Australian Government
National Health and
Medical Research Council
NHMRC
WORKING TO BUILD A HEALTHY AUSTRALIA
www.nhmrc.gov.au

NHMRC additional levels of evidence and grades for recommendations for developers of guidelines

STAGE 2 CONSULTATION

Early 2008 – end June 2009

Introduction

The National Health and Medical Research Council (NHMRC) in Australia has, over recent years, developed a suite of handbooks to support organisations involved in the development of evidence-based clinical practice guidelines (www.nhmrc.gov.au/publications/synopses/cp65syn.htm).

Reflecting the general impetus of the past decade, these handbooks focus predominantly on assessing the clinical evidence for interventions. As a consequence, the handbooks present 'levels of evidence' appropriate mainly for intervention studies. However, feedback from guideline developers received by the NHMRC has indicated that the levels of evidence used by the NHMRC for intervention studies have been found to be restrictive. This is particularly so where the areas of study do not lend themselves to research designs appropriate to intervention studies (i.e. randomised controlled trials).

As an interim measure to a review of the handbooks, this paper presents a forward-thinking approach to classifying levels of evidence, and grading evidence recommendations, which should be relevant to any clinical guideline (not just those dealing with interventions).

The NHMRC is committed to addressing this issue and any other relevant concerns when all of the handbooks are reviewed.

This consultation draft has been developed based on a Pilot Program on 'NHMRC additional levels of evidence and grades for recommendations for developers of guidelines', which was initially released for public consultation in 2005, until mid-2006 with feedback sought until 30 June 2007 on their usability and applicability.

Levels of evidence
Guidelines can have different purposes, dealing with clinical questions such as intervention, diagnosis, prognosis, aetiology and screening. To address these clinical questions adequately, guideline developers need to include different research designs. This consequently requires different evidence hierarchies that recognise the importance of research designs relevant to the purpose of the guideline. A new evidence hierarchy was subsequently developed in 2005 by the NHMRC Guideline Assessment Register (GAR) consultants. This hierarchy assigns levels of evidence according to the

type of research question, recognising the importance of appropriate research design to that question. As well as the current NHMRC levels of evidence for interventions, new levels have been developed for studies relevant for guidelines on diagnosis, prognosis, aetiology and screening.

This consultation framework outlines the expanded levels of evidence, and provides additional information in the form of table notes, a study design glossary and a summary of how the levels of evidence and other NHMRC dimensions of evidence should be used (see Part A).

Grades of recommendations

To further assist guideline developers to make judgments on the basis of the body of evidence relevant to a research question, a grading system for recommendations has been developed (see Part B). This takes the form of an evidence matrix, which lists the evidence components that should be considered when judging the body of evidence. The grade of a recommendation is based on an overall assessment of the rating of individual components of the evidence matrix.

Feedback

A draft of the revised levels of evidence and new grading system for recommendations was posted on the NHMRC website from 2005 until mid-2007, with feedback being sought internationally on their usability and applicability. Several guideline development teams, with guidance from their GAR consultant, tested the revised levels and grades of evidence in guidelines that were developed during the pilot period. The website feedback and the practical experience of guideline developers support the clinical utility and academic rigour of the new levels of evidence, and the grades of recommendation.

This revised levels of evidence and grading system for recommendations which has been amended based on submissions received from the first pilot consultation, is now released for a second stage consultation from early 2008 to 30 June 2009. Submissions on the Stage 2 consultation draft can be submitted to the NHMRC at: nhc@nhmrc.gov.au, with the heading: Att Project Officer, Levels & Grades public consultation Stage 2.

NHMRC or other guidelines that are developed using this consultation framework must include a statement at the front of the document explaining that the guidelines were developed using this consultation framework, which blends the official NHMRC levels with the 'interim' levels of evidence and grading system for recommendations.

