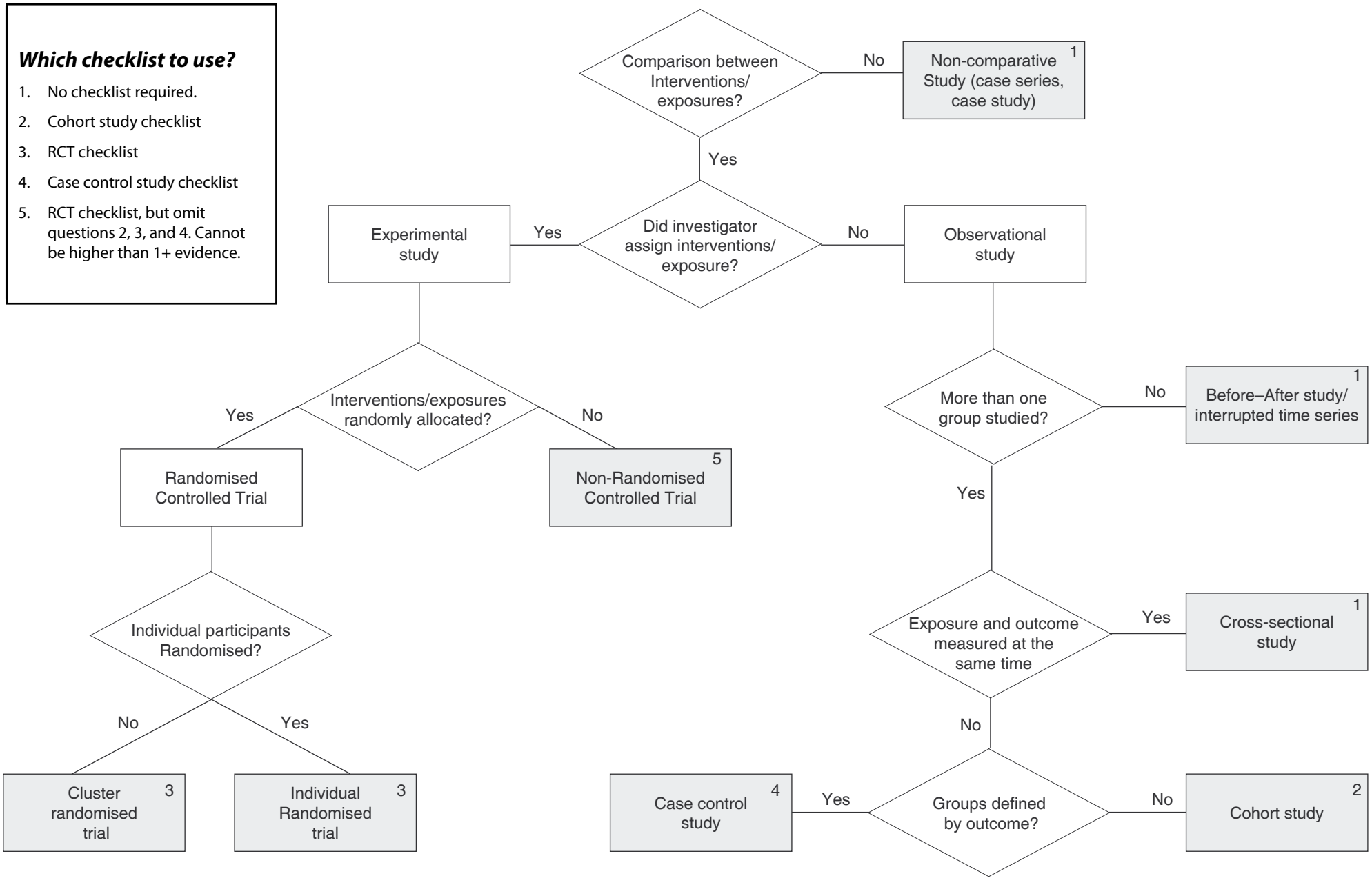
Algorithm for classifying study design for questions of effectiveness



Adapted from NICE (www.nice.org.uk)

Australian National Health and Medical Research Council (ANHMRC)

Table 2. Designation of levels of evidence (Australian NHMRC 199X)

Level of evidence	Study design
I	Evidence obtained from a systematic review of all relevant randomised controlled trials.
II	Evidence obtained from at least one properly-designed randomised controlled trial.
III-1	Evidence obtained from well-designed pseudorandomised controlled trials (alternate allocation or some other method).
III-2	Evidence obtained from comparative studies (including systematic reviews of such studies) with concurrent controls and allocation not randomised, cohort studies, case-control studies, or interrupted time series with a control group.
III-3	Evidence obtained from comparative studies with historical control, two or more single arm studies, or interrupted time series without a parallel control group.
IV	Evidence obtained from case series, either post-test or pretest/post-test.

