results of all the included studies?
  - Were the results consistent from study to study?

Possible answers to these questions include "yes", "no" or "unclear". A study is considered to have good validity if the investigators take care of each issue above and tell the readers clearly that they have done so. Unclear or negative answer in any of the question above casts some doubts on the study, as it alerts one on the possibility of biases from different sources. Some of these biases can be major. For example, lack of randomisation in a study on therapy, and lack of an independent and blind comparison with a reference standard test in a study on a diagnostic test raise major concerns on its overall validity.

RESULTS: ARE THEY CLINICALLY IMPORTANT?

This part of the article covers at a basic level, concepts and terms that are essential in understanding the results of an article. For more in-depth understanding of these and other related concepts, readers are advised to refer to many good resources that are available either online or in print.[2, 6-9]

When we determine the internal validity of a study, we look mainly at its methods. After deciding that a piece of research has good enough validity, we next determine the clinically importance of its messages by looking at the results. If a study has many possible sources of bias in the methods, hence poor validity, it will be a waste of time looking at the results. On the other hand, a valid piece of research with clinically trivial results may at best generate some interest among clinicians and researchers. It will not find a place in the current day-to-day care of the patients.

How to assess clinical importance? In different types of study, there are different things to look out for, as listed below.

i) Studies (and systematic reviews) on therapy: Clinical relevance of the outcome, the magnitude of effects, and the precision of the effect estimates.
ii) Studies on diagnosis: The performance of the diagnostic test in question (including its sensitivity, specificity, predictive values and likelihood ratio).
iii) Studies on prognosis: Clinical relevance of the outcome, the precision of the estimates, and the comprehensiveness of the information relating to the outcome over time.

Following is a more detailed explanation of each issue according to the type of study.

STUDIES (OR SYSTEMATIC REVIEWS) ON THERAPY

We look at the study results in three ways:

a) The outcome itself: whether it is of major relevance to patient care
b) The magnitude of effects: whether large enough to be considered worthwhile in actual clinical practice
c) The precision of the effect estimates: whether the range is narrow enough for us to be confident on the magnitude of effects.

The first of these is straightforward. Let us explore the second and third in more depth.

The magnitude of effects: beyond the p value
Take the example of two studies on antihypertensive therapy, one testing Drug A, another Drug B, both against a currently popular drug, Drug C. We are interested in whether Drugs A

Malaysian Family Physician 2009; Volume 4, Number 2&3
ISSN: 1985-207X (print), 1985-2274 (electronic)

Online version: http://www.e-mfp.org

or B further reduces one-year cardiovascular mortality in hypertensive adults compared to Drug C. Imagine that both studies were equally well-conducted i.e. validity is not a concern. At the end, both studies show a significant difference in one-year cardiovascular mortality, with p values of less than 0.05. In the first study, patients receiving Drug A had lower mortality by an average of 15% compared to those receiving Drug C. In the second study, those receiving Drug B had lower mortality by an average of 1% compared to Drug C.

With p values of less than 0.05, both Drugs A and B produce statistically significant lower one-year cardiovascular mortality compared to Drug C. Does it mean that both Drugs A and B have clinically important effects and are worthwhile to be considered in practice?

To answer this question, we should think about what constitutes clinically important reduction in one-year cardiovascular mortality. Most will not dispute that 15% is a remarkable and important drop in mortality, as opposed to 1%. So, by considering the magnitude of the effects alone, we can claim that despite statistically significant reductions in mortality produced by both drugs, Drug A has more clinically important effect in one-year cardiovascular mortality compared to Drug B.

To determine the magnitude of effect of a therapy, we need to understand how the outcomes are expressed. In a study on therapy, there are two major types of outcome:

a) **Continuous**, like length of hospital stay or blood pressure. Example: in a study comparing treatment A with placebo, patients receiving treatment A had shorter hospital stay compared to those receiving placebo, **mean difference: 3.2 days** ……...
b) **Dichotomous**, like being dead or alive. Example: In a study comparing treatment A with placebo, patients receiving treatment A were less likely to die at one year compared to those receiving placebo, **Relative Risk (RR) of dying: 0.85** ……. In the example of antihypertensive studies above, we use a dichotomous outcome of one-year cardiovascular mortality, i.e. being dead from cardiovascular causes, or being alive at one year after the commencement of the studies.

A continuous outcome can be converted to a dichotomous outcome. For example, length of hospital stay: e" 3 days or < 3 days; and systolic blood pressure of e" 130 mmHg, or < 130 mmHg.

Interpreting the difference in a continuous outcome, for example, mean difference, is usually straightforward. For dichotomous outcomes, however, we need to be careful, because the same results can be expressed in different ways, making the intervention appear to have a large or small effect.