Authors

Kristina Coleman, Sarah Norris, Adele Weston - Health Technology Analysts Pty Ltd
Karen Grimmer-Somers, Susan Hillier - Division of Health Sciences, University of South Australia
Tracy Merlin - Adelaide Health Technology Assessment (AHTA), Discipline of Public Health, University of Adelaide
Philippa Middleton, Rebecca Tooher - ASERNIP-S
Janet Salisbury - Biotext

Acknowledgments:

(a) Feedback has been provided during this document's development phase from the following:
Marita Broadstock – New Zealand Health Technology Assessment
Suzanne Dyer – NHMRC Clinical Trials Centre
Paul Glasziou – Oxford University, United Kingdom
Brian Haynes – McMaster University, Canada
Paul Ireland – National Institute of Clinical Studies

Nicki Jackson –Deakin University
Sally Lord and Les Irwig – University of Sydney
Skye Newton and Janet Hiller – University of Adelaide
Andrew Oxman – Oslo, Norway (GRADE Working Group)

(b) Significant contribution to the revision of the pilot for the Stage 2 Consultation was provided during July-December 2007 by members of the NHMRC Guideline Assessment Register (GAR) panel organisations:

• Division of Health Sciences, University of South Australia 1) Quality Use of Medicines and Pharmacy Research Centre - Dr Agnes Vitry - Ms Simone Rossi 2) Centre for Allied Health Evidence - Prof Karen Grimmer-Somers - Dr Susan Hillier - Dr Caroline Smith - Dr Saravana Kumar - Dr Nicola Massy-Westropp - Mr Peter Lekkas	• Faculty of Health Sciences, The University of Adelaide: 1) Australian Research Centre for Health of Women and Babies (ARCH), Discipline of Obstetrics & Gynaecology, School of Paediatrics and Reproductive Health - Prof Caroline Crowther - Ms Philippa Middleton - Dr Rebecca Tooher 2) Adelaide Health Technology Assessment (AHTA), Discipline of Public Health, School of Population Health and Clinical Practice - Prof Janet Hiller - Ms Tracy Merlin - Dr Peng Bi - Ms Skye Newton
• Health Technology Analysts Pty Ltd - Dr Adèle Weston - Dr Sarah Norris - Dr Kristina Coleman	• Biotext Pty Ltd - Dr Janet Salisbury - Dr Hilary Cadman - Dr Fiona Mackinnon

The work on this project is being managed by the Evidence Translation Section, and supported by National Institute of Clinical Studies Officers of the NHMRC.

PART A

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.

Table 2 Assessment of individual study quality

Study type	Location of NHMRC checklist[1]	Additional/supplemental quality assessment tool
Intervention	Page 45	
Diagnosis	Page 62	QUADAS (Whiting et al., 2003)
Prognosis	Page 81	GATE checklist for prognostic studies (NZGG, 2001)
Aetiology	Page 73	
Screening	Page 45	UK National Screening Committee Guidelines (2000)
Systematic Review	Page 16[2]	SIGN checklist (SIGN, 2006), CASP checklist (CASP, 2006)

[1] Included in *How to review the evidence: systematic identification and review of the scientific literature* (NHMRC 2000a)

[2] Included in *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b)

Study design glossary (alphabetic order)

Adapted from NHMRC 2000ab, Glasziou et al 2001, Elwood 1998

Note: This is a specialised glossary that relates specifically to the study designs mentioned in the NHMRC Evidence Hierarchy. Glossaries of terms that relate to wider epidemiological concepts and evidence based medicine are also available – see http://www.inahta.org/HTA/Glossary/; http://www.ebmny.org/glossary.html

All or none –- all or none of a series of people (case series) with the risk factor(s) experience the outcome. The data should relate to 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. This is a rare situation.

A study of test accuracy with: an independent, blinded comparison with a valid reference standard, among consecutive patients with a defined clinical presentation – a cross-sectional study where a consecutive group of people from an appropriate (relevant) population receive the test under study (index test) and the reference standard test. The index test result is not incorporated in (is independent of) the reference test result/final diagnosis. The assessor determining the results of the index test is blinded to the results of the reference standard test and vice versa.