BMC Health Services Research

Research article

Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches The GRADE Working Group

David Atkins[1], Martin Eccles[2], Signe Flottorp[3], Gordon H Guyatt[4], David Henry[5], Suzanne Hill[5], Alessandro Liberati[6], Dianne O'Connell[7], Andrew D Oxman[3], Bob Phillips[8], Holger Schünemann[4,9], Tessa Tan-Torres Edejer[10], Gunn E Vist*[3], John W Williams Jr[11] and The GRADE Working Group[3]

Address: [1]Center for Practice and Technology Assessment, Agency for Healthcare Research and Quality, 540 Gaither Rd. Rokville, MD 20852, USA, [2]Centre for Health Services Research, University of Newcastle upon Tyne, 21 Claremont Place, Newcastle upon Tyne NE2 4AA, UK, [3]Informed Choice Research Department, Norwegian Health Services Research Centre, Pb. 7004 St. Olavs Plass, 0130 Oslo, Norway, [4]Departments of Clinical Epidemiology and Biostatistics and Medicine, McMaster University, 1200 Main Street West, Hamilton, Ontario L8N 3Z5, Canada, [5]Department of Clinical Pharmacology, Faculty of Medicine and Health Sciences, University of Newcastle, Level 5, New Med 2 Building, Newcastle Mater Hospital, Waratah, NSW 2298, Australia, [6]Department of Oncology and Hematology, Università di Modena e Reggio Emilia, Azienda Ospedaliera Policlinico, Via dal Pozzo 41, 41100 Modena, Italia and Centro per la Valutazione della Efficacia della Assistenza Sanitaria (CeVEAS), Modena, Italy, [7]Cancer Epidemiology Research Unit, Cancer Research and Registers Division, The Cancer Council NSW, PO Box 572, Kings Cross NSW 1340, Australia, [8]Centre for Evidence-based Medicine, University Department of Psychiatry, Warneford Hospital, Oxford OX3 7JX, UK, [9]Departments of Medicine and Social & Preventive Medicine, University at Buffalo, State University of New York, ECMC-CC142, 462 Grinder St, Buffalo, NY 14215, USA, [10]Global Programme on Evidence for Health Policy, World Health Organisation, CH-1211 Geneva 27, Switzerland and [11]The Center for Health Services Research in Primary Care, HSR&D, Department of Veterans Affairs Medical Center and Duke University Medical Center, 508 Fulton St., Durham, NC 27705, USA

Email: David Atkins - DAtkins@AHRQ.GOV; Martin Eccles - Martin.Eccles@newcastle.ac.uk; Signe Flottorp - signe.flottorp@nhsrc.no; Gordon H Guyatt - guyatt@mcmaster.ca; David Henry - mddah@mail.newcastle.edu.au; Suzanne Hill - hillsu@mail.newcastle.edu.au; Alessandro Liberati - alesslib@tin.it; Dianne O'Connell - dianneo@nswcc.org.au; Andrew D Oxman - oxman@online.no; Bob Phillips - bob.phillips@doctors.org.uk; Holger Schünemann - hjs@buffalo.edu; Tessa Tan-Torres Edejer - tantorrest@who.ch; Gunn E Vist* - gev@nhsrc.no; John W Williams - jw.williams@duke.edu; The GRADE Working Group - gev@nhsrc.no

* Corresponding author

Published: 22 December 2004

BMC Health Services Research 2004, **4**:38 doi:10.1186/1472-6963-4-38

Received: 23 January 2004
Accepted: 22 December 2004

This article is available from: http://www.biomedcentral.com/1472-6963/4/38

Abstract

Background: A number of approaches have been used to grade levels of evidence and the strength of recommendations. The use of many different approaches detracts from one of the main reasons for having explicit approaches: to concisely characterise and communicate this information so that it can easily be understood and thereby help people make well-informed decisions. Our objective was to critically appraise six prominent systems for grading levels of evidence and the strength of recommendations as a basis for agreeing on characteristics of a common, sensible approach to grading levels of evidence and the strength of recommendations.

Methods: Six prominent systems for grading levels of evidence and strength of recommendations were selected and someone familiar with each system prepared a description of each of these. Twelve assessors independently evaluated each system based on twelve criteria to assess the sensibility of the different approaches. Systems used by 51 organisations were compared with these six approaches.

Results: There was poor agreement about the sensibility of the six systems. Only one of the systems was suitable for all four types of questions we considered (effectiveness, harm, diagnosis and prognosis). None of the systems was considered usable for all of the target groups we considered (professionals, patients and policy makers). The raters found low reproducibility of judgements made using all six systems. Systems used by 51 organisations that sponsor clinical practice guidelines included a number of minor variations of the six systems that we critically appraised.

Conclusions: All of the currently used approaches to grading levels of evidence and the strength of recommendations have important shortcomings.

Background

In 1979 the Canadian task Force on the Periodic Health Examination published one of the first efforts to explicitly characterise the level of evidence underlying healthcare recommendations and the strength of recommendations [1]. Since then a number of alternative approaches has been proposed and used to classify clinical practice guidelines [2-28].

