eunethta

EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT

Scoping study of reviews about evidence grading systems

Date: October 15, 2019

Written by: NIPHNO

1 Background

One aim of the EUnetHTA task Group for Common Phrases and GRADE (Grading of Recommendations Assessment, Development and Evaluation) is to formulate recommendations on the use or non-use of GRADE or other internationally adopted rating systems in Joint Assessments.

As a first step, this task group will examine existing evaluations of evidence grading systems. The task group will use these findings to inform the project plan. The findings will also be used as one of the elements to make recommendations upon the most appropriate evidence grading system for use within EUnetHTA.

The objective of this scoping study is to describe the breadth and key findings of reviews on evidence grading systems. This study does not aim to provide an in-depth analysis of individual pieces of research. Neither does this study aim to provide an overview of how the different evidence grading systems function.

2 Methods

As to provide timely information to the task group, this scoping study has a pragmatic nature. As opposed to a full systematic review, the search was not as comprehensive and the selection/extraction process was done by only one reviewer. Readers should use this report for internal EUnetHTA purposes only.

One reviewer (Stijn Van de Velde) searched PubMed for existing systematic reviews using the search terms (((((((certainty[ti] OR grading[ti] OR strength[ti] OR quality[ti]) AND evidence[ti])) OR ("grading system"[ti] OR "grading systems"[ti]))) AND ("Systematic Reviews as Topic"[Majr] OR "Practice Guidelines as Topic"[MAJR] OR "Review Literature as Topic"[MAJR] OR "Evidence-Based Practice"[MAJR]))). One reviewer (Sari Ormstad) also scanned the first 100 hits in Google Scholar to identify any reports that were not indexed in Pubmed. Here we used the following search terms ("grading system" or "grading systems" or ((certainty or strength or quality) and evidence) and review). We limited the search to studies that were published in the last 10 years (2009-2019).

We defined evidence grading systems as systems to assess the quality of the accumulated bodies of evidence and to communicate (un)certainty about the estimated effects of the evaluated interventions.[1] We excluded papers that focused only on the evaluation of risk of bias in individual studies. We only included papers that evaluated two or more evidence grading systems and excluded any primary studies (for example to collect empirical data based on head-to-head comparisons of two or more grading systems).

We extracted and synthesized information about strengths and weaknesses for evaluating the quality of evidence and information about actual usage of the systems. We did not extract information about the set-up for each system. Neither did we extract information related to characteristics for expressing strength of recommendations. This scoping review only

extracted key information and readers are invited to read the full papers for further details. Findings were taken directly from the selected reports with only minor edits.

3 Results

Overview of the included studies

Corabian 2018[1]	
Context: Report by the Institute of Health Economics to inform on the use of evidence grading systems when doing systematic reviews about safety and effectiveness of healthcare technologies. **Definitions:** -Prominent evidence grading systems are defined here as those that are top-rated by published comprehensive systematic reviews of existing evidence grading systems for fully covering at least three of the evidence domains identified by experts in review methodology as important for grading the strength of a body of evidence: quality, quantity, and consistency. -Generic grading systems are defined as those that are not deliberately focused on a specific clinical condition or healthcare technology or used to answer only specific research questions. **Methods:** Use of published systematic reviews to identify prominent evidence grading systems; Screening of systematic reviews and HTAs published in 2014 to summarize which of the prominent evidence grading systems were used; Survey with INAHTA members about approaches taken; Evaluation of inter-rater reliability studies based on a systematic review of evaluation studies reporting on reproducibility of evidence grading systems when used by researchers.	
Questions:	**Results:**
1. What are the prominent evidence grading systems?	Based on appropriateness criteria for evidence grading systems (i.e. “do the systems consider quality, quantity and consistency of studies for a given question?”), the reviews by AHRQ in 2002 and by CADTH in 2012 top rated 10 evidence grading systems out of 60 systems in total. This included ICSI, USPSTF, OCEBM, GRADE, SIGN50, NHMRC, Cochrane handbook, and three other guidebooks.
2. Which of the existing prominent generic evidence grading systems are used by researchers in systematic review and HTA organizations/agencies?	1242 systematic reviews and HTA’s were identified. 640 (52%) systematic reviews and HTA’s mentioned the use of prominent evidence grading systems in 604 studies (GRADE n=547, GRADE modified version n=45, NHMRC n=7, USPSTF n=5). In 36 studies, GRADE was planned to be used but not applied (authors judged that it was not feasible to use GRADE n=11, empty review n=25). Among 50 INAHTA members, 12 replied to the survey about the use of evidence grading systems (GRADE n=7, modified GRADE version n=1, other grading system n=2, no grading system used n=3).

