TRUST UPON EVIDENCE

By

Dr. Syeda Uroosa Hashmi

- In this chapter we consider how to decide if a study is of sufficient quality that its findings are likely to be valid.
- We begin with a general discussion of approaches to appraising validity and then describe specific methods for appraising validity of studies of the effects of interventions, experiences, prognosis and the accuracy of diagnostic tests.

- Many physiotherapists experience a common frustration.

- When they consult the research literature for answers to clinical questions they are confronted by a range of studies with very different conclusions.

- Consider, for example, the findings that confront a physiotherapist who would like to know whether acupuncture protects against exercise induced asthma.

One study, by Fung et al (1986) concluded ‘acupuncture provided better protection against exercise-induced asthma On the other hand, Gruber et al (2002) concluded ‘acupuncture treatment offers no protection against exercise-induced bronchoconstriction

Why is the literature apparently so inconsistent? There are several possible explanations.

First, there may be important differences between studies in the type of patients included, the way in which the intervention was administered, and the way in which outcomes were measured.

Simple conclusions may obscure important details about patients, interventions and outcomes. However, as we shall see later, it may be difficult to draw more precise conclusions from clinical research.

It seems implausible that both could be true. Situations like this, where similar studies draw contradictory conclusions, often arise.

CRITICAL APPRAISAL OF EVIDENCE ABOUT THE EFFECTS OF INTERVENTION

- In Chapter 3 it was argued that the preferred source of evidence of the effects of a therapy is usually a recent systematic review.

- But for some questions there are no relevant, recent systematic reviews, in which case it becomes necessary to consult individual randomized trials.

- We first consider how to assess the validity of randomized trials even though the reader is encouraged to look first for systematic reviews,

- because it is easier to understand critical appraisal of systematic reviews

RANDOMIZED TRIALS

- Readers of clinical trials can ask three questions to discriminate coarsely between those trials that are likely to be valid and those that are potentially seriously biased.

Were treated and control groups comparable?

- It is essential that the groups are comparable, and comparability can only be assured by randomly assigning subjects to groups.

- ‘Matching’ of subjects in the treatment and control groups cannot, on its own, ensure that the groups are comparable, regardless of how diligently the matching is carried out.

- Subjects may be allocated to groups on the basis of their birth dates (for example, subjects with even-numbered birth dates could be assigned to the treatment group and subjects with odd-numbered birth dates assigned to the control group), or medical record numbers, or

- the date of entry into the trial.

- It is likely that, if carried out carefully, all of these procedures could assign subjects to groups in a way that is effectively random in the sense that all the procedures could generate comparable groups.

- Most clinical trials involve interventions that are implemented over days or weeks or months. Usually outcomes are assessed at the end of the intervention, and they are often also assessed at one or several times after the intervention has ceased.

- Trials of chronic conditions may assess outcomes several years after the intervention period has ceased.

- A problem that arises in most trials is that it is not always possible to obtain outcome measures as planned. Occasionally subjects die.

- Others become too sick to measure, or they move out of town, or go on long holidays.

- Some may lose interest in participating in the study or simply be too busy to attend for follow-up appointments. For these and a myriad of other reasons it may be impossible for the researchers to obtain outcome measures from all subjects as planned, no matter how hard the researchers try to obtain follow-up measures from all patients.
- Randomization is undone. Estimates of the effect of treatment become contaminated by differences between groups due to loss to follow-up.

intervention and chance. Randomization is undone. Estimates of the effect of treatment become contaminated by differences between groups due to loss to follow-up.

It is quite plausible that dropouts from one group will differ systematically from dropouts in the other group. This is because it is quite plausible that subjects' experiences of the intervention or its outcomes will influence whether they attend for follow-up.[10] Imagine a hypothetical trial of treatment for cervical headache. The trial compares the effect of six sessions of manual therapy to a no-intervention control condition, and outcomes in both groups are assessed 2 weeks after randomization. Some subjects in the control group may experience little resolution of their symptoms. Understandably, these subjects may become dissatisfied with participation in the trial and may be reluctant to return for outcome assessment after not having received any intervention. The consequence is that there may be a tendency for those subjects in the control group with the worst outcomes to be lost to follow-up, more so than in the treated group. In that case, estimates of the effects of intervention (the difference between the outcomes of treated and control groups) are likely to be biased and the treatment will appear less effective than it really is.