The original approach used by the Canadian Task Force was based on study design alone, with randomised controlled trials (RCTs) being classified as good (level I) evidence, cohort and case control studies being classified as fair (level II) evidence and expert opinion being classified as poor (level III) evidence. The strength of recommendation was based on the level of evidence with direct correspondence between the two; e.g. a strong recommendation (A) corresponded to there being good evidence. A strength of the original Canadian Task Force approach was that it was simple; the main weakness was that it was too simple. Because of its simplicity, it was easy to understand, apply and present. However, because it was so simple there were many implicit judgements, including judgements about the quality of RCTs, conflicting results of RCTs, and convincing results from non-experimental studies.

For example:

• Should a small, poorly designed RCT be considered level I evidence?

• Should RCTs with conflicting results still be considered level I evidence?

• Should observational studies always be considered level II evidence, regardless of how convincing they are?

The original approach by the Canadian Task Force also did not include explicit judgements about the strength of recommendations, such as how trade-offs between the expected benefits, harms and costs were weighed and taken account of in going from an assessment of how good the evidence is to determining the implications of the results for practice.

The GRADE Working Group is an informal collaboration of people with an interest in addressing shortcomings such as these in systems for grading evidence and recommendations. We describe here a critical appraisal of six prominent systems and the results of the critical appraisal.

Methods

We selected systems for grading the level of evidence and the strength of recommendations that we considered prominent and that included features not captured by other prominent systems. These were selected based on the experience and knowledge of the authors through informal discussion. A description of the most recent version (as of summer 2000) of each of these systems (Appendix 1 to 6), was prepared by one of the authors familiar with the system, and used in this exercise. The following six systems were appraised: the American College of Chest Physicians (ACCP, [see Additional file 1]) [21], Australian National Health and Medical Research Council (ANHMRC, [see Additional file 2]) [17], Oxford Centre for Evidence-Based Medicine (OCEBM, [see Additional file 3]) [16], Scottish Intercollegiate Guidelines Network (SIGN, [see Additional file 4]) [18], US Preventive Services Task Force (USPSTF, [see Additional file 5]) [22], US Task Force on Community Preventive Services (USTFCPS, [see Additional file 6]) [25].

These descriptions of the systems were given to the twelve people who independently appraised the six systems, all

of the authors minus GEV appraised the six systems, three of the authors (DH, SH and DO'C) appraised as a group and reported as one (see contributions). The 12 assessors all had experience with at least one system and most had helped to develop one of the six included systems. Twelve criteria described by Feinstein [29] provided the basis for assessing the sensibility of the six systems.

Criteria used to assess the sensibility of systems for grading evidence and recommendations

1. To what extent is the approach applicable to different types of questions? -effectiveness, harm, diagnosis and prognosis (No, Not sure, Yes)

2. To what extent can the system be used with different audiences? -patients, professionals and policy makers (Little extent, Some extent, Large extent)

3. How clear and simple is the system? (Not very clear, Somewhat clear, Very clear)

4. How often will information not usually available be necessary? (Often, Sometimes, Seldom)

5. To what extent are subjective decisions needed? (Often, Sometimes, Seldom)

6. Are dimensions included that are not within the construct (level of evidence or strength of recommendation)? (Yes, Partially, No)

7. Are there important dimensions that should have been included and are not? (No, Partially, Yes)

8. Is the way in which the included dimensions are aggregated clear and simple? (No, Partially, Yes)

9. Is the way in which the included dimensions are aggregated appropriate? (No. Partially, Yes)

10. Are the categories sufficient to discriminate between different levels of evidence and strengths of recommendations? (No, Partially, Yes)

11. How likely is the system to be successful in discriminating between high and low levels of evidence or strong and weak recommendations? (Not very likely, Somewhat likely, Highly likely)

12. Are assessments reproducible? (Probably not, Not sure, Probably)

No training was provided and we did not discuss the 12 criteria prior to applying them to the six systems.

Our independent appraisal of the six systems were summarised and discussed. The discussion focused on differences in the interpretation of the criteria, disagreement about the judgements that we made and sources of these disagreements, the strengths and weaknesses of the six systems, and inferences based on the appraisals and subsequent discussion.

In order to identify important systems that we might have overlooked following our appraisal of these six systems we also searched the US Agency for Health Care Research and Quality (AHRQ) National Guidelines Clearing House for organisations that have graded two or more guidelines in the Clearing House using an explicit system [30]. These systems were compared with the six systems that we critically appraised.

Results

There was poor agreement among the 12 assessors who independently assessed the six systems. A summary of the assessments of the sensibility of the six approaches to rating levels of evidence and strength of recommendation is shown in Table 1.

Discussion

The poor agreement among the assessors likely reflects several factors. Some of us had practical experience using one of the systems or used additional background information related to one or more grading systems, and we may have been biased in favour of the system with which we were most familiar. Each criterion was applied to grading both evidence and recommendations. Some systems were better for one of these constructs than the other and we may have handled these discrepancies differently. In addition each criterion may have been assessed relative to different judgements about the evidence, such as an assessment of the overall quality of evidence for an important outcome (across studies) versus the quality of an individual study. Some of the criteria were not clear and were interpreted or applied inconsistently. For example, a system might be clear and not simple or visa versa. We likely differed in how stringently we applied the criteria. Finally, there was true disagreement.