3. What is the degree of agreement (consistency) among researchers who apply the same prominent system to the same body of evidence?	Three studies were selected that evaluated either GRADE or a modified version of GRADE. The inter-rater reliability estimates varied from slight to almost perfect agreement for the domain and overall quality of evidence scores.

Conclusions as formulated by the authors (extracts): Uptake of prominent evidence grading systems varies. GRADE and its modified versions is most commonly used. The evidence indicates that GRADE and its modified versions needs to be improved (guidance, new tools or modifications to existing tools) in order to obtain acceptable reliability scores. The reliability studies were based on older versions of GRADE when detailed guidance and supporting tools were not yet properly developed.

Irving 2017 [2]

Context: To analyse grading systems that are used to inform health policy and for the development of clinical practice guidelines, with focus on their use and potential for misuse.

Methods: A narrative review (snowball approach) of papers that reviewed grading instruments.

Questions:	**Results:**
What are the limitations of grading systems for public health?	Few systems provide evidence of item validity or reliability of use. There is poor concurrent validity and the use of different instruments may lead to different conclusions and recommendations. Grading instruments may focus solely on scientific robustness and not evaluate the external validity of findings. Grading systems may not be inherently logical when trading off different elements against each other to establish the level of quality. Grading systems are susceptible to subjectivity and their grading guidelines may be interpreted differently by different assessors. Instructions of grading systems may be inadequate or overly complex. Grading systems may be biased towards RCTs. Even with the ability to up/downgrade ratings, flawed RCTs may be rated higher than strong non-RCTs. There may be a lack of recognition of the large methodological differences that fall under the non-RCT or “observational” umbrella.

Conclusions as formulated by the authors (extracts): Grading instruments are susceptible to misuse because of their complexity, insufficient instructions, and their reliance on the traditional evidence hierarchy that places RCT’s at the apex irrespective of context. The majority of instruments have not been validated, and of those that have been subjected to tests of reliability,

the results have tended to be unfavourable. The consequences of inaccurate grading are serious. Rating of research also provides a possible avenue for public or parties with vested interests to misinterpret or misuse evidence grades. There is a need apply the most appropriate grading instrument to both the research question being asked and the type of evidence being used.

Andreyeva 2012 [3]

Context: To analyse systems for grading evidence and recommendations created and currently used in other countries as to inform use or creation of similar grading systems in Russia.

Methods: Selection of grading systems from well-known international agencies for health technology assessment and organizations responsible for the production of clinical guidelines and elaboration of a comparative analysis based on the following criteria for assigning levels of evidence (quality, quantity and consistency of evidence). The covered systems include SIGN, OCEBM, GRADE, NICE, NHMRC.

Questions:	**Results:**
Comparative analysis of different systems for grading evidence and recommendations	The main difference in grading evidence concerns the object that is graded: OCEBM: level of evidence is assigned to each individual study, not intended for development of clinical practice guidelines. SIGN: a group of studies are assigned an overall level of evidence GRADE, NICE: level of evidence is assigned to pooled evidence relating to each individual treatment outcome from all studies. In GRADE the overall level of evidence corresponds to the lowest level of evidence among all critical and important outcomes (GRADE). NHMRC: assesses pooled evidence in a number of separate domains. This system lacks a clear distinction between grading of evidence and grading of recommendations. None of these systems completely eliminated the need for judgments (often subjective) by expert members of the task force. All the systems emphasize that every decision about the level of evidence must be documented in detail.
Prospects for implementing a unified system for grading evidence and recommendations in Russia and in other countries	A uniform grading system would eliminate any confusion about how to interpret and implement clinical guidelines, and would make it impossible to “go fishing” for a grading system that would assign the highest level of evidence and grade of recommendation to a particular intervention. The most important argument against a unified grading system is doubt about the feasibility of having one adequate system for the entire range of medical problems.