We could imagine many such scenarios that would illustrate that loss to follow-up can bias estimates of the effects of intervention in either direction. Unfortunately, while statistical techniques have been formulated to

- **Studies which do not provide data on loss to follow-up and which do not explicitly state that there was no loss to follow-up should be considered potentially biased.**

- A problem that is closely related to loss to follow-up is the problem of protocol violation.

- Protocol violations occur when the trial is not carried out as planned. In trials of physiotherapy interventions, the most common protocol violation is the failure of subjects to receive the intended intervention.

- For example, subjects in a trial of exercise may be allocated to an exercise group but may fail to do their exercises, or fail to exercise according to the protocol (this is sometimes called ‘non-compliance’ or ‘non-adherence’), or subjects allocated to the control condition may take up exercise.

Was there blinding to allocation of patients and assessors?

- There is reason to prefer that, in clinical trials, subjects are unaware of whether they received the intervention or control condition.

- This is called blinding of subjects. Blinding of subjects is considered important because it provides a means of controlling for placebo effects.

Placebo effects

- Placebo effects are effects of intervention attributable to patients' expectations of a beneficial effect of therapy.
- The placebo effect is demonstrated when patients benefit from interventions that could have no direct physiological effects, such as detuned ultrasound.

- **Blinding of subjects ensures that estimates of the effects of intervention (the difference between outcomes of treated and control groups) cannot be due to placebo effects**

groups will be, on average, similar. Consequently,

Blinding of subjects ensures that estimates of the effects of intervention (the difference between outcomes of treated and control groups) cannot be due to placebo effects.

How is it possible to blind patients to allocation? How can subjects not know if they received the intervention or control? The general approach involves giving a 'sham' intervention to the control group. Sham interventions are those that look, feel, sound, smell and taste like the intervention but could not effect the presumed mechanism of the intervention. The clearest examples in physiotherapy come from studies of electrotherapies. Several clinical trials (for example, McLachlan et al 1991, Ebenbichler et al 1999, van der Heijden et al 1999) have used sham interventions in studies of pulsed ultrasound. In these studies the ultrasound machine is adapted so that it either emits pulsed ultrasound (the intervention) or does not (the sham intervention). In the study by McLachlan et al (1991), the sham ultrasound transducer was designed to become warm when turned on, so the patient was unable to distinguish between intervention and sham. The intervention and sham could not be distinguished by the patient, and yet the sham could not effect the presumed mechanisms of ultrasound therapy because no ultrasound was emitted. Consequently this is a near-perfect sham. Other near-perfect shams used in clinical trials of physiotherapy interventions include the use of coloured light as sham low-level laser therapy (for example, de Bie et al 1998), and the use of specially constructed collapsing needles in studies of acupuncture (Kleinhenz et al 1999).

SYSTEMATIC REVIEWS OF RANDOMIZED TRIALS

- If a systematic review is to produce valid conclusions it must identify most of the relevant studies that exist and produce a balanced synthesis of their findings.
- To determine if this goal has been achieved, readers can ask three questions.

Was it clear which trials were to be reviewed?

- When we read systematic reviews we need to be satisfied that the reviewer has not selectively reviewed those trials which support his or her own point of view.
- An example of a systematic review which provides clear inclusion and exclusion criteria is the review by Green et al (1998) of interventions for shoulder pain.
- In their review the authors indicated that they ‘identified trials independently according to predetermined criteria (that the trial be randomized, that the outcome assessment be blinded, and that the intervention was one of those under review).
- Randomized controlled trials which investigated common interventions for shoulder pain in adults (age greater than or equal to 18 years) were included provided that there was a blinded assessment of outcome.’
- Systematic reviews which specify clear inclusion and exclusion criteria provide stronger evidence of effects of therapy than those that do not.

Were most relevant studies reviewed?

- Well-conducted reviews identify most trials relevant to the review question.
- There are two reasons why it is important that reviews identify most
- relevant trials. First, if it may conclude that there is less evidence than there really is.

Was the quality of the reviewed studies taken into account?