There was agreement that the OCEBM system works well for all four types of questions. There was disagreement about the extent to which the other systems work well for questions other than effectiveness. It was noted that some systems are not intended to address other types of questions and it is not clear that it is important that a system should address all four types of questions that we considered (effectiveness, harm, diagnosis, prognosis), although criteria for assessing individual studies must take this into account [31,32].

Table 1: Summary of assessments of the sensibility of six approaches to rating levels of evidence and strength of recommendation

Criteria[1]	ACCP			ANHMRC[2]			USTFCPS			OCEBM			SIGN			USPSTF[3]		
	No		Yes	No		Yes	No		Yes	No		Yes	No		Yes	No		Yes
1. Applicable to different questions:																		
Effectiveness			12		2	8		1	11			12	1		11		2	9
Harm		1	11		5	5	1	7	4		1	11	1	3	8	2	2	7
Diagnosis	7	3	2	4	4	2	9	3				12	5	2	5	2	2	7
Prognosis	6	3	3	2	5	3	9	2	1			11	4	3	5	3	3	5
2. Can be used by:																		
Professionals		1	11	1	5	3		7	4	1	6	5		5	7		3	8
Policy makers	1	5	6	1	5	3	1	2	9	3	7	2	2	6	4	1	4	6
Patients	4	5	3	5	5		6	3	3	9	3		7	5		4	6	1
3. Clear and simple	1	5	6	2	6	1	2	8	2	2	4	5	1	5	6	1	4	7
4. Information not available		8	4	1	5	3	1	6	5		4	8	1	7	4	1	9	2
5. Subjective decisions	2	10		5	2	2	5	5	2		7	5	5	7		2	9	
6. Inappropriate dimensions	1	3	8		1	6	2	4	6		1	10	1	2	8	1	4	6
7. Missing dimensions	1	6	5	2	2	4	5	4	3	9	3	1	5	4	3	2	5	4
Aggregation of dimensions:																		
8. Clear and simple	1	5	6	4	1	2	2	2	7	3	4	4		6	6	2	7	2
9. Appropriate		6	5	3	1	1	3	4	4	2	5	4	1	4	6	1	5	5
10. Sufficient categories	1	4	6	4	2	1		5	7	2	2	7	1	2	9		1	10
11. Likely to discriminate		7	5	2	5	1	1	9	2	2	4	6		5	7		4	7
12. Assessments reproducible	1	8	3	4	4		2	7	2		7	4	1	8	2		10	

[1]See Criteria described in Methods.
[2]Two people did not assess the ANHMRC because it was more descriptive and others responded not applicable for some questions.
[3]One person did not assess the USPST and one person had two responses on questions 3 and 4.

Most of us did not find that any of the systems are likely to be suitable for use by patients. Almost all agreed that the ACCP system was suitable for professionals and most considered that the USPSTF system was suitable for professionals. There was not much agreement about the suitability of any of the other systems for professionals or about the suitability of any of the systems for policy makers, although most assessed the USTFCPS system to be suitable for policy makers.

There was no agreement that any of the systems are clear and simple, although USPSTF, ACCP and SIGN systems were generally assessed more favourably in this regard. It was generally agreed that the clearer a system was the less simple it was; e.g. the OCEBM system is clear but not simple for categorising the level of evidence. There was some confusion regarding whether we were assessing how clear and simple the system was to guideline developers (as some interpreted this criterion) or how clear and simple the outcome of applying the system was to guideline users (as others interpreted this criterion). Either way, the simpler a system is the less clear it is likely to be.

Most of us judged that for most of the systems necessary information would not be available at least sometimes. The OCEBM system came out somewhat better than the other systems and lack of availability of necessary information was considered to be less of a problem for the USTFCPS system. However, the OCEBM and USTFCPS systems were considered by most to be missing dimensions which may, in part, explain why missing information was considered to be less of a problem. This would be the case to the extent the missing dimensions were the ones for which information would often or sometimes not be available. The dimension for which we considered that information would most often be missing was trade-offs; i.e. knowledge of the preferences or utility values of those affected. Additional problems were identified in relationship to complex interventions and counselling, particularly with the USTFCPS and USPSTF systems. It was pointed out that the USTFCPS system addressed this problem by including availability of information about the intervention as part of its assessment of the quality of evidence.

Most of the systems were assessed to require subjective decisions at least to some extent. The OCEBM system again stood out as being assessed more favourably, although it may be related to omission of dimensions that require more subjective decisions. Judgement is clearly needed with any system. The aim should be to make judgements transparent and to try to protect against bias in the judgements that are made by being systematic and explicit.

Inclusion of dimensions that are not within the constructs being graded was not considered a problem for most of the systems by most of us. Several people considered that it might be a problem for the USTFCPS and USPSTF systems. On the other hand, all of the systems were evaluated to be missing at least one important dimension by at least one person. The challenge of missing dimensions were considered less of a problem for the ACCP and ANHMRC systems. There was not agreement about any of the systems having a clear and simple approach to aggregating the dimensions, although this was considered to be less of a problem for the ACCP, SIGN and USTFCPS systems.