A study of test accuracy with: an independent, blinded comparison with a valid reference standard, among non-consecutive patients with a defined clinical presentation – a cross- sectional study where a non-consecutive group of people from an appropriate (relevant) population receive the test under study (index test) and the reference standard test. The index test result is not incorporated in (is independent of) the reference test result/final diagnosis. The assessor determining the results of the index test is blinded to the results of the reference standard test and vice versa.

Adjusted indirect comparisons – an adjusted indirect comparison compares single arms from two or more interventions from two or more separate studies via the use of a common reference ie A versus B and B versus C allows a comparison of A versus C when there is statistical adjustment for B. This is most commonly done in meta-analyses (see Bucher et al 1997). Such an indirect comparison should only be attempted when the study populations, common comparator/reference, and settings are very similar in the two studies (Song et al 2000).

Case-control study – people with the outcome or disease (cases) and an appropriate group of controls without the outcome or disease (controls) are selected and information obtained about their previous exposure/non-exposure to the intervention or factor under study.

Case series – a single group of people exposed to the intervention (factor under study).

Post-test – only outcomes after the intervention (factor under study) are recorded in the series of people, so no comparisons can be made.

Pre-test/post-test – measures on an outcome are taken before and after the intervention is introduced to a series of people and are then compared (also known as a 'before- and-after study').

Cohort study – outcomes for groups of people observed to be exposed to an intervention, or the factor under study, are compared to outcomes for groups of people not exposed.

Prospective cohort study – where groups of people (cohorts) are observed at a point in time to be exposed or not exposed to an intervention (or the factor under study) and then are followed prospectively with further outcomes recorded as they happen.

Retrospective cohort study – where the cohorts (groups of people exposed and not exposed) are defined at a point of time in the past and information collected on subsequent outcomes, eg. the use of medical records to identify a group of women using oral contraceptives five years ago, and a group of women not using oral contraceptives, and then contacting these women or identifying in subsequent medical records the development of deep vein thrombosis.

Cross-sectional study – a group of people are assessed at a particular point (or cross-section) in time and the data collected on outcomes relate to that point in time ie proportion of people with asthma in October 2004. This type of study is useful for hypothesis-generation, to identify whether a risk factor is associated with a certain type of outcome, but more often than not (except when the exposure and outcome are stable eg. genetic mutation and certain clinical symptoms) the causal link cannot be proven unless a time dimension is included.

Diagnostic (test) accuracy – in diagnostic accuracy studies, the outcomes from one or more diagnostic tests under evaluation (the *index test*/s) are compared with outcomes from a *reference standard test*. These outcomes are measured in individuals who are suspected of having the condition of interest. The term *accuracy* refers to the amount of agreement between the index test and the reference standard test in terms of outcome measurement. Diagnostic accuracy can be expressed in many ways, including sensitivity and specificity, likelihood ratios, diagnostic odds ratio, and the area under a receiver operator characteristic (ROC) curve (Bossuyt et al 2003)

Diagnostic case-control study – the index test results for a group of patients already known to have the disease (through the reference standard) are compared to the index test results with a separate group of normal/healthy people known to be free of the disease (through the use of the reference standard). 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 because the spectrum of study participants will not be representative of patients seen in practice. *Note: this does not apply to well-designed population based case-control studies.*

Historical control study – outcomes for a prospectively collected group of people exposed to the intervention (factor under study) are compared with either (1) the outcomes of people treated at the same institution prior to the introduction of the intervention (ie. control group/usual care), or (2) the outcomes of a previously published series of people undergoing the alternate or control intervention.

Interrupted time series with a control group – trends in an outcome or disease are measured over multiple time points before and after the intervention (factor under study) is introduced to a group of people, and then compared to the outcomes at the same time points for a group of people that do not receive the intervention (factor under study).

Interrupted time series without a parallel control group – trends in an outcome or disease are measured over multiple time points before and after the intervention (factor under study) is introduced to a group of people, and compared (as opposed to being compared to an external control group).

