SUPPORTING DOCUMENT FOR THE JOANNA BRIGGS INSTITUTE LEVELS OF EVIDENCE AND GRADES OF RECOMMENDATION

Developed by the JBI Levels of Evidence and Grades of Recommendation Working Party January 2014

jbi.global

CONTENTS

INTRODUCTION

It is of the utmost importance for the clinician attempting to implement evidence based research to understand the strength of evidence of a piece of research by critically appraising or assessing its methodological quality. By applying a level of evidence to a certain piece of information based on its study design, the clinician is able to make a preliminary judgement on the methodological quality and rigour of the evidence.

Hierarchies to rank evidence have existed for a number of years, with many organisations developing their own unique ranking and grading systems. These systems have come under criticism due to their superficial nature as they automatically promote evidence from experimental studies over observational studies. This does not necessarily reflect reality where at times evidence from observational studies may in fact be of more value than that from experimental studies. Due to this, there has been a push internationally to adopt the approach of the GRADE (Grading of Recommendations Assessment, Development and Evaluation) working group[1] who have developed a grading of evidence and recommendation system that has been endorsed by many evidence-based healthcare organisations, including Cochrane, WHO, AHRQ, NICE, BMJ Clinical Evidence and SIGN, amongst others.

The approach of GRADE is not to classify findings based only on study design but other factors as well. These include critical appraisal/risk of bias, publication bias, inconsistency, indirectness, and imprecision of evidence, effect size, dose-response relationships, and confounders. The evidence is then ranked into one of four levels (High, Moderate, Low, Very Low). This process begins with studies being pre-ranked based on their design (High = RCTs, Low = observational studies), and then downgraded or upgraded based on the aforementioned factors. A new, more nuanced ranking can then be assigned to an individual finding or outcome. In this way, evidence from observational studies can be ranked above that of randomised controlled trials where appropriate. This process is often presented in a summary of findings table.[1]

The Joanna Briggs Institute and its collaborating entities have recently decided to adopt the GRADE approach for systematic reviews of effectiveness. Guidance for conducting systematic reviews is provided in the JBI reviewer's manual. However, it is the view of the Institute that a system to designate Levels of Evidence is still necessary considering the work conducted by the Institute and its collaborating entities. The main reasons for continuing with a Levels of Evidence system are as follows:

- To assist in assigning GRADE pre-rankings to studies when conducting systematic reviews.
- For resources such as evidence summaries which require a rapid review and classification of literature (for example, the Levels of Evidence can provide information on the most appropriate study design to search for when asking a clinical question). Following the GRADE guidance for

developing a full summary of findings table/evidence profile for each evidence summary is beyond the scope of these documents. These tables can be included in JBI systematic reviews of effects.

- For educational purposes for health professionals.

Therefore, JBI, due to its broader view of what constitutes research evidence for practice, has developed its own unique Levels of Evidence and Grades of Recommendation. These Levels of Evidence are utilised across JBI, its international collaboration and related entities and are incorporated in many of its evidence-based resources, specifically Systematic Reviews, Evidence Summaries, Best Practice Information Sheets and Recommended Practices.

It is important to note that these Levels of Evidence provide a ranking based on the ***likely*** best available evidence, and should not be used as a definitive measure of the best available evidence. As mentioned above, it may be that evidence that comes from observational studies should sometimes be preferred over that which comes from experimental studies. Although the Levels provide useful information to initially rank a study design, they should not act as a substitute for critical appraisal and clinical reasoning.

These levels have been designed with the following in mind:

- Ability to be easily incorporated into the GRADE approach.
- Consistent across evidence types to be based solely on study design (i.e. they do not relate to strength of findings).
- To provide clarity to users of levels of evidence and address common feedback.
- To incorporate additional types of evidence not included in previous hierarchies.

USING THE LEVELS OF EVIDENCE

Effectiveness

The levels of evidence for therapy/ interventions are designed to align with the GRADE approach to preranking findings based on the study design, which are then upgraded or downgraded depending on a number of factors.

Diagnosis

The levels of evidence for diagnosis have been designed to align with the GRADE approach to preranking findings based on the study design, which are then upgraded or downgraded based on a number of factors. When discussing diagnostic studies, the GRADE group state there are two main study designs which are used;

- RCTs, which investigate patient important outcomes as a result of diagnosis with two different methods
- Studies of test accuracy which evaluate test specificity and sensitivity.[2]

If RCTs are used that assess the effectiveness of diagnostic tests on patient important outcomes, the levels of evidence for effectiveness can be used. The levels of evidence for diagnosis apply to studies assessing only test accuracy.

Prognosis

The levels of evidence for prognosis have been designed to align with the GRADE approach where study findings are pre-ranked findings based on the study design and then upgraded or downgraded based on a number of factors.

Costs

The levels of evidence for costs are different than the other levels of evidence and are not based purely on study design. As costs are always unique to a certain setting and location, these levels are ranked to reflect applicability in the decision making context.