There was also not agreement on the appropriateness of how the dimensions were aggregated. This was considered to be more of a problem for the ANHMRC and USTFCPS systems than the other four systems, all of which were considered to have taken an approach to aggregating the dimensions that was at least partially inappropriate by more than half of us.

Most of us considered that most of the systems had sufficient categories, with the exception of the ANHMRC system. There was almost agreement that the USPSTF system has sufficient categories. We agreed that it is possible to have too many categories as well as too few, the OCEBM system being an example of having too many categories.

There was not agreement that any of the systems are likely to discriminate successfully, although everyone thought that the ACCP, SIGN and USPSTF systems are somewhat to highly likely to discriminate. Lastly, we largely agreed that we were not sure how reproducible assessments are using any of the systems, although half of us considered that assessments using the ANHMRC system are unlikely to be reproducible and about 1/3 considered that assessments using the OCEBM and ACCP systems are likely to be reproducible.

We identified 22 additional organisations that have produced 10 or more practice guidelines using an explicit approach to grade the level of evidence or strength of recommendations. Another 29 have produced between two and nine guidelines using an explicit approach. These systems include a number of minor variations of the six systems that we appraised in detail.

There was generally poor agreement between the individual assessors about the scoring of the six approaches using the 12 criteria. However, there was general agreement that none of these six prominent approaches to grading the levels of evidence and strength of recommendations adequately addressed all of the important concepts and dimensions that we thought should be considered. Although we limited our appraisal to six systems all of the additional approaches to grading levels of evidence and strength of recommendations that we identified were, in essence, variations of the six approaches that we had critically appraised. Therefore we are confident that we did not miss any important grading systems available at the time when these assessments were undertaken.

Based on discussions following the critical appraisal of these six approaches, we agreed on some conclusions:

• Separate assessments should be presented for judgements about the quality of the evidence and judgements about the balance of benefits and harms.

• Evidence for harms should be assessed in the same way as evidence for benefits, although different evidence may be considered relevant for harms than for benefits; e.g. local evidence of complication rates may be considered more relevant than evidence of complication rates from trials for endarterectomy.

• Judgements about the quality of evidence should be based on a systematic review of the relevant research.

• Systematic reviews should not be included in a hierarchy of evidence (i.e. as a level or category of evidence). The availability of a well-done systematic review does not correspond to high quality evidence, since a well-done review might include anything from no studies to poor quality studies with inconsistent results to high quality studies with consistent results.

• Baseline risk should be taken into consideration in defining the population to whom a recommendation applies. Baseline risk should also be used transparently in making judgements about the balance of benefits and harms. When a recommendation varies in relationship to baseline risk, the evidence for determining baseline risk should be assessed appropriately and explicitly.

• Recommendations should not vary in relationship to baseline risk if there is not adequate evidence to guide reliable determinations of baseline risk.

Conclusions

Based on discussions of the strengths and limitations of current approaches to grading levels of evidence and the strength of recommendations, we agreed to develop an approach that addresses the major limitations that we identified. The approach that the GRADE Working Group has developed is based on the discussions following the critical appraisal reported here and a pilot study of the GRADE approach [33]. Based on the pilot testing and the discussions following the pilot, the GRADE Working Group has further developed the GRADE system to its present format [34].

The GRADE Working Group has continued to grow as an informal collaboration that meets one or two times per year. The group maintains web pages http://www.gradeworkinggroup.org and a discussion list.

Competing interests

DA has competing interests with the US Preventive Services Task Force (USPSTF), PAB has a competing interest with the US Task Force on Community Preventive Services (USTFCPS), GHG and HS have competing interests with the American College of Chest Physicians (ACCP), DH, SH and DO'C have competing interests with the Australian National Health and Medical Research Council (ANHMRC), BP has competing interests with the Oxford Centre for Evidence-Based Medicine (OCEBM). Most of the other members of the GRADE Working Group have experience with the use of one or more systems of grading evidence and recommendations.

Contributions

DA, PAB, ME, SF, GHG, DH, SH, AL, DO'C, ADO, BP, HS, TTTE, GEV & JWW Jr as members of the GRADE Working Group have contributed to the preparation of this manuscript and the development of the ideas contained herein, participated in the critical assessment, and read and commented on drafts of this article. GHG and ADO have led the process. GEV has had primary responsibility for coordinating the process.

Additional material

Additional File 1

American College of Chest Physicians (ACCP), a brief description of the ACCP approach.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6963-4-38-S1.doc]

Additional File 2

Australian National Health and Medical Research Council (ANHMRC), a brief description of the ANHMRC approach.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6963-4-38-S2.doc]

Additional File 3

Oxford Centre for Evidence-based Medicine (OCEBM), a brief description of the OCEBM approach.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6963-4-38-S3.doc]

Additional File 4

Scottish Intercollegiate Guidelines (SIGN), a brief description of the SIGN approach.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6963-4-38-S4.doc]

Additional File 5

U.S. Preventive Services Task Force (USPSTF), a brief description of the USPSTF approach.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6963-4-38-S5.doc]

Additional File 6

U.S. Task Force on Community Preventive Services (USTFCPS), a brief description of the USTFCPS approach.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6963-4-38-S6.doc]

Acknowledgements

We wish to thank Peter A Briss for participating in the critical assessment and for providing constructive comments on the process. The institutions with which members of the Working Group are affiliated have provided intramural support. Opinions expressed in this paper do not necessarily represent those of the institutions with which the authors are affiliated.

