Implementing NHMRC dimensions of evidence including new 'interim' levels of evidence

This part of the document outlines how individual studies included in a systematic literature review should be assessed using the NHMRC dimensions of evidence and provides levels of evidence appropriate for the most common types of research questions. The basic principles of systematic reviewing and assessing evidence are set out in the NHMRC handbook series on the development of clinical practice guidelines (NHMRC 2000ab).

Dimensions of evidence for assessing included studies

Each included study in a systematic review should be assessed according to the following three dimensions of evidence:

1. Strength of evidence

a. *Level of evidence*: Each study design is assessed according to its place in the research hierarchy. The hierarchy reflects the potential of each study included in the systematic review to adequately answer a particular research question, based on the probability that its design has minimised the impact of bias on the results. See page 6–10 of *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b).
The currently available NHMRC levels of evidence for intervention studies (NHMRC 2000b), together with the new levels of evidence for questions on diagnosis, prognosis, aetiology and screening are shown in the evidence hierarchy in Table 1.

b. *Quality of evidence* (risk of bias): The methodological quality of each included study is critically appraised. Each study is assessed according to the likelihood that bias, confounding and/or chance may have influenced its results. The NHMRC toolkit *How to review the evidence: systematic identification and review of the scientific literature* (NHMRC 2000a) lists examples of ways that methodological quality can be assessed. In cases where other critical appraisal approaches may be required, there are a number of alternatives. The NHMRC Guideline Assessment Register consultant can advise on the choice of an alternative to supplement and/or replace those in the NHMRC handbook (see Table 2).

c. *Statistical precision*: The primary outcomes of each included study are evaluated to determine whether the effect is real, rather than due to chance (using a level of significance expressed as a *P*-value and/or a confidence interval). See page 17 of *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b).

2. Size of effect

This dimension is useful for assessing the clinical importance of the findings of each study (and hence clinical impact). This is a different concept to statistical precision and specifically refers to the measure of effect or point estimate provided in the results of each study (eg mean difference, relative risk, odds ratio, hazard ratio, sensitivity, specificity). In the case of a meta-analysis it is the pooled measure of effect from the studies included in the systematic review (eg weighted mean difference, pooled relative risk). These point estimates are calculated in comparison to either doing nothing or versus an active control.

Size of the effect therefore refers to the distance of the point estimate from its null value for each outcome (or result) and the values included in the corresponding 95% confidence interval. For example, for a ratio such as a relative risk the null value is 1.0 and so a relative of risk of 5

is a large point estimate; for a mean difference the null value is zero (indicating no difference) and so a mean difference of 1.5kg may be small. The size of the effect indicates just how much clinical impact that particular factor or intervention will have on the patient and should always be taken in the context of what is a clinically relevant difference for the patient. The upper and lower point estimates in the confidence interval can then be used to judge whether it is likely that most of the time the intervention will have a clinically important impact, or that it is possible that in some instances the impact will be clinically unimportant or that there will be no impact. See pages 17–23 of *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b).

3. Relevance of evidence

This dimension deals with the translation of research evidence into clinical practice and is potentially the most subjective of the evidence assessments. There are two key questions.

a. *Appropriateness of the outcomes*: Are the outcomes measured in the study relevant to patients? This question focuses on the patient-centredness of the study. See pages 23–27 of *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b).
b. *Relevance of study question*: How closely do the elements of the research question ('PICO'[1]) match those of the clinical question being considered? This is important in determining the extent to which the study results are relevant (generalisable) for the population who will be the recipients of the clinical guideline.

The results of these assessments for each included study should be entered into a data extraction form described in the *NHMRC standards and procedures for externally developed guidelines* (NHMRC 2007). Once each included study is assessed according to these dimensions of evidence, a summary can be made that is relevant to the whole body of evidence, which can then be graded as described in Part B of this document. The data extraction process provides the evidence base on which the systematic review, and subsequent guideline recommendations are built.

[1] P=Population, I=Intervention/index test/indicator, C=Comparison, O=Outcome

Table 1 NHMRC Evidence Hierarchy: designations of 'levels of evidence' according to type of research question (including explanatory notes)

Level	Intervention [1]	Diagnostic accuracy [2]	Prognosis	Aetiology [3]	Screening Intervention
I [4]	A systematic review of level II studies	A systematic review of level II studies	A systematic review of level II studies	A systematic review of level II studies	A systematic review of level II studies
II	A randomised controlled trial	A study of test accuracy with: an independent, blinded comparison with a valid reference standard,[5] among consecutive persons with a defined clinical presentation[6]	A prospective cohort study[7]	A prospective cohort study	A randomised controlled trial
III-1	A pseudorandomised controlled trial (i.e. alternate allocation or some other method)	A study of test accuracy with: an independent, blinded comparison with a valid reference standard,[5] among non-consecutive persons with a defined clinical presentation[6]	All or none[8]	All or none[8]	A pseudorandomised controlled trial (i.e. alternate allocation or some other method)
III-2	A comparative study with concurrent controls: ▪ Non-randomised, experimental trial[9] ▪ Cohort study ▪ Case-control study ▪ Interrupted time series with a control group	A comparison with reference standard that does not meet the criteria required for Level II and III-1 evidence	Analysis of prognostic factors amongst persons in a single arm of a randomised controlled trial	A retrospective cohort study	A comparative study with concurrent controls: ▪ Non-randomised, experimental trial ▪ Cohort study ▪ Case-control study
III-3	A comparative study without concurrent controls: ▪ Historical control study ▪ Two or more single arm study[10] ▪ Interrupted time series without a parallel control group	Diagnostic case-control study[6]	A retrospective cohort study	A case-control study	A comparative study without concurrent controls: ▪ Historical control study ▪ Two or more single arm study
IV	Case series with either post-test or pre-test/post-test outcomes	Study of diagnostic yield (no reference standard)[11]	Case series, or cohort study of persons at different stages of disease	A cross-sectional study or case series	Case series