	A growing number of international organizations are switching to GRADE. It is not clear whether GRADE will emerge as the new standard.

Conclusions as formulated by the authors (extracts): No specific conclusions extracted.

Gopalakrishna 2013 [4]

Context: To make an inventory of evidence grading systems for medical tests and to compare the methods in each of these systems. The objective of this review was not to make an analytical appraisal of the different grading systems available within the context of guideline development.

Methods: Review and description of systems that included a 'levels of evidence' and 'strength of recommendations' table.

Questions:	**Results:**
1. Which evidence-grading systems for medical tests exist?	We identified 12 eligible evidence grading systems that could be used by guideline developers to develop guidelines for medical tests.
2. Which methodological and process criteria (23 items that are derived from the AGREE checklist) does each system address?	The EGAPP, USPSTF, NICE, GRADE, and NHMRC systems addressed more items than the other grading systems.

Conclusions as formulated by the authors (extracts): Five systems for grading evidence about medical tests in guideline development addressed to differing degrees of explicitness the need for and appraisal of different bodies of evidence, the linking of such evidence, and its translation into recommendations. At present, no one system addressed the full complexity of gathering, assessing and linking different bodies of evidence.

Bai et al, 2012 [5]

Context: To identify appropriate evidence grading systems for use by CADTH.

Definitions: An appropriate evidence grading system was defined as a system that is most feasible and efficient for CADTH work.

Methods: Update of the AHRQ 2002 report, appraisal of the identified systems, expert consultation and collection of stakeholder input. The systems were appraised on the domains: quality (concept of validity), quantity (number of studies and subjects included in those studies), and consistency (extent to which findings are similar between different studies on the same topic).

Questions:	**Results:**

What is the most appropriate quality assessment tool for grading evidence?	The authors identified 60 grading systems and assessed them according to predefined criteria. Six systems received top-scores and were further assessed by nine experts. The GRADE system was identified as the most preferred system. Although the grading system of SIGN 50 also received a high score, GRADE was selected because: more international recognition, better match with needs of the COMPUS expert committee to make optimal drug use recommendations, more focus on relevant/critical outcomes (while also considering other outcomes in the decision process), financial costs, and factors in the lowest quality studies in the decision process, continuous improvement of grading system.

Conclusions as formulated by the authors (extracts): GRADE was selected as the most appropriate evidence grading system. Applying quality assessment instruments and evidence grading systems systematically and consistently can make our evaluations more transparent, and thus, can help reviewers, expert panels, or government agencies more effectively translate evidence into more comprehensive, reliable, and practical recommendations.

Steelman 2011 [6]

Context: To identify the most applicable rating method for perioperative nursing practice, evaluate the reliability of this method, and identify barriers and facilitators to adoption of this method for AORN recommendations.

Methods: A literature search to find systematic evaluations of methods of rating scientific evidence, expanded to include rating methods with which the task force members had positive professional experience. The AHRQ appropriateness criteria for evidence grading systems (quality, quantity, consistency) were used as selection criteria. Selected systems were evaluated on ease of application, ease of teaching to others, understandability, credibility, applicability to non-RCTs. The reliability and determinants for adoption were only evaluated for the rating system that scored best on the five previously listed criteria.

Questions:	**Results:**
1. Which systems are available that match the AHRQ appropriateness criteria?	The authors identified 46 systems and 10 systems met the three criteria.
2. What is the applicability of the selected systems?	ASPAN, OCEBM, GRADE, ONS, USPSTF scored better on one or more of the criteria. ONS was the system that scored best overall. GRADE scored best on credibility and lowest on applicability to non-RCTs.
3. What is the reliability of the most applicable system?	Reliability was further evaluated for the ONS system only.
4. What are the determinants for adoption for the most applicable system?	Determinants were further evaluated for the ONS system only.

Conclusions as formulated by the authors (extracts): However, selection of an evidence-rating method is only the first step. An implementation plan will be developed to achieve integration of evidence rating into AORN documents. This plan should address education of those who will implement the new process as well as the end users of AORN documents. Resources must be allocated to provide the time to evaluate the quality of individual studies as well as to rate the collective evidence that supports AORN recommendations.

Baker 2010 [7]

Context: Appraisal of grading systems for use within the development of clinical practice guidelines.