References

1. Canadian Task Force on the Periodic Health Examination: **The periodic health examination.** *Can Med Assoc J* 1979, **121:**1193-254.
2. Sackett DL: **Rules of evidence and clinical recommendations on the use of antithrombotic agents.** *Chest* 1986, **89(suppl 2):**2S-3S.
3. Sackett DL: **Rules of evidence and clinical recommendations on the use of antithrombotic agents.** *Archives Int Med* 1986, **146:**464-465.
4. Sackett DL: **Rules of evidence and clinical recommendations on the use of antithrombotic agents.** *Chest* 1989, **95:**2S-4S.
5. Cook DJ, Guyatt GH, Laupacis A, Sackett DL: **Rules of evidence and clinical recommendations on the use of antithrombotic agents. Antithrombotic Therapy Consensus Conference.** *Chest* 1992, **102(suppl 4):**305S-311S.
6. US Department of Health and Human Services, Public Health Service, Agency Health Care Policy and Research: **Acute Pain Management: Operative or Medical Procedures and Trauma.** *Agency for Health Care Policy and Research Publications, Rockville, MD. (AHCPR Pub 92-0038)* 1992.

7. Gyorkos TW, Tannenbaum TN, Abrahamowicz M, Oxman AD, Scott EA, Millson ME, Rasooly I, Frank JW, Riben PD, Mathias RG: **An approach to the development of practice guidelines for community health interventions.** *Can J Public Health* 1994, **85(suppl 1):**S8-S13.
8. Hadorn DC, Baker D: **Development of the AHCPR-sponsored heart failure guideline: methodologic and procedural issues.** *J Quality Improvement* 1994, **20:**539-54.
9. Cook DJ, Guyatt GH, Laupacis A, Sackett DL, Goldberg RJ: **Clinical recommendations using levels of evidence for antithrombotic agents.** *Chest* 1995, **108(4 Suppl):**227S-230S.
10. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ, for the Evidence-Based Medicine Working Group: **User's guides to the medical literature.IX. A method for grading health care recommendations. Evidence-Based medicine working group.** *JAMA* 1995, **274:**1800-4.
11. Petrie J, Barnwell E, Grimshaw J: **Criteria for appraisal for national use. Pilot Edition.** *Scottish Intercollegiate Guidelines Network* 1995 [http://www.sign.ac.uk/methodology/index.html].
12. US Preventive Services Task Force: **Guide to Clinical Preventive Services.** 2nd edition. Baltimore: Williams & Wilkins; 1996:xxxix-lv.
13. Eccles M, Clapp Z, Grimshaw J, Adams PC, Higgins B, Purves I, Russell I: **North of England evidence based guidelines development project: methods of guideline development.** *BMJ* 1996, **312:**760-2.
14. *Centro per la Valutazione della Efficacia della Assistenza Sanitaria (CeVEAS). Linee Guida per il trattamento del tumore della mammella nella provincia di Modena (Luglio 2000)* [http://www.ceveas.it/ceveas/viewpage.do?idp=3]. accessed December 29, 2002
15. Guyatt GH, Cook DJ, Sackett DL, Eckman M, Pauker S: **Grades of recommendation for antithrombotic agents.** *Chest* 1998, **114(5 Suppl**441S-4S [http://www.chestjournal.org/content/vol119/1_suppl/].
16. Ball C, Sackett D, Phillips B, Straus S, Haynes B: **Levels of evidence and grades of recommendations. Last revised 17 September 1998.** [http://www.cebm.net/levels_of_evidence.asp]. Centre for Evidence-Based Medicine
17. National Health and Medical Research Council: **How to use the evidence: assessment and application of scientific evidence.** *Commonwealth of Australia* 2000 [http://www.nhmrc.gov.au/publications/synopses/cp65syn.htm].
18. Harbour R, Miller J: **A new system for grading recommendations in evidence based guidelines.** *BMJ* 2001, **323:**334-6.
19. Roman SH, Silberzweig SB, Siu AL: **Grading the evidence for diabetes performance measures [see comments].** *Eff Clin Pract* 2000, **3:**85-91.
20. Woloshin S: **Arguing about grades.** *Eff Clin Pract* 2000, **3:**94-5.
21. Guyatt GH, Schünemann H, Cook D, Pauker S, Sinclair J, Bucher H, Jaeschke R: **Grades of recommendation for antithrombotic agents.** *Chest* 2001, **119:**3S-7S.
22. Atkins D, Best D, Shapiro EN: **The third U.S. Preventive Services Task Force : background, methods and first recommendations.** *Am J Preventive Medicine* 2001, **20(3 (supplement 1)):**1-108.
23. Woolf SH, Atkins D: **The evolving role of prevention in health care: Contributions of the U.S. Preventive Services Task Force.** *Am J Preventive Medicine* 2001, **20(3 (supplement 1)):**13-20.
24. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, Atkins D, for the Methods Work Group of the Third U.S. Preventive Services Task Force: **Current methods of the U.S. Preventive Services Task Force: A review of the process.** *Am J Preventive Medicine* 2001, **20(3 (Supplement 1)):**21-35.
25. Briss PA, Zaza S, Pappaioanou M, Fielding J, Wright-De Aguero L, Truman BI, Hopkins DP, Mullen PD, Thompson RS, Woolf SH, Carande-Kulis VG, Anderson L, Hinman AR, McQueen DV, Teutsch SM, Harris JR: **Developing an evidence-based Guide to Community Preventive Services – methods. The Task Force on Community Preventive Services.** *Am J Preventive Medicine* 2000, **18:**35-43.
26. Zaza S, Wright-De Aguero LK, Briss PA, Truman BI, Hopkins DP, Hennessy MH, Sosin DM, Anderson L, Carande-Kulis VG, Teutsch SM, Pappaioanou M: **Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Task Force on Community Preventive Services.** *American Journal of Preventive Medicine* 2000, **18:**44-74.
27. Greer N, Mosser G, Logan G, Halaas GW: **A practical approach to evidence grading.** *Joint Commission J Qual Improv* 2000, **26:**700-12.
28. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, Lux L: **Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011).** In *AHRQ Publication No. 02-E016* Rockville, MD: Agency for Healthcare Research and Quality; 2002:64-88.
29. Feinstein AR: *Clinimetrics* New Haven, CT: Yale University Press; 1987:141-66.
30. **National Guidelines Clearing House** [http://www.guideline.gov/resources/guideline_index.aspx]. Accessed April 19, 2001
31. Guyatt G, Drummond R, eds: **Users' Guide to the Medical Literature.** Chicago, IL: AMA Press; 2002:55-154.
32. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, Lux L: **Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011).** In *AHRQ Publication No. 02-E016* Rockville, MD: Agency for Healthcare Research and Quality; 2002:51-63.
33. Atkins D, Briss PA, Eccles M, Flottorp S, Guyatt GH, Harbour RT, Hill S, Jaeschke R, Liberati A, Magrini N, Mason J, O'Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer TT, Vist GE, Williams JW Jr, GRADE Working Group: **Systems for grading the quality of evidence and the strength of recommendations II: Pilot study of a new system.** *BioMed Central* in press.
34. Atkins D, Best D, Briss PA, Eccles M, Falck Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini N, Mason J, Middleton P, Mrukowicz J, O'Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer TT, Varonen H, Vist GE, Williams JW Jr, Zaza S, Grade Working Group: **Grading quality of evidence and strength of recommendations.** *BMJ* **328(7454):**1490. 2004 Jun 19