Non-randomised, experimental trial - the unit of experimentation (eg. people, a cluster of people) is allocated to either an intervention group or a control group, using a non-random method (such as patient or clinician preference/availability) and the outcomes from each group are compared.

This can include:

(1) a controlled before-and-after study, where outcome measurements are taken before and after the intervention is introduced, and compared at the same time point to outcome measures in the (control) group.

(2) an adjusted indirect comparison, where two randomised controlled trials compare different interventions to the same comparator ie. the placebo or control condition. The outcomes from the two interventions are then compared indirectly. *See entry on adjusted indirect comparisons.*

Pseudo-randomised controlled trial - the unit of experimentation (eg. people, a cluster of people) is allocated to either an intervention (the factor under study) group or a control group, using a pseudo-random method (such as alternate allocation, allocation by days of the week or odd-even study numbers) and the outcomes from each group are compared.

Randomised controlled trial – the unit of experimentation (eg. people, or a cluster of people[2]) is allocated to either an intervention (the factor under study) group or a control group, using a random mechanism (such as a coin toss, random number table, computer-generated random numbers) and the outcomes from each group are compared.

Reference standard - the reference standard is considered to be the best available method for establishing the presence or absence of the target condition of interest. The reference standard can be a single method, or a combination of methods. It can include laboratory tests, imaging tests, and pathology, but also dedicated clinical follow-up of individuals (Bossuyt et al 2003).

Screening intervention – a screening intervention is a public health service in which members of

[2] Known as a cluster randomised controlled trial

a defined population, who do not necessarily perceive that they are at risk of, or are already affected by a disease or its complications (asymptomatic), are asked a question or offered a test, to identify those individuals who are more likely to be helped than harmed by further tests or treatment to reduce the risk of a disease or its complications (UK National Screening Committee, 2007). A screening intervention study compares the implementation of the screening intervention in an asymptomatic population with a control group where the screening intervention is not employed or where a different screening intervention is employed. The aim is to see whether the screening intervention of interest results in improvements in patient-relevant outcomes eg survival.

Study of diagnostic yield – these studies provide the yield of diagnosed patients, as determined by the index test, without confirmation of the accuracy of the diagnosis (ie. whether the patient is actually diseased) by a reference standard test.

Systematic review – systematic location, appraisal and synthesis of evidence from scientific studies.

Test - any method of obtaining additional information on a person's health status. It includes information from history and physical examination, laboratory tests, imaging tests, function tests, and histopathology (Bossuyt et al 2003).

Two or more single arm study – the outcomes of a single series of people receiving an intervention (case series) from two or more studies are compared. *Also see entry on unadjusted indirect comparisons.*

Unadjusted indirect comparisons – an unadjusted indirect comparison compares single arms from two or more interventions from two or more separate studies via the use of a common reference ie A versus B and B versus C allows a comparison of A versus C but there is no statistical adjustment for B. Such a simple indirect comparison is unlikely to be reliable (see Song et al 2000).

How to assess the body of evidence and formulate recommendations

This part of the document describes how to grade the ‘body of evidence’ for each guideline recommendation. The body of evidence considers the evidence dimensions of all the studies relevant to that recommendation. To assist guideline developers, the NHMRC GAR consultants have developed an approach for assessing the body of evidence and formulating recommendations. This will ensure that while guidelines may differ in their purpose and formulation, their developmental processes are consistent, and their recommendations are formulated in a consistent manner.

Consequently, the NHMRC Evidence Statement Form is intended to be used for each clinical question addressed in a guideline. Before completing the form, each included study should be critically appraised and the relevant data extracted and summarised as shown in the *NHMRC standards and procedures for externally developed guidelines* (NHMRC 2007). This information assists in the formulation of the recommendation, and in determining the overall grade of the ‘body of evidence’ that supports that recommendation.

The NHMRC Evidence Statement Form sets out the basis for rating five key components of the ‘body of evidence’ for each recommendation. These components are:

1. The evidence base, in terms of the number of studies, level of evidence and quality of studies (risk of bias).
2. The consistency of the study results.
3. The potential clinical impact of the proposed recommendation.
4. The generalisability of the body of evidence to the target population for the guideline.
5. The applicability of the body of evidence to the Australian healthcare context.