Methods: A group of experts appraised a selected number of grading systems, i.e. SIGN50 (because of its established use by societies and the familiarity of guideline development groups with the system), GRADE (because of its methodological rigour and the extensive resources used to produce its appraisal system), GATE (due to its simplicity and clarity, and its ability to be used to critically appraise different types of studies) and NSF-LTC (due to its ability to offer a real alternative to SIGN and GRADE through its holistic interpretation of medical research; it also aims at a new approach to critically appraising RCT, non-RCT and qualitative studies as well as expert opinion).

Questions:	**Results:**
1. What is the suggested appraisal system for different research fields?	Therapy: SIGN or GRADE Diagnosis: GRADE or NSF-LTC Screening: GRADE or NSF-LTC Prognosis: NSF-LTC Causation: GRADE Psychometric studies: NSF-LTC Qualitative studies: NSF-LTC
2. What are the strengths and weaknesses for each system?	GRADE Strengths: established system, robust appraisal system, allows the assessment of a number of variables, appraisal focus is on RCT's, More robust at appraising observational studies than SIGN; emphasizes explicit judgements to increase transparency Weaknesses: Classifies study types by hierarchy, training is required, weak on case reports SIGN Strengths: established system, appraisal focus is on RCT's Weaknesses: training is required NSF-LTC Strengths: Easy to use; flexible, acknowledges qualitative studies and expert opinion

Weaknesses: fewer variables assessed, does not explicitly take into account confounding and size of effect, places expert opinion on equal status to other studies

GATE

Strengths: Excellent for teaching critical appraisal of papers

Weaknesses: does not assign a grade to papers or recommendations and therefore its use in guideline development is limited

Conclusions as formulated by the authors (extracts): The decision on which grading system should be used for specialist society guidelines depends on the research area to which the guideline questions pertain. If the research field and study designs for a guideline are largely homogenous, then one system needs only be used. If, as is often the case, the study designs are heterogeneous, the specialist society will need to carefully consider the options for critical appraisal systems. While it is possible to consider using differing appraisal systems for different study designs, this is likely to be confusing and impractical in reality. Specialist societies would be better advised to select the one that will most effectively address the predominant type of study design being appraised.

Owens 2010 [8]

Context: To establish guidance on grading strength of evidence for the EPC program of the US AHRQ.

Methods: Review of authoritative grading systems, identification of domains and methods that should be considered when grading bodies of evidence, public consultation and discussion with GRADE working group.

Questions:	**Results:**
What are the domains and methods that should be considered when grading bodies of evidence in systematic reviews?	Grading bodies of evidence in systematic reviews requires assessment of four domains: risk of bias, consistency, directness, and precision. Additional domains to be used when appropriate include dose-response association, presence of confounders that would diminish an observed effect, strength of association, and publication bias.
How should EPC staff apply the selected grading system?	The EPC program uses a modified version of GRADE, which is further described in the paper.

Conclusions as formulated by the authors (extracts): No specific conclusions extracted.

Faggion 2010 [9]

Context: To critically describe and evaluate two prominent approaches that might be used to grade levels of evidence and the strength of recommendations in clinical dentistry.

Methods: Selection and appraisal of two grading systems (i.e. GRADE, SORT) based on the following criteria: Separation of grades of recommendations from quality of evidence, Simplicity and transparency of use, Explicit methodology, Consistent with general trends in grading systems, Explicit approach to different levels of evidence for different outcomes

Questions:	**Results:**
What are the strengths and weaknesses of each system?	Both systems seem to fulfil the criteria for an optimum grading system for clinicians GRADE: can offer a more robust picture of the grade of current evidence because the quality of evidence is not only dependent on study design, However, clinicians (mainly new professionals) might initially face difficulties in using the system, because a good understanding of weighting all factors when grading evidence is necessary. SORT: The main criteria for determining the level of evidence and the strength of a recommendation (type of evidence – disease or patient-oriented) might facilitate use of this system in clinical dentistry. Currently, most evidence on dental treatments is disease-oriented and determination of the weakest grades (evidence level C, recommendation C) by the clinician is straightforward.

Conclusions as formulated by the authors (extracts): no specific conclusions extracted.