Take the example mentioned above under the dichotomous outcome. Treatment A reduces the risk of dying at one year compared to placebo, relative risk: 0.85. This means treatment A reduces the risk of dying by 15% compared to placebo (1.00 -0.85 = 0.15, which in percentage is 15%). This seems a good figure, a clinically important effect size. Is that really so?

The truth is, we are not sure. It depends on what the actual mortality figures are. A relative risk of dying of 0.85 may mean either a reduction of mortality rate from 100% to 85% (15% absolute risk difference, which is clinically important), from 10% to 8.5% (1.5% absolute risk difference, which may or may not be clinically important), or from 1% to 0.85% (0.15% absolute risk difference, which is hardly considered clinically important). The possible confusion arises because relative risk is an "adjusted" figure. It is the absolute risk difference divided by the risk of the control group. From the three examples above, (100-85)/100, (10-8.5)/10 and (1-0.85)/1 all give us the same relative risk of 0.85. This commonly used expression of relative risk, on its own, gives us no idea on the absolute figures (**Absolute Risk Difference** in EBM terms), which are necessary to decide whether the results are clinically important.

How large should a treatment effect be to be considered as clinically important? There is no fixed rule and no scientific formula - at least none so far. It depends on what the care providers perceive as a worthwhile benefit, taking into account the severity of the condition, the harms associated with the treatment, and the known effects of the currently available alternatives. These in turn depend on our experience and judgment. Different thresholds for clinical importance are therefore possible, provided that they are well-justified.

The precision of the estimates

Let us take the example of a study comparing a new antihypertensive, X with a currently used antihypertensive, Y. The study shows that at the end, patients on X had on average lower systolic blood pressure by 15mmHg, compared to Y. The figure of 15 mmHg is the mean reduction obtained in this study.

In truth, the authors are not absolutely sure that 15mmHg is the exact mean for the population studied, because the study, like any other study, has a finite number of participants, which the authors hope represents the entire population of patients for whom this new drug is intended. With a limited sample, there is always a degree of uncertainty on the estimates. The authors normally give a range of values on top of the point estimate to address the uncertainty. This range of values is usually expressed as 95% Confidence Interval (95% CI), meaning that the authors are "95% sure" that the true mean lies within the range. By looking at the range of the estimates, we will know how confident we should be on the effect size. A narrow range gives us more confidence than a wide range.

Malaysian Family Physician 2009; Volume 4, Number 2&3
ISSN: 1985-207X (print), 1985-2274 (electronic)

Online version: http://www.e-mfp.org

From the example above, we may see the results being reported more or less like this: mean reduction in blood pressure of 15 mmHg (95% Confidence Interval: 13-18 mmHg). By reporting as such, the authors tell us that "In this study, we found a mean reduction in blood pressure of 15 mmHg, and we are 95% sure that the true mean reduction lies between 13 and 18mmHg".

Looking at the data, we are confident that the true mean reduction in blood pressure is around a narrow range of 13-18mmHg. If instead the 95% CI is 2 to 35mmHg, we are much less confident about the estimate, as the true mean value might be as small as 2mmHg, or as large as 35mmHg. This explains why precision of the estimates is an important, although indirect measure of the clinical importance of a study.

STUDIES ON DIAGNOSIS

A study examining a diagnostic tool typically compares a new tool against an established, well-accepted tool for the same condition –commonly called the reference standard. We look at the performance of the tool in question against the reference standard. The better the tool performs, the more clinically important is the results of the study.

There are several measures commonly used to indicate the performance of a diagnostic tool. These include sensitivity, specificity, positive and negative predictive values and likelihood ratios. The multitude of terms may be confusing, but they are necessary, and so are continued studies on new diagnostic tools, because there is no single tool that can accurately identify all diseased persons as diseased, and all healthy persons as healthy. There are always some "errors" associated with any diagnostic tool. There are two major sides of the "errors": first, mistakenly showing a diseased person as not diseased (missing the diagnosis), and second, mistakenly showing a disease-free person as diseased (giving false alarm). The lower is the probability of these "errors", the better is the diagnostic tool.

To understand the concepts of sensitivity, specificity, predictive values and likelihood ratios, it is best to orientate ourselves with a common two-by-two table, as follows:

Table 1. Two-by-two table depicting the states of disease and test results, commonly used to illustrate concepts related to the performance of a diagnostic test

	Diseased	Not diseased
Test result positive	a	c
Test result negative	b	d

In this table, the "test results" refer to the results of the new diagnostic tool being assessed. The terms "diseased" and "not diseased" refer to the disease status of the participants, as determined by the results of the current reference standard.

Let us take the table apart as we go through each term.