The first two components give a picture of the internal validity of the study data in support of efficacy (for an intervention), accuracy (for a diagnostic test), or strength of association (for a prognosis or aetiological question). The third component addresses the likely clinical impact of the proposed recommendation. The last two components consider external factors that may influence the effectiveness of the proposed recommendation in practice, in terms of the generalisability for the intended population and setting of the proposed recommendation, and applicability to the Australian (or other local) health care system.

Definitions of the components of the evidence statement[3]

1. Evidence base

The evidence base is assessed in terms of the quantity, level and quality (risk of bias) of the included studies:

- *Quantity of evidence* reflects the number of the studies that have been included as the evidence base for each guideline (and listed in the evidence summary table or text). The quantity assessment also takes into account the number of patients in relation to the frequency of the outcomes measured (ie the statistical power of the studies). Small, underpowered studies that are otherwise sound may be included in the evidence base if their findings are generally similar — but at least some of the studies cited as evidence must be large enough to detect the size and direction of any effect. Alternatively, the results of the studies could be considered in a meta-analysis to increase the power and statistical precision of the effect estimate.

- *Level of evidence* reflects the best study types for the specific type of question (see Table 1). The most appropriate study design to answer each type of clinical question (intervention, diagnostic accuracy, aetiology or prognosis) is level II evidence. Level I studies are systematic reviews of the appropriate level II studies in each case. Study designs that are progressively less robust for answering each type of question are shown at levels III and IV.

- *Quality of evidence* reflects how well the studies were designed in order to eliminate bias, including how the subjects were selected, allocated to groups, managed and followed up (see Part A, Dimensions of evidence, and Table 2 for further information).

2. Consistency

The consistency component of the 'body of evidence' assesses whether the findings are consistent across the included studies (including across a range of study populations and study designs). Ideally, for a meta-analysis of randomised studies, there should be a statistical analysis of heterogeneity showing little statistical difference between the studies. However, given that statistical tests for heterogeneity are underpowered, presentation of an I^2 statistic[4], as well as an appraisal of the reasons for the heterogeneity between studies, would be useful. Heterogeneity between studies may be due to differences in the study design, the quality of the studies (risk of bias), the population studied, the definition of the outcome being assessed, as well as many other factors. Non-randomised studies may have larger estimates of effect as a result of the greater bias in such studies; however, such studies may also be important for confirming or questioning results from randomised trials in larger populations that may be more representative of the target population for the proposed guideline.

3. Clinical impact

Clinical impact is a measure of the potential benefit from application of the guideline to a population. Factors that need to be taken into account when estimating clinical impact include:

- the relevance of the evidence to the clinical question, the statistical precision and size of the effect (including clinical importance) of the results in the evidence-base, and the relevance of the effect to the patients, compared with other management options (or none)

[3] Adapted from the Scottish Intercollegiate Guidelines Network (SIGN) guide to using their Considered Judgement Form (available from http://www.sign.ac.uk/guidelines/fulltext/50/annexd.html Accessed 19.10.07)

[4] whereas most statistical tests of heterogeneity (eg Cochran's Q) assess whether heterogeneity *exists* between studies, I^2 is a statistic that quantifies *how much* heterogeneity exists between the studies (see Higgins & Thompson, 2002)

- the duration of therapy required to achieve the effect, and
- the balance of risks and benefits (taking into account the size of the patient population concerned).

4. Generalisability

This component covers how well the subjects and settings of the included studies match those of the recommendation. Population issues that might influence the relative importance of recommendations include gender, age or ethnicity, baseline risk, or the level of care (eg community or hospital). This is particularly important for evidence from randomised controlled trials (RCTs), as the setting and entry requirements for such trials are generally narrowly based and therefore may not be representative of all the patients to whom the recommendation may be applied in practice. Confirmation of RCT evidence by broader-based population studies may be helpful in this regard (see '2. Consistency').