Ali 2009 [10]

Context: Study ordered by the New Zealand Ministry of Health as to help determine weighting or scoring that should be placed on results of an analysis when making a funding decision. The study was conducted by the Health Services Assessment Collaboration.

Methods: Briefing report based on a systematic search strategy. No systematic review methods were applied for other steps in the report.

Questions:	**Results:**
1. Which are the most commonly used tools for grading of evidence in New Zealand?	The New Zealand Guidelines Group uses a self-created evidence grading system that is described in their handbook.

2. What are the most commonly used tools internationally as reported by literature?	The report refers to the following publications -First results of the CADTH study (Bai 2012, summarized above) -Results of an appraisal of evidence grading systems by SIGN (Baker 2010, summarized above). -A study by Palda et al from 2007 that compared three systems: GRADE Strengths: Working group is an international collaboration interested in developing a common grading system to address limitations and draw on strengths of existing systems. System sequentially assesses quality of evidence, balance between risks and benefits, and judgment about the strength of recommendations. Weaknesses: Application is complicated, Developers use formulaic approaches to global judgments about evidence. SIGN Strengths: Represents a collaboration to improve the quality of health care for patients in Scotland by reducing variation in practice and outcomes, through the development and dissemination of national clinical guidelines. Levels of evidence depend on type and quality of study design. “Considered judgment” forms are used to help guideline development if decisions must be made according to experience as well as knowledge of evidence and underlying methods; forms address quantity, quality and consistency of evidence, generalisability of study findings, directness and clinical impact. Weaknesses: System lacks transparency; no rationale provided to clarify which factors are weighted more heavily for any particular recommendation. Use of numbers and letters may not be intuitive. SORT Strengths: Developed by the US family medicine and primary care journals and the Family Practice Inquiries Network to address the need for a single consistently applied taxonomy of evidence. Emphasizes patient-oriented outcomes.

	Weaknesses: Limited guidance for developers on how to classify studies within numeric categories (1, 2 or 3). Use of numbers and letters may not be intuitive. -A paper by Schünemann et al from 2006 that provided background for advice to WHO Advisory Committee on Health Research. This paper addressed the following questions: 1. Should WHO grade the quality of evidence? 2. What criteria should be used to grade evidence? 3. Should WHO use the same grading system for all of its recommendations? Taking into account this advice, WHO has decided to use GRADE for grading the quality of evidence and strength of recommendations in their guidelines. -A study by Atkins et al from 2004 that evaluated six evidence grading systems (ACCP, OCEBM, NHMRC, SIGN, USPSTF, USTFCPS) as part of the work of the GRADE working group. The working group found that there was poor agreement about the sense of the systems; all of the systems used were considered to have important shortcomings when attempting to grade levels of evidence and the strength of clinical recommendations. The OCEBM system worked well for all four types of questions (studies of diagnosis, effectiveness, harm, and prognosis) considered for the appraisal, although it was not without its faults. -A review by AHRQ in 2002 where the authors identified seven systems that fully addressed all three domains for grading the strength of a body of evidence. This report was updated by Bai et al and findings are described higher in this table.
Conclusions as formulated by the authors (extracts): The evidence grading tools that are more frequently used and highly rated worldwide are.(in alphabetical order) OCEBM, GRADE, NICE and SIGN. There is significant heterogeneity among different 'interest groups'. There is, therefore, a need for a uniform system of grading the rapidly generated evidence so that it can be effectively utilized in clinical practice. This review identified the following several desirable attributes of a grading system: ease of use, perceived quality or validity of the grading system, and clarity of the output or time taken.	

Acronyms: see separate section below.

4 Discussion

Key findings

The identified reviews show that the use of evidence grading systems has become an important step in conducting evidence synthesis. Multiple organisations consider it as an essential process to accurately and transparently move from research findings to conclusions

and to communicate certainty or uncertainty about the effect estimates of healthcare interventions. The goals is to help policymakers, healthcare providers and patients make well-informed decisions. One review warned about the limitations of current grading systems in general and the potentially serious consequences of inaccurate grading, misinterpretation of grading or of misuse of evidence grades.[2]

Many different evidence grading systems exist and uptake of the systems varies. Multiple institutes active with the development of HTAs and guidelines have appraised the available grading systems in order to identify the most sensible approach. From the studies included in this review, we identified a number of desirable attributes for evidence grading systems. In the table below, we use these attributes to summarize the findings of this scoping study.