Meaningfulness

By assigning levels of evidence to qualitative studies, JBI addresses one of the most difficult problems in qualitative research, 'that of defining clear criteria for selecting high-quality qualitative studies.[(p.43)3]' This is an area where the JBI levels of evidence differ in comparison to many other institutions as they offer a way to rank qualitative research.

The levels of evidence for qualitative research have been designed to fit with a modified GRADE approach where a study's findings are pre-ranked based on the study design and then upgraded or downgraded depending on a number of factors. The old JBI levels considered credibility of the findings; however, it was deemed by the working party that this should not be considered when assigning a level of evidence, but rather when creating a summary of findings table and moving to recommendations. The new levels reflect study designs only.

USING THE GRADES OF RECOMMENDATION

Grades of Recommendation are used to assist healthcare professionals when implementing evidence into practice. JBI currently assigns a grade of recommendation to all recommendations made in its resources, including Evidence Summaries, Systematic Reviews and Best Practice Information Sheets.

The new JBI grades of recommendation are informed by the GRADE working party, which has a binary system for recommendations, with only the two options: 'strong' or 'weak.' The benefit of such a system is its ease of interpretation by both clinicians and patients. When forming a recommendation, GRADE recommends the following four key factors be considered: the balance between desirable and undesirable effects, the quality of the evidence, values and preferences, and costs.[4, 5] Recommendations can be made for or against particular management approaches.[4, 5] Due to negative connotations associated with the term 'weak,' GRADE have provided the alternative terms of conditional, discretionary or qualified recommendations which can substitute for the term weak.[5]

Recommendations should be actionable. When wording recommendations, the following need to be specified as much as possible, as the more specific a recommendation is, the easier it is to implement and the more likely it is that it will be acted upon: [4-8]

- The population (i.e. age, sex, condition)
- Intervention (i.e. dose, timing, intensity, professional)
- Any comparator (where applicable)
- The setting (where applicable)

Wording for strong recommendations should be in the active voice. This can be achieved by using phrases such as 'we recommend/ Health professionals should/ or Do, or must' [4, 5, 9] For weak recommendations, phrases such as 'we suggest/health professionals might (could/may) /we conditionally recommend' can be used. [4, 5, 9] An example of a strong recommendation is: 'Health professionals should provide written information detailing methods of self-management of blood glucose levels for patients with type 2 diabetes living in the community.' An example of a weak recommendation is: 'Health professionals may provide information regarding self- management of blood glucose levels for patients with type 2 diabetes living in the community.'

As mentioned above, recommendations can be made for or against particular management approaches. When making strong recommendations against a certain strategy, terms such as 'we recommend against, health professionals should not, or don't' can be used.

The use of the term 'consider' has been advised against due to its difficulty of interpretation when determining if a certain activity was considered.[10] Other terms to avoid include the use of phrases such as 'where necessary' or 'when clinically indicated.'[8]. Whatever words are chosen to convey the recommendation, the connection between the strength of the recommendation and the wording needs to be explicit, which can be achieved through consistent use of the same wording structure.[11]

GRADE recommends that symbols are used when presenting recommendations. They suggest the symbol ↑↑ for strong recommendations whilst ↑? is used for weak recommendations. [4, 5] However, JBI will continue using letters to represent the strength of recommendations, with Grade 'A' being a 'strong' recommendation, and Grade 'B' representing a 'weak' recommendation, as this is the approach most familiar to JBI reviewers.

The FAME (Feasibility, Appropriateness, Meaningfulness and Effectiveness) scale may help inform the wording of a recommendation. The following elements should be taken into consideration when applying the evidence, and therefore recommendations should be graded accordingly.

F – Feasibility; specifically:

- What is the cost effectiveness of the practice?
- Is the resource/practice available?
- Is their sufficient experience/levels of competency available?

A – Appropriateness; specifically:

- Is it culturally acceptable?
- Is it transferable/applicable to the population of interest?
- Is it easily adaptable to a variety of circumstances?

M – Meaningfulness; specifically:

- Is it associated with positive experiences?
- Is it not associated with negative experiences?

E – Effectiveness; specifically:

- Was there a beneficial effect?
- Is it safe? (i.e. is there a lack of harm associated with the practice?)

GLOSSARY

Definitions are taken verbatim from the NHMRC guidance,[12] American College of Physicians,[13] and the Centre for Evidence Based Medicine.[14, 15]

All or none studies:[12]

'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. [12]

Alternatively, this is met when all patients died before the treatment became available, but some now survive on it; or when some patients died before the treatment became available, but none now die on it.[15]

Bench research

Studies that have been conducted with nonhuman subjects in a laboratory setting.

Case – controlled studies[12]

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[14]

A group or series of case reports involving patients who were given similar treatment. Reports of case series usually contain detailed information about the individual patients. This includes demographic information (for example, age, gender, ethnic origin) and information on diagnosis, treatment, response to treatment, and follow-up after treatment.

Case study/report

A description of a single case.