In the case of studies of diagnostic accuracy, a number of additional criteria also need to be taken into account, including the stage of the disease (eg early versus advanced), the duration of illness and the prevalence of the disease in the study population as compared to the target population for the guideline.

5. Applicability

This component addresses whether the evidence base is relevant to the Australian health care setting generally, or to more local settings for specific recommendations (such as rural areas or cities).

Factors that may reduce the direct application of study findings to the Australian or more local settings include organisational factors (eg availability of trained staff, clinic time, specialised equipment, tests or other resources) and cultural factors (eg attitudes to health issues, including those that may affect compliance with the recommendation).

How to use the NHMRC Evidence Statement Form

Step 1 — Rate each of the five components

Applying evidence in real clinical situations is not usually straightforward. Consequently guideline developers find that the body of evidence supporting a recommendation rarely consists of entirely one rating for all the important components (outlined above). For example, a body of evidence may contain a large number of studies with a low risk of bias and consistent findings, but which are not directly applicable to the target population or Australian healthcare context and have only a limited clinical impact. Alternatively, a body of evidence may only consist of one or two randomised trials with small sample sizes that have a moderate risk of bias but have a very large clinical impact and are directly applicable to the Australian healthcare context and target population. The NHMRC evidence grading system is designed to allow for this mixture of components, while still reflecting the overall body of evidence supporting a guideline recommendation.

The components described above should be rated according to the matrix shown in Table 3. Enter the results into the NHMRC Evidence Statement Form along with any further notes relevant to the discussions for each component.

Table 3 Body of evidence matrix

Component	A	B	C	D
	Excellent	Good	Satisfactory	Poor
Evidence base[1]	several level I or II studies with low risk of bias	one or two level II studies with low risk of bias or a SR/multiple level III studies with low risk of bias	level III studies with low risk of bias, or level I or II studies with moderate risk of bias	level IV studies, or level I to III studies with high risk of bias
Consistency[2]	all studies consistent	most studies consistent and inconsistency may be explained	some inconsistency reflecting genuine uncertainty around clinical question	evidence is inconsistent
Clinical impact	very large	substantial	moderate	slight or restricted
Generalisability	population/s studied in body of evidence are the same as the target population for the guideline	population/s studied in the body of evidence are similar to the target population for the guideline	population/s studied in body of evidence differ to target population for guideline but it is clinically sensible to apply this evidence to target population[3]	population/s studied in body of evidence differ to target population and hard to judge whether it is sensible to generalise to target population
Applicability	directly applicable to Australian healthcare context	applicable to Australian healthcare context with few caveats	probably applicable to Australian healthcare context with some caveats	not applicable to Australian healthcare context

[1] Level of evidence determined from the NHMRC evidence hierarchy

[2] If there is only one study, rank this component as 'not applicable'.

[3] For example, results in adults that are clinically sensible to apply to children OR psychosocial outcomes for one cancer that may be applicable to patients with another cancer

The Evidence Statement Form also provides space to enter any other relevant factors that were taken into account by the guideline developers when judging the body of evidence and developing the wording of the recommendation.

Step 2 — Prepare an evidence statement matrix

In the 'Evidence statement matrix ' section of the form, summarise the guideline developers' synthesis of the evidence relating to each component at the right hand side of the form, and fill in the evidence matrix at the left hand side of the form. Each recommendation should be accompanied by this matrix as well as the overall grade given to the recommendation (see Step 3). Developers should indicate dissenting opinions or other relevant issues in the space provided under the evidence matrix.

Step 3 — Formulate a recommendation based on the body of evidence

Develop wording for the recommendation. This should address the specific clinical question and ideally be written as an action statement. The wording of the recommendation should reflect the strength of the body of evidence. Words such as 'must' or 'should' are used when the evidence underpinning the recommendation is strong, and words such as 'might' or 'could' are used when the evidence body is weaker.