Desirable attribute	**Summary of Findings**
Consideration of at least quality, quantity and consistency of studies for a given question	Studies by AHRQ and CADTH identified 10 evidence grading systems that adequately incorporate these domains. EPC AHRQ and the GRADE working group defined the following desired domains: risk of bias, consistency, directness, and precision. Additional domains to be used when appropriate include dose-response association, presence of confounders that would diminish an observed effect, strength of association, and publication bias.
Perceived quality or validity of the grading system	Input from experts on the prominent grading systems lead CADTH to the selection of GRADE. WHO also decided to use GRADE in their guideline programme. SIGN has switched to using GRADE and NICE switched to using GRADE for its guideline programme. One study concluded that the majority of grading systems are not validated.
Reproducibility of evidence grading judgements	Reliability studies that were based on older versions of GRADE indicated that reproducibility of judgements could be improved. Detailed guidance and supporting tools have been developed since then.
Potential to use the same system for every type of question (studies of diagnosis, effectiveness, harm, prognosis and public health questions)	OCEBM, GRADE and NSF-LTC were mentioned as systems that can be applied for a diverse set of questions. Although their might not be a system that works well for every question. Instead of using different grading systems for each type of question, it has been suggested to select one system that will

	most effectively address the predominant type of study design(s) being appraised. One evaluation lead to the selection of a domain specific grading system for nursing. The criteria included applicability to non-RCTs for which the authors gave GRADE a lower score.[6]
Ease of use and time required to use the system.	GRADE was mentioned as a more complicated system, but the availability of supporting tools might remediate this. Time required to use the system was not addressed in the identified studies.
Clarity of the output	Systems that use letters and numbers to grade the evidence might be less intuitive.
Amount of uptake of the grading system internationally	GRADE and its modified versions appears to be the system that is most commonly used. This finding is based on a sample of published of systematic reviews and HTA's and a survey among INAHTA partners.[1] An advantage with GRADE is that it is the product of an international working group.
Continuous improvement process	The presence of a continuous improvement process was one of the reasons why CADTH selected GRADE.

Although there is no strong direct evidence, it appears from the table above that GRADE has multiple advantages over other systems. Its complexity and uncertainty in relation to reproducibility of judgements are weaknesses, but this might be compensated by the availability of supportive tools.

An additional element is the call for harmonizing evidence grading systems as to establish a uniform approach. Advantages of such a uniform approach includes that confusion about how to interpret the grades could be eliminated and that it would make it impossible to select the grading system that would lead to the highest evidence grades for a specific question.

Today, GRADE and its modified versions appear to be the most often used evidence grading system for systematic reviews and HTA's. While the modified versions might help address specific needs, such modifications also undermine the goal of achieving a uniform approach internationally.[3] Within the appendix we provide an overview of the modifications to GRADE that were described in the included studies.

Strengths and limitations of this study

The search and extraction process of this scoping review was not as comprehensive as a full systematic review. However, for the purpose of this scoping study, we did not consider this limitation to be significant.

A strength of this review of reviews is that the total of all the reviews allows a multi-perspective summary about the diversity of desirable features for evidence grading systems and how the most important systems perform in relation to these features.

The information presented in this scoping review is limited to the content that was included in the selected reviews. During the search we encountered empirical evaluations of evidence grading systems, which were not included in any of the reviews.[11, 12]

We presented the findings as they were reported in the included studies without any further appraisal. We cannot guarantee that the reviews always presented the evidence grading systems as they are intended by the developers of the grading systems.

Some of the findings might also be outdated. Baker et al tried to identify the best appraisal system for different research fields (Therapy, Screening, Diagnosis, Prognosis, Causation, Psychometric studies, Qualitative studies).[7] However, in the meantime important work has occurred. For example, GRADE has now also developed approaches to appraise qualitative studies and prognostic studies. Another example is that some reviews mentioned the SIGN system for grading evidence. In the meantime, SIGN has decided to use GRADE as their grading system.

In some situations, we felt that the findings and conclusions were not fully in line. For example, Corabian et al found that inter-rater reliability for GRADE and its modified versions varied from slight to almost perfect agreement.[1] However, in the conclusions they state that reliability scores are not acceptable.