Explanatory notes

[1] Definitions of these study designs are provided on pages 7-8 *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b).

[2] The dimensions of evidence apply only to studies of diagnostic accuracy. To assess the effectiveness of a diagnostic test there also needs to be a consideration of the impact of the test on patient management and health outcomes (Medical Services Advisory Committee 2005, Sackett and Haynes 2002).

[3] If it is possible and/or ethical to determine a causal relationship using experimental evidence, then the 'Intervention' hierarchy of evidence should be utilised. If it is only possible and/or ethical to determine a causal relationship using observational evidence (ie. cannot allocate groups to a potential harmful exposure, such as nuclear radiation), then the 'Aetiology' hierarchy of evidence should be utilised.

[4] A systematic review will only be assigned a level of evidence as high as the studies it contains, excepting where those studies are of level II evidence. Systematic reviews of level II evidence provide more data than the individual studies and any meta-analyses will increase the precision of the overall results, reducing the likelihood that the results are affected by chance. Systematic reviews of lower level evidence present results of likely poor internal validity and thus are rated on the likelihood that the results have been affected by bias, rather than whether the systematic review itself is of good quality. Systematic review *quality* should be assessed separately. A systematic review should consist of at least two studies. In systematic reviews that include different study designs, the overall level of evidence should relate to each individual outcome/result, as different studies (and study designs) might contribute to each different outcome.

[5] The validity of the reference standard should be determined in the context of the disease under review. Criteria for determining the validity of the reference standard should be pre-specified. This can include the choice of the reference standard(s) and its timing in relation to the index test. The validity of the reference standard can be determined through quality appraisal of the study (Whiting et al 2003).

[6] Well-designed population based case-control studies (eg. population based screening studies where test accuracy is assessed on all cases, with a random sample of controls) do capture a population with a representative spectrum of disease and thus fulfil the requirements for a valid assembly of patients. However, in some cases the population assembled is not representative of the use of the test in practice. In diagnostic case-control studies a selected sample of patients already known to have the disease are compared with a separate group of normal/healthy people known to be free of the disease. In this situation patients with borderline or mild expressions of the disease, and conditions mimicking the disease are excluded, which can lead to exaggeration of both sensitivity and specificity. This is called spectrum bias or spectrum effect because the spectrum of study participants will not be representative of patients seen in practice (Mulherin and Miller 2002).

[7] At study inception the cohort is either non-diseased or all at the same stage of the disease. A randomised controlled trial with persons either non-diseased or at the same stage of the disease in *both* arms of the trial would also meet the criterion for this level of evidence.

[8] All or none of the people with the risk factor(s) experience the outcome; and the data arises from an unselected or representative case series which provides an unbiased representation of the prognostic effect. For example, no smallpox develops in the absence of the specific virus; and clear proof of the causal link has come from the disappearance of small pox after large-scale vaccination.

[9] This also includes controlled before-and-after (pre-test/post-test) studies, as well as adjusted indirect comparisons (ie. utilise A vs B and B vs C, to determine A vs C with statistical adjustment for B).

[10] Comparing single arm studies ie. case series from two studies. This would also include unadjusted indirect comparisons (ie. utilise A vs B and B vs C, to determine A vs C but where there is no statistical adjustment for B).

[11] Studies of diagnostic yield provide the yield of diagnosed patients, as determined by an index test, without confirmation of the accuracy of this diagnosis by a reference standard. These may be the only alternative when there is no reliable reference standard.

Note A: Assessment of comparative harms/safety should occur according to the hierarchy presented for each of the research questions, with the proviso that this assessment occurs within the context of the topic being assessed. Some harms are rare and cannot feasibly be captured within randomised controlled trials; physical harms and psychological harms may need to be addressed by different study designs; harms from diagnostic testing include the likelihood of false positive and false negative results; harms from screening include the likelihood of false alarm and false reassurance results.

Note B: When a level of evidence is attributed in the text of a document, it should also be framed according to its corresponding research question eg. level II intervention evidence; level IV diagnostic evidence; level III-2 prognostic evidence.

Source: Hierarchies adapted and modified from: NHMRC 1999; Bandolier 1999; Lijmer et al. 1999; Phillips et al. 2001.