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1472-6963/4/38/prepub

Levels of Evidence For Primary Research Question[1]
As Adopted by the North American Spine Society January 2005*

	Types of Studies			
	Therapeutic Studies – Investigating the results of treatment	Prognostic Studies – Investigating the effect of a patient characteristic on the outcome of disease	Diagnostic Studies – Investigating a diagnostic test	Economic and Decision Analyses – Developing an economic or decision model
Level I	• High quality randomized trial with statistically significant difference or no statistically significant difference but narrow confidence intervals • Systematic Review[2] of Level I RCTs (and study results were homogenous[3])	• High quality prospective study[4] (all patients were enrolled at the same point in their disease with ≥ 80% follow-up of enrolled patients) • Systematic review[2] of Level I studies	• Testing of previously developed diagnostic criteria on consecutive patients (with universally applied reference "gold" standard) • Systematic review[2] of Level I studies	• Sensible costs and alternatives; values obtained from many studies; with multiway sensitivity analyses • Systematic review[2] of Level I studies
Level II	• Lesser quality RCT (e.g. < 80% follow-up, no blinding, or improper randomization) • Prospective[4] comparative study[5] • Systematic review[2] of Level II studies or Level 1 studies with inconsistent results	• Retrospective[6] study • Untreated controls from an RCT • Lesser quality prospective study (e.g. patients enrolled at different points in their disease or <80% follow-up.) • Systematic review[2] of Level II studies	• Development of diagnostic criteria on consecutive patients (with universally applied reference "gold" standard) • Systematic review[2] of Level II studies	• Sensible costs and alternatives; values obtained from limited studies; with multiway sensitivity analyses • Systematic review[2] of Level II studies
Level III	• Case control study[7] • Retrospective[6] comparative study[5] • Systematic review[2] of Level III studies	• Case control study[7]	• Study of non-consecutive patients; without consistently applied reference "gold" standard • Systematic review[2] of Level III studies	• Analyses based on limited alternatives and costs; and poor estimates • Systematic review[2] of Level III studies
Level IV	Case Series[8]	Case series	• Case-control study • Poor reference standard	• Analyses with no sensitivity analyses
Level V	Expert Opinion	Expert Opinion	Expert Opinion	Expert Opinion

1. A complete assessment of quality of individual studies requires critical appraisal of all aspects of the study design.
2. A combination of results from two or more prior studies.
3. Studies provided consistent results.
4. Study was started before the first patient enrolled.
5. Patients treated one way (e.g. cemented hip arthroplasty) compared with a group of patients treated in another way (e.g. uncemented hip arthroplasty) at the same institution.
6. The study was started after the first patient enrolled.
7. Patients identified for the study based on their outcome, called "cases"; e.g. failed total arthroplasty, are compared to those who did not have outcome, called "controls"; e.g. successful total hip arthroplasty.
8. Patients treated one way with no comparison group of patients treated in another way.