5 Conclusions

Most agree that evidence grading systems offer important benefits when conducting evidence synthesis and when communicating the findings. In the past 10 years, various organisations conducted evaluation studies of grading systems in order to identify which approach would serve their activities best. The included studies formulated desirable attributes for evidence grading systems such as quality and validity, applicability to various research questions, ease of use, understandibility, uptake internationally and the availability of a continuous improvement process.

GRADE appears to be the system that is most often evaluated favourably and that is most commonly used by organisations that are active with HTAs, systematic reviews and clinical practice guidelines. Some organisations have chosen to use GRADE with modifications, although this is an obstacle in the move to more global uniformity.

Acronyms

ACCP	American College of Clinical Pharmacy
AGREE	Appraisal of Guidelines for Research and Evaluation
AHRQ	Agency for Healthcare Research and Quality (United States)
AORN	Association of periOperative Registered Nurses
CADTH	Canadian Agency for Drugs and Technologies in Health
COMPUS	Canadian Optimal Medication Prescribing and Utilization Service
EGAPP	Evaluation of Genomic Applications in Practice and Prevention initiative
EPC	Evidence-based Practice Center
GATE	Graphic Appraisal Tool for Epidemiology
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HTA	Health Technology Assessment
INAHTA	International Network of Agencies for Health Technology Assessment
NHMRC	National Health and Medical Research Council (Australia)
NICE	National Institute of Clinical Excellence
NSF-LTC	National Service Framework for Long Term Conditions
OCEBM	Oxford Centre for Evidence Based Medicine
ONS	Oncology Nursing Society
RCT	Randomised Controlled Trial
SIGN	Scottish Intercollegiate Guideline Network
SORT	Strength of Recommendation Taxonomy
USPSTF	United States Preventive Services Task Force
USTFCPS	United States Task Force on Community Preventive Services

References

1. Corabian P, Tjosvold L, Harstall C. Evidence grading systems used in health technology assessment practice. Edmonton (AB): Institute of Health Economics; 2018.
2. Irving M, Eramudugolla R, Cherbuin N, Anstey KJ. A critical review of grading systems: implications for public health policy. Evaluation & the health professions. 2017;40(2):244-62.
3. Andreyeva N, Rebrova O, Zorin N, Avxentyeva M, Omelyanovsky V. Systems for Grading Evidence and Recommendations: Comparison and Prospects for Unification. Medical Technologies. 2012.
4. Gopalakrishna G, Langendam MW, Scholten RJ, Bossuyt PM, Leeflang MM. Guidelines for guideline developers: a systematic review of grading systems for medical tests. Implementation science. 2013;8(1):78.
5. Bai A, Shukla V, Bak G, Wells G. Quality assessment tools project report. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2012. Report No.: 1897465882.
6. Steelman VM, Pape T, King CA, Graling P, Gaberson KB. Selection of a method to rate the strength of scientific evidence for AORN recommendations. AORN journal. 2011;93(4):433-44.
7. Baker A, Young K, Potter J, Madan I. A review of grading systems for evidence-based guidelines produced by medical specialties. Clinical medicine. 2010;10(4):358-63.
8. Owens DK, Lohr KN, Atkins D, Treadwell JR, Reston JT, Bass EB, et al. Grading the strength of a body of evidence when comparing medical interventions-Agency for Healthcare Research and Quality and the Effective Health Care Program. J Clin Epidemiol. 2010;63(5):513-23.
9. Faggion Jr CM. Grading the quality of evidence and the strength of recommendations in clinical dentistry: a critical review of 2 prominent approaches. Journal of Evidence Based Dental Practice. 2010;10(2):78-85.
10. Ali W. What assessment tools are used both in New Zealand and in other countries for grading of evidence? HSAC Report. Health Services Assessment Collaboration, University of Canterbury; 2009.
11. Roos C, Borowiack E, Kowalska M, Zapalska A, Mol B, Mignini L, et al. What do we know about tocolytic effectiveness and how do we use this information in guidelines? A comparison of evidence grading. BJOG: An International Journal of Obstetrics & Gynaecology. 2013;120(13):1588-98.
12. Cooper NA, Khan KS, Clark TJ. Evidence quality in clinical guidelines: a comparison of two methods. Acta obstetricia et gynecologica Scandinavica. 2015;94(12):1283-9.