Step 4 — Determine the grade for the recommendation

Determine the overall grade of the recommendation based on a summation of the rating for each individual component of the body of evidence. **A recommendation cannot be graded A or B unless the evidence base and consistency of the evidence are both rated A or B.**

NHMRC overall grades of recommendation are intended to indicate the strength of the body of evidence underpinning the recommendation. This should assist users of the clinical practice guidelines to make appropriate and informed clinical judgments. Grade A or B recommendations are generally based on a body of evidence that can be trusted to guide clinical practice, whereas Grades C or D recommendations must be applied carefully to individual clinical and organisational circumstances and should be interpreted with care (see Table 4).

Table 4 Definition of NHMRC grades of recommendations

Grade of recommendation	Description
A	Body of evidence can be trusted to guide practice
B	Body of evidence can be trusted to guide practice in most situations
C	Body of evidence provides some support for recommendation(s) but care should be taken in its application
D	Body of evidence is weak and recommendation must be applied with caution

Implementing guideline recommendations

How the guideline will be implemented should be considered at the time that the guideline recommendations are being formulated. Guidelines require an implementation plan that ensures appropriate roll out, supports and evaluation of guideline effectiveness in improving practice, and guideline uptake. The Evidence Statement Form asks developers to consider four questions related to the implementation of each recommendation:

- Will this recommendation result in changes in usual care?
- Are there any resource implications associated with implementing this recommendation?
- Will the implementation of this recommendation require changes in the way care is currently organised?
- Are the guideline development group aware of any barriers to the implementation of this recommendation?

Conclusion

This paper outlines an approach piloted and refined over two years by NHMRC GAR consultants. This approach reflects the concerted input of experience in assisting a range of guideline developers to develop guidelines for a range of conditions and purposes. This approach provides a way forward for guideline developers to appraise, classify and grade evidence relevant to the purpose of a guideline. With further application of these levels and grades of evidence, modifications will inevitably be made to further improve guideline development processes.

ATTACHMENT 1

NHMRC Evidence Statement

(If rating is not completely clear, use the space next to each criteria to note how the group came to a judgment.)

Key question(s):		Evidence table ref:
1. Evidence base *(number of studies, level of evidence and risk of bias in the included studies)*		
	A	Several Level I or II studies with low risk of bias
	B	one or two Level II studies with low risk of bias or SR/multiple Level III studies with low risk of bias
	C	Level III studies with low risk of bias or Level I or II studies with moderate risk of bias
	D	Level IV studies or Level I to III studies with high risk of bias
2. Consistency *(if only one study was available, rank this component as 'not applicable')*		
	A	All studies consistent
	B	Most studies consistent and inconsistency can be explained
	C	Some inconsistency, reflecting genuine uncertainty around question
	D	Evidence is inconsistent
	NA	Not applicable (one study only)
3. Clinical impact *(Indicate in the space below if the study results varied according to some <u>unknown</u> factor (not simply study quality or sample size) and thus the clinical impact of the intervention could not be determined)*		
	A	Very large
	B	Moderate
	C	Slight
	D	Restricted
4. Generalisability		
	A	Evidence directly generalisable to target population
	B	Evidence directly generalisable to target population with some caveats
	C	Evidence not directly generalisable to the target population but could be sensibly applied
	D	Evidence not directly generalisable to target population and hard to judge whether it is sensible to apply
5. Applicability		
	A	Evidence directly applicable to Australian healthcare context
	B	Evidence applicable to Australian healthcare context with few caveats
	C	Evidence probably applicable to Australian healthcare context with some caveats
	D	Evidence not applicable to Australian healthcare context

Other factors *(Indicate here any other factors that you took into account when assessing the evidence base (for example, issues that might cause the group to downgrade or upgrade the recommendation)*

EVIDENCE STATEMENT MATRIX

Please summarise the development group's synthesis of the evidence relating to the key question, taking all the above factors into account.