Sensitivity and specificity
To understand sensitivity, we need only to look at the left column of the table, as follows:

Table 2. The "left-half" of table 1, used for illustrating the concept of sensitivity

	Diseased
Test result positive	a
Test result negative	b

Here, the total number of patients who actually have the disease is represented by the whole column, the sum of a and b. The upper cell, a, is the number of patients with the disease, in whom the test shows up ("correctly") as positive. The lower cell, b, is the number of patients with the disease, in whom the test shows up ("incorrectly") as negative, i.e. the patients missed.

Sensitivity is represented by the formula: a/a+b. In words, it is the percentage of people with the disease that are correctly identified by the test. We can see that when dealing with the concept of sensitivity, we only talk about patients with disease. The higher is the sensitivity, the smaller is b, which means the fewer patients, missed by the test.

Next, we go through the concept of specificity by looking at the other half of the table.

Table 3. The "right-half" of Table 1, used for illustrating the concept of specificity

	Not diseased
Test result positive	c
Test result negative	d

Here, we can see that we only deal with patients without the disease. The total number of patients without the disease is represented by the whole column, the sum of c and d. The upper cell, c, is the number of patients without the disease, in whom the test shows up ("incorrectly") as positive. The lower cell, d, is the number of patients without the disease, in whom the test shows up ("correctly") as negative.

Malaysian Family Physician 2009; Volume 4, Number 2&3
ISSN: 1985-207X (print), 1985-2274 (electronic)

Online version: http://www.e-mfp.org

Specificity is represented by the formula: d/c+d. In words, it is the percentage of people without the disease that are correctly identified by the test. The higher is the specificity, the smaller is c, which means fewer disease-free persons would be tested positive, and get a false alarm.

The concept of sensitivity and specificity is useful when we assess patients at the bedside, thinking through their likelihood of having the disease from our clinical assessment, and pondering on the added value of a diagnostic test in helping us make our diagnosis. A diagnostic tool with high sensitivity is necessary in primary setting where screening takes place, and where we refer patients who are suspected to have the disease for further work-up. On the other hand, a tool with high specificity is preferable in tertiary centres where patients have already undergone some form of prior assessments, to avoid unduly labeling healthy people as diseased and subject them to unnecessary anxiety, and even treatment. In general, for both sensitivity and specificity, figures above 80% are taken as acceptable, and above 90% as excellent.

Predictive values

Once a diagnostic test is ordered and the results available, the concepts of sensitivity and specificity are no longer directly applicable. Instead, we look to see how accurate the available results of the test are in predicting the likelihood of having the disease. These measures are called predictive values. There are two types of predictive value, namely positive and negative predictive values. Let us explore each of these terms using again, parts of Table 1.

Table 4. The "top-half" of Table 1, used for illustrating the concept of positive predictive value

	Diseased	Not diseased
Test result positive	a	c

In dealing with sensitivity and specificity, we look down the columns. In dealing with predictive values, we look across the rows. Shown here is the upper row of the table, where everyone has a positive test result, while some truly have the disease (a), and some do not have the disease (c).

Positive predictive value of a test, represented by the formula a/a+c, is the proportion of those with positive test result who truly have the disease.

Table 5. The "bottom-half" of Table 1, used for illustrating the concept of negative predictive value

	Diseased	Not diseased
Test result negative	b	d

Similarly, looking at Table 5, negative predictive value, represented by the formula d/b+d, is the proportion of those with negative test result who truly do not have the disease.

Likelihood ratios

Compared to earlier terms, likelihood ratios are less intuitive, but probably more complete measures of the performance of a diagnostic test. From the notes above, one may notice that sensitivity, specificity and predictive values, if considered alone, tell us only part of the story about a diagnostic test (only parts of the 2x2 table is used when illustrating each of these terms). In likelihood ratios, we combine both sensitivity and specificity into a single measure.

There are two types of likelihood ratio – the likelihood ratio for a positive result (**LR+**) and the likelihood ration for a negative result (**LR -**).

LR+ is represented by the formula: **Sensitivity/(1-specificity).** It tells us how much the **odds** of the disease **increases** when the test is **positive**.

LR- is represented by the formula: **(1-sensitivity)/specificity**. It tells us how much the **odds** of the disease **decreases** when the test is **negative**.

The **odds** of a disease is the ratio between the number who have the disease over the number who do not have the disease. This is different from the usual terms of "risk" or "probability", in which the denominator is the total number of patients. Although less intuitive, measures involving odds are widely used for different reasons, the details of which are beyond the scope of this article.