Cohort studies[12]

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,

e.g. 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 studies [12]

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 i.e. 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 e.g. genetic

mutation and certain clinical symptoms) the causal link cannot be proven unless a time dimension is included.

Diagnostic case-control study[12]

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.

Diagnostic yield study

These studies provide the yield of diagnosed patients, as determined by the index test, without confirmation of the accuracy of the diagnosis (i.e. whether the patient is actually diseased) by a reference standard test.

Expert consensus

Evidence arising from the consensus of experts in the field.

Historic/retrospective control group studies [12]

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 (i.e. control group/usual care), or (2) the outcomes of a previously published series of people undergoing the alternate or control intervention.

Inception Cohort Studies[14]

A group of individuals identified for subsequent study at an early, uniform point in the course of the specified health condition, or before the condition develops.

N-of-1 trial[14]

A variation of a randomized controlled trial in which a sequence of alternative treatment regimens is randomly allocated to a patient. The outcomes of regimens are compared, with the aim of deciding on the optimum regimen for the patient.

Pseudo-Randomised Controlled Trials [12]

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

Quasi-experimental study [12]

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.

Randomised Controlled Trials [12]

The unit of experimentation (e.g. people, or a cluster of people) is allocated to either an intervention (the fact or 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.

Sensitivity [13]

The proportion of patients with the target disorder who have a positive test result

Specificity [13]

The proportion of patients without the target disorder who have a negative test result

Systematic reviews [12]

Systematic location, appraisal and synthesis of evidence from scientific studies

Test Accuracy [12]

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 curve (ROC).

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.

HOW TO CITE THIS DOCUMENT:

JBI Levels of Evidence and Grades of Recommendation Working Party*. Supporting Document for the JBI Levels of Evidence and Grades of Recommendation. JBI. 2014. https://jbi.global

*The Levels of Evidence and Grades of Recommendation Working party consists of:

Dr Zachary Munn, Senior Research Fellow, Implementation Science, JBI, The University of Adelaide.
Dr Kylie Porritt, Research Fellow, Implementation Science, JBI, The University of Adelaide.

Associate Professor Edoardo Aromataris, Director Synthesis Science, JBI, The University of Adelaide.

Associate Professor Craig Lockwood, Director Implementation Science, JBI, The University of Adelaide.

Dr Micah Peters, Research Fellow, Synthesis Science, JBI, The University of Adelaide.

REFERENCES

1. GRADE Working group. Education and debate - grading quality of evidence and strength of recommendations. BMJ. 2004;328(1490).
2. Thornton J, Alderson P, Tan T et al. Introducing GRADE across the NICE clinical guideline program. Journal of Clinical Epidemiology. 2013;66(2):124-31.
3. Daly J, Willis K, Small R et al. A hierarchy of evidence for assessing qualitative health research. Journal of Clinical Epidemiology. 2007;60:43-9.
4. Guyatt GH, Oxman AD, Kunz R et al. Going from evidence to recommendations. BMJ. 2008;336(7652):1049-51.
5. Andrews J, Guyatt G, Oxman AD et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol. 2013;66(7):719-25.
6. Michie S, Johnston M. Changing clinical behaviour by making guidelines specific. BMJ. 2004;328(7435):343-5.
7. Michie S, Lester K. Words matter: increasing the implementation of clinical guidelines. Quality & safety in health care. 2005;14(5):367-70.
8. Woolf S, Schunemann HJ, Eccles MP, Grimshaw JM, Shekelle P. Developing clinical practice guidelines: types of evidence and outcomes; values and economics, synthesis, grading, and presentation and deriving recommendations. Implementation science : IS. 2012;7:61.
9. Hillier S, Grimmer-Somers K, Merlin T et al. FORM: an Australian method for formulating and grading recommendations in evidence-based clinical guidelines. BMC medical research methodology. 2011;11:23.
10. Lomotan EA, Michel G, Lin Z, Shiffman RN. How "should" we write guideline
11. Recommendations? Interpretation of deontic terminology in clinical practice guidelines: survey of the health services community. Quality & safety in health care. 2010;19(6):509-13.
12. Akl EA, Guyatt GH, Irani J et al. "Might" or "suggest"? No wording approach was clearly superior in conveying the strength of recommendation. J Clin Epidemiol. 2012;65(3):268-75.
13. Merlin T, Weston A, Tooher R et al. NHMRC additional levels of evidence and grades for recommendations for developers of guidelines In: Council NHaMR, ed.: Australian Government 2009.
14. Physicians ACo. Glossary. 2013 [cited 2014 10th January]; Available from: http://acpjc.acponline.org/shared/glossary.htm
15. Law K, Howick J. OCEBM Table of Evidence Glossary. 2013 [cited 2014 10th January]; Available from: http://www.cebm.net/index.aspx?o=1116
16. Oxford Centre for Evidence Based Medicine. Oxford Centre for Evidence-based Medicine - Levels of Evidence 2009 [cited 5/5/2013]; Available from:
17. http://www.cebm.net/?o=1025