Component	Rating	Description
1. Evidence base		
2. Consistency		
3. Clinical impact		
4. Generalisability		
5. Applicability		

Indicate any dissenting opinions

RECOMMENDATION *What recommendation(s) does the guideline development group draw from this evidence? Use action statements where possible.*	GRADE OF RECOMMENDATION	

IMPLEMENTATION OF RECOMMENDATION *Please indicate yes or no to the following questions. Where the answer is yes please provide explanatory information about this. This information will be used to develop the implementation plan for the guidelines.*	
Will this recommendation result in changes in usual care?	YES
	NO
Are there any resource implications associated with implementing this recommendation?	YES
	NO
Will the implementation of this recommendation require changes in the way care is currently organised?	YES
	NO
Are the guideline development group aware of any barriers to the implementation of this recommendation?	YES
	NO

References

Bandolier editorial. Diagnostic testing emerging from the gloom? *Bandolier*, 1999;70. Available at: http://www.jr2.ox.ac.uk/bandolier/band70/b70-5.html

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HCW for the STARD Group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *AJR*, 2003; 181:51-56

Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*, 1997;50:683-91.

CASP (2006). Critical Appraisal Skills Programme (CASP) - making sense of evidence: 10 questions to help you make sense of reviews. England: Public Health Resource Unit. Available at: http://www.phru.nhs.uk/Doc_Links/S.Reviews%20Appraisal%20Tool.pdf

Elwood M. (1998) *Critical appraisal of epidemiological studies and clinical trials*. Second edition. Oxford: Oxford University Press.

Glasziou P, Irwig L, Bain C, Colditz G. (2001) *Systematic reviews in health care. A practical guide.* Cambridge: Cambridge University Press.

Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*, 2002; 21(11):1539-58.

Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, Bossuyt PMM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*, 1999; 282(11):1061-6.

Medical Services Advisory Committee (2005). *Guidelines for the assessment of diagnostic technologies*. [Internet] Available at: www.msac.gov.au

Mulherin S, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med*, 2002;137:598-602.

NHMRC (1999). *A guide to the development, implementation and evaluation of clinical practice guidelines*. Canberra: National Health and Medical Research Council.

NHMRC (2000a). *How to review the evidence: systematic identification and review of the scientific literature*. Canberra: National Health and Medical Research Council.

NHMRC (2000b). *How to use the evidence: assessment and application of scientific evidence*. Canberra: National Health and Medical Research Council.

NHMRC (2007). *NHMRC standards and procedures for externally developed guidelines*. Canberra: National Health and Medical Research Council. http://www.nhmrc.gov.au/publications/synopses/_files/nh56.pdf

NZGG (2001). *Handbook for the preparation of explicit evidence-based clinical practice guidelines*. Wellington: New Zealand Guidelines Group. Available at: http://www.nzgg.org.nz

Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B, Dawes M (2001). *Oxford Centre for Evidence-Based Medicine levels of evidence (May 2001).* Oxford: Centre for Evidence-Based Medicine. Available at: http://www.cebm.net/levels_of_evidence.asp

Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ*, 2002;324:539-41.

SIGN. *SIGN 50. A guideline developers' handbook*. Methodology checklist 1: Systematic reviews and meta-analyses. Edinburgh: Scottish Intercollegiate Guidelines Network. Available at: http://www.sign.ac.uk/guidelines/fulltext/50/checklist1.html

Song F, Glenny A-M, Altman DG. Indirect comparison in evaluating relative efficacy illustrated by antimicrobial prophylaxis in colorectal surgery. *Controlled Clinical Trials*, 2000;21(5):488-497.

UK National Screening Committee (2000). *The UK National Screening Committee's criteria for appraising the viability, effectiveness and appropriateness of a screening programme*. In: Second Report of the UK National Screening Committee. London: United Kingdom Departments of Health. Pp. 26-27. Available at: http://www.nsc.nhs.uk/

UK National Screening Committee. *What is screening?*. [Internet]. Available at - http://www.nsc.nhs.uk/whatscreening/whatscreen_ind.htm [Accessed August 2007].

Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS:a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003; 3(1): 25. Available at: http://www.biomedcentral.com/1471-2288/3/25