Appendix

Overview of GRADE modifications

In the table below, we describe which GRADE modifications were mentioned in the reports that we selected for this scoping study. Under GRADE modifications we also include partial use of the system.

This summary does not account for organisations that may have discussed GRADE and eventually decided to adopt it without modifications.

While some evidence grading systems have comparable elements, this overview only includes those systems for which it is explicitly stated that they are a modified GRADE version or that they used parts of the GRADE approach.

AHRQ ECP (as described in Corabian 2018,[1] Owens 2010[8])
Context: The AHRQ EPC approach is primarily used for systematic reviews on effectiveness of preventive and therapeutic interventions, and may relate to research on diagnostic tests, screening strategies, and health services interventions, as well as effects of exposures (characteristics or risk factors) on health outcomes. Diverse stakeholders use EPC reviews for developing guidelines or making clinical or health policy decisions and they may have quite different views on how much, or little, the evidence applies to populations of interest to them. **Modifications:** The differences between the two approaches involve some terminology, purposes of grading the body of evidence, and instructions on how to assess evidence domain characteristics. The AHRQ EPC guidance is designed to separate the raters of the strength of evidence from the decision-makers. EPC researchers grade the strength of evidence for individual outcomes not across outcomes, and do not make or grade clinical recommendations. EPCs may either move up the initial rating of strength of evidence based on observational studies to moderate or move down the initial rating based on RCTs to moderate or low. EPCs can take into account criteria other than those specified by GRADE in assessing the risk of bias of observational (nonrandomized) studies to moderate, but changing the assessment of observational studies for risk of bias (from low to moderate) should be done judiciously. A wide array of groups use EPC reports and the context and populations these users consider relevant may differ. For this reason, the AHRQ EPC approach has chosen to make judgments about applicability explicit and separate from assessments of other domains of strength of evidence. In doing so, AHRQ EPC aim to make it clear when statements about the evidence are based on applicability rather than on other aspects of the evidence. GRADE also addresses applicability and incorporates it within the general concept of directness. EPC reports will have a discussion and information about applicability, and the intention is for the various users and audiences to read this section of the report and make their own judgments. In the EPC approach, the directness evaluation is limited to appraising if intermediate or surrogate outcomes were used

instead of ultimate health outcomes and if more than one body of evidence is required to link interventions to the most important health outcomes (e.g. studies of A vs. C, B vs C, but no studies on A vs B)

*A detailed comparison of the EPC and GRADE approaches is available in the EPC methods handbook.

NICE (as described in Ali 2009[10])

Context: Questions about interventions in NICE clinical guidelines. NICE recommends GRADE as the first approach to quality assessment for all guidelines, including those covering public health and social care topics. NICE recommends GRADE-CERQual for qualitative evidence.

Modifications:
The approach taken by NICE differs from the standard GRADE and GRADE-CERQual system in 2 ways:
It also integrates a review of the quality of cost-effectiveness studies,
It does not use 'overall summary' labels for the quality of the evidence across all outcomes or for the strength of a recommendation, but uses the wording of recommendations to reflect the strength of the evidence.

In addition, although GRADE does not yet cover all types of review questions, GRADE principles can be applied and adapted to other types of questions.

Further the NICE manual states that any substantial changes, made by the developer, to GRADE should be agreed with NICE staff with responsibility for quality assurance before use. If GRADE or GRADE-CERQual is not appropriate for the evidence review, evidence statements should be included. Examples of where evidence statements may be needed are review questions covering prognosis/clinical prediction models (where data cannot be pooled), review questions covering service delivery, or where formal consensus approaches have been taken to answer a review question.

**Description supplemented with info from NICE guideline manual.*

Cochrane Collaboration (as described in Ali 2009[10])

Context: Development of Cochrane systematic reviews

Modifications: The Cochrane Collaboration has adopted the principles of the GRADE system for evaluating the quality of evidence for outcomes reported in systematic reviews. This assessment was phased in together with the introduction of the 'Summary of findings' table.

** Description supplemented with additional info from Cochrane Handbook.*

Other organizations (as described in Ali 2009[10])

Context: Diverse organisations that use the GRADE approach.

Modifications: Minor modifications include collapsing low and very low quality evidence into a single category.