- Many randomized trials are poorly designed and provide potentially seriously biased estimates of the effects of intervention.
- Consequently, if a systematic review is to obtain an unbiased estimate of the effects of intervention, it must ignore low quality studies.

Appraisal about experience
Was the sampling strategy appropriate?

- In qualitative research we are not interested in an ‘on average’ view of a population. We want to gain an in-depth understanding of the experience of particular individuals or groups. The characteristics of individual study participants are therefore of particular interest.

- The author should explain and justify why the participants in the study were the most appropriate to provide access to the type of knowledge sought by the study. If there have been any problems with recruitment (for example, if there were many people that were invited to participate but chose not to take part), this should be reported.

- People may be selected because they are typical or atypical, because they have some important relationship, or just because they are the most available subjects.

- Sometimes sampling occurs in an opportunistic way: one person leads the researcher to another person, and that person to one more, and so on. This is called snowball sampling (Seers 1999).

- Often the goal of sampling is to obtain as many perspectives as possible.

Was the method used to collect data relevant?

- A range of very different methods is used to collect data in qualitative research. These vary from, for example, participant observations, to in-depth interviews, to focus groups, to document analysis.
- The data collection method should be relevant and address the questions raised, and should be justified in the research report.
- A common method in physiotherapy research involves the use of observations or in-depth interviews to explore communication and interactions of physiotherapists and patients.
- In-depth interviews are also used to explore experiences, meanings, attitudes, views and beliefs, for example the experiences of being a patient, or of having a certain condition, as in a study that explored stroke patients' motivation for rehabilitation

- Another important question to ask about data collection is whether ethical issues have been taken into consideration.
- **Were the data analysed in a rigorous way?**
- The process of analysis in qualitative research should be rigorous.
- This is a challenging, complex and time-consuming job.
- The aim of this process is often to make sense of an enormous amount of text, tape recordings or video materials by reducing, summarizing and interpreting the data.
- The researchers often extend their conceptual frameworks into themes, patterns, hypotheses or theories; but ultimately they must communicate what their data mean.
- An in-depth description of the decision trail gives the reader a chance to follow the interpretations that have been made and to assess these interpretations in the light of the data.
- An indication of a rigorous analysis is that the data are presented in a way that is clearly separated from the interpretation of the data.
- There should be sufficient data (e.g. transcripts) to justify the interpretation.

CRITICAL APPRAISAL OF EVIDENCE ABOUT PROGNOSIS

- we considered two sorts of questions about prognosis: questions about what a person's outcome will be, and questions about how much we should modify our estimates of prognosis on the basis of particular prognostic characteristics.

Was there representative sampling from a well-defined population?

prognosis and then consider, very briefly, systematic reviews of prognosis.

INDIVIDUAL STUDIES OF PROGNOSIS

Was there representative sampling from a well-defined population?

If we are to derive useful information about prognoses from clinical research, we must be able to use the findings of the research to make inferences about prognoses of some larger population. We can only do this if the subjects participating in the research (the 'sample') are representative of the population we are interested in.

When we read studies of prognosis we first need to know which population the study is seeking to provide a prognosis for (the 'target population'). The target population is defined by the criteria used to determine who was eligible to participate in the study. Most studies of prognosis describe a list of inclusion and exclusion criteria that clearly identify the target population. For example, Coste et al (1994) conducted an inception cohort study of the prognosis of people presenting for primary medical care for acute low back pain. They stated that 'all consecutive patients aged 18 and over, self-referring to participating doctors ($n = 39$) for a primary complaint of back pain between 1 June and 7 November 1991

?

were eligible. Only patients with pain lasting less than 72 hours and without radiation below the gluteal fold were included. Patients with malignancies, infections, spondylarthropathies, vertebral fractures, neurological signs, and low back pain during the previous 3 months were excluded, as were non-French speaking and illiterate patients.′ The target population for this study is clear.

will provide a biased estimate of prognosis in the target population. When a study recruits 'all' subjects or 'consecutive cases' that satisfy inclusion criteria (as in the study by Coste, cited in the last paragraph) we can be relatively confident that the findings of the study apply to a defined population. The greater the proportion of eligible subjects that participates in the study, the more representative the sample is likely to be.