STUDIES ON PROGNOSIS

In a study on prognosis, the authors look at the association between a factor, called the prognostic indicator, and the outcome. For example, patients with "stage 2BX, high grade" Non-Hodgkins Lymphoma (NHL) (prognostic indicator) and five-year disease-free survival (outcome). Among the three major elements that determine the clinical importance of such studies, clinical relevance of the outcome is straightforward, and the precision of the estimates is covered in the previous

Malaysian Family Physician 2009; Volume 4, Number 2&3
ISSN: 1985-207X (print), 1985-2274 (electronic)

Online version: http://www.e-mfp.org

section on therapy. We shall now focus on the remaining element: the comprehensiveness of the information relating to the outcome over time.

Take the example above on NHL and survival. Let's say the study finds that the five-year disease-free survival rate for patients with such specific staging and grading of NHL is 50%. We know that at the end of the five-year period 50% of the patients survive. That is all we know. But surely, we and our patients would be interested to find out what happens along the five years. We may ask: among the 50% who did not survive, was there a steady increase in mortality over the years of follow-up? When did most of this 50% of patients die? Between the fourth and fifth year, or during the first year?.... These are crucial information for our patients and their family, who may well be planning their future.

A study on prognosis provides clinically useful and important results by providing us the information on the likelihood of the outcome over time. This is often best demonstrated using graphs called the survival curves (see Figure 1). The curves

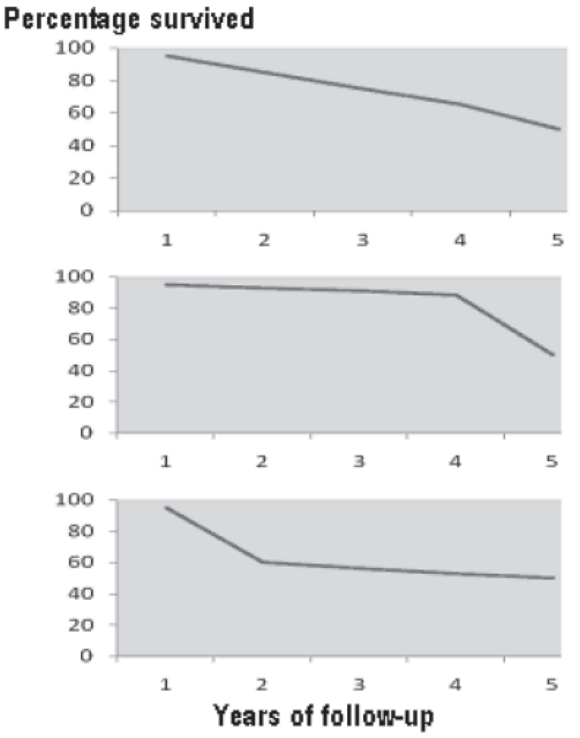

Figure 1. Three survival curves, each showing a five-year survival rate of 50% but different trends of survival over the five years

are commonly used to depict all outcomes over time in general, not only survival.

The three graphs contained in Figure 1 all end up with 50% survival rate at the end of the five-year period, but each has different trend of survival over time. The top graph shows a steady decrease in survival rate over the five years, the middle shows a minimal decrease in the first four years followed by a sharp drop between the fourth and the fifth year, while the bottom graph shows a sharp decrease over the first year and a slow drop thereafter. It should be self-evident that such information is critical in counselling patients about their prognosis.

Summary

It may be obvious to the readers now that critical appraisal, especially evaluating the clinical importance of evidence, is not a robotic process. Apart from understanding the various means of expressing a study result, it takes into account factors other than the evidence itself, like the clinical circumstances and the individual judgment of the clinicians who make decisions on whether the treatment, diagnostic tool or prognostic information is worth using in practice.

REFERENCES

1. Altman DG. The scandal of poor medical research. *BMJ*. 1994;308(6924):283-4
2. Straus SE, Richardson WS, Glasziou P, Haynes RB. Evidence-based medicine. How to practice and teach EBM. 3rd ed.: Edinburgh: Churchill Livingstone; 2005.
3. Teaching EBM: Teaching materials: Critical appraisal worksheets. Toronto: Centre for Evidence Based Medicine: University Health Network. Available from: http://www.cebm.utoronto.ca/teach/materials/caworksheets.htm
4. Appraising. Available from: http://www.shef.ac.uk/scharr/ir/netting/
5. Critical appraisal. Available from: http://www.cebm.net/index.aspx?o=1157.
6. Evidence-Based Medicine for Primary Care and Internal Medicine. BMJ Publishing Group; 2009. Available from: http://ebm.bmj.com
7. Glasziou P, Del Mar C, Salisbury J. Evidence-based practice workbook: bridging the gap between health care research and practice. Wiley-Blackwell; 2007
8. Dans A, Dans L, Silvestre M. Painless Evidence-Based Medicine. Chichester: John Wiley and Sons Ltd; 2007
9. Edwards A, Elwyn G, Mulley A. Explaining risks: turning numerical data into meaningful pictures. *BMJ*. 2002;324(7341):827- 30