Researchers may find it difficult to gather data from consecutive cases, particularly when participation in the study requires extra measurements be made over and above those that would normally be made as part of routine clinical practice. An example of a study that did not sample in a representative way is a study of the 'outcomes' (prognosis) of children with developmental torticollis (Taylor & Norton 1997). The researchers sampled 'twenty-three children (14 male, nine female) ... diagnosed with developmental torticollis by a physician. ... Most of the children (74%) were referred to physical therapy by pediatricians ... Data were collected retrospectively from the initial physical therapy evaluations of the 23 children whose parents agreed to a follow-up evaluation.' Such samples may not always be representative; they may comprise subjects with particularly good or particularly bad prognoses. Consequently, samples of convenience can provide biased prognoses for the target population.

---

[38] There are two ways to claim representativeness. The first approach is to clearly define the population of interest and then sample from that population in a representative way, or in as representative a way as possible. The alternative approach is to sample in a non-representative way and then use the characteristics of the sample to dictate about whom inferences can be made. With the former approach, inferences can be made about the

- Studies in which all (or nearly all) eligible subjects enter the study are sometimes said to have sampled ‘consecutive cases’.

- **When you read studies looking for information about prognosis, start by looking to see whether the study recruited ‘all’ patients or ‘consecutive cases’.**

- **If it did not, the study may provide biased estimates of the true prognosis.**

Was there an inception cohort?

- A study of prognosis could sample from the whole population of people who currently have the condition of interest.
- But samples obtained from the whole population of people who currently have the condition (called ‘survivor cohorts’) will tend to consist largely of people who have had the condition for a long time, and that introduces a potential bias.
- The solution is to recruit subjects at a uniform (usually early) point in the course of the disease.
- Studies which recruit subjects in this way are said to recruit ‘inception cohorts’ because subjects were identified as closely as possible to the inception of the condition.
- The advantage of inception cohorts is that they are not exposed to the biases inherent in studies of survivor cohorts.

- **Studies that recruit inception cohorts may provide less biased estimates of prognosis than studies that recruit survivor cohorts.**

Was there complete or near-complete follow-up?

- prognostic studies can be biased by loss to follow-up. Bias occurs if those lost to follow-up have, on average, different outcomes to those who were followed up.

- Look for information about the proportion of subjects for whom follow-up data were available at key time points.
- Alternatively, calculate loss to follow-up from numbers of subjects entered into the study and the numbers followed up.

CRITICAL APPRAISAL OF EVIDENCE ABOUT DIAGNOSTIC TESTS

INDIVIDUAL STUDIES OF DIAGNOSTIC TESTS

Was there comparison with an adequate reference standard?

Interpretation of studies of diagnostic accuracy is most straightforward if the reference standard is perfectly accurate, or close to it. But it is difficult to know if the reference standard is accurate. Assessment of the accuracy of the reference standard would require comparing its findings with another reference standard, and we would then need to know *its* accuracy. So, realistically, we have to live with imperfect knowledge of the reference standard. Claims of the adequacy of a reference standard cannot be based on data. Instead they must rely on face validity. That is, ultimately our assessments of the adequacy of the reference standard must rely on our assessment of whether the reference standard appears to be the sort of measurement that would be more-or-less perfectly accurate.

An example of a reference standard that has apparent face validity is open surgical or arthroscopic confirmation of a complete tear of the anterior cruciate ligament. It is reasonable to believe that the diagnosis of a complete tear could be made unambiguously at surgery. On the other hand, the diagnosis of partial tears is more difficult, and the surgical presentation may be ambiguous. Thus open surgical exploration and arthroscopic examination are excellent reference standards for diagnosis of complete tears, but less satisfactory reference standards for partial tears.

When the reference standard is imperfect, the accuracy of the diagnostic test of interest will tend to be underestimated. This is because when the reference standard is imperfect we are asking the clinical test to do

Box 5.8 Assessing validity of individual studies of diagnostic tests

Was there comparison with an adequate reference standard?
Were the findings of the test compared with the findings of a reference standard that is considered to have near-perfect accuracy?

Was the comparison blind?
Were the clinicians who applied the clinical tests unaware of the findings of the reference standard?

Did the study sample consist of subjects in whom there was diagnostic uncertainty?
Was there sampling of consecutive cases satisfying clear inclusion and exclusion criteria?