*These documents have also been adopted by the American Academy of Orthopaedic Surgeons, Pediatric Orthopaedic Society of North America, *Clinical Orthopaedics and Related Research, Journal of Bone & Joint Surgery* and *Spine*. September 28, 2004

Oxford Centre for Evidence-based Medicine – Levels of Evidence (March 2009)

What are we to do when the irresistible force of the need to offer clinical advice meets with the immovable object of flawed evidence? All we can do is our best: give the advice, but alert the advisees to the flaws in the evidence on which it is based.

The CEBM 'Levels of Evidence 1' document sets out one approach to systematising this process for different question types.

(For definitions of terms used see our glossary)

Level	Therapy / Prevention, Aetiology / Harm	Prognosis	Diagnosis	Differential diagnosis / symptom prevalence study	Economic and decision analyses
1a	SR (with homogeneity*) of RCTs	SR (with homogeneity*) of inception cohort studies; CDR" validated in different populations	SR (with homogeneity*) of Level 1 diagnostic studies; CDR" with 1b studies from different clinical centres	SR (with homogeneity*) of prospective cohort studies	SR (with homogeneity*) of Level 1 economic studies
1b	Individual RCT (with narrow Confidence Interval"¡)	Individual inception cohort study with > 80% follow-up; CDR" validated in a single population	Validating** cohort study with good" " " reference standards; or CDR" tested within one clinical centre	Prospective cohort study with good follow-up****	Analysis based on clinically sensible costs or alternatives; systematic review(s) of the evidence; and including multi-way sensitivity analyses
1c	All or none§	All or none case-	Absolute SpPins	All or none	Absolute better-

		series	and SnNouts” “	case-series	value or worse-value analyses ” ” ” “
2a	SR (with homogeneity*) of cohort studies	SR (with homogeneity*) of either retrospective cohort studies or untreated control groups in RCTs	SR (with homogeneity*) of Level >2 diagnostic studies	SR (with homogeneity*) of 2b and better studies	SR (with homogeneity*) of Level >2 economic studies
2b	Individual cohort study (including low quality RCT; e.g., <80% follow-up)	Retrospective cohort study or follow-up of untreated control patients in an RCT; Derivation of CDR” or validated on split-sample§§§ only	Exploratory** cohort study with good” ” ” reference standards; CDR” after derivation, or validated only on split-sample§§§ or databases	Retrospective cohort study, or poor follow-up	Analysis based on clinically sensible costs or alternatives; limited review(s) of the evidence, or single studies; and including multi-way sensitivity analyses
2c	“Outcomes” Research; Ecological studies	“Outcomes” Research		Ecological studies	Audit or outcomes research
3a	SR (with homogeneity*) of case-control studies		SR (with homogeneity*) of 3b and better studies	SR (with homogeneity*) of 3b and better studies	SR (with homogeneity*) of 3b and better studies
3b	Individual Case-Control Study		Non-consecutive study; or without consistently applied reference standards	Non-consecutive cohort study, or very limited population	Analysis based on limited alternatives or costs, poor quality estimates of data, but including

					sensitivity analyses incorporating clinically sensible variations.
4	Case-series (and poor quality cohort and case-control studies§§)	Case-series (and poor quality prognostic cohort studies***)	Case-control study, poor or non-independent reference standard	Case-series or superseded reference standards	Analysis with no sensitivity analysis
5	Expert opinion without explicit critical appraisal, or based on physiology, bench research or “first principles”	Expert opinion without explicit critical appraisal, or based on physiology, bench research or “first principles”	Expert opinion without explicit critical appraisal, or based on physiology, bench research or “first principles”	Expert opinion without explicit critical appraisal, or based on physiology, bench research or “first principles”	Expert opinion without explicit critical appraisal, or based on economic theory or “first principles”

Produced by Bob Phillips, Chris Ball, Dave Sackett, Doug Badenoch, Sharon Straus, Brian Haynes, Martin Dawes since November 1998. Updated by Jeremy Howick March 2009.

Notes

Users can add a minus-sign “-” to denote the level of that fails to provide a conclusive answer because:

- ***EITHER*** a single result with a wide Confidence Interval
- ***OR*** a Systematic Review with troublesome heterogeneity.

Such evidence is inconclusive, and therefore can only generate Grade D recommendations.

*	By homogeneity we mean a systematic review that is free of worrisome variations (heterogeneity) in the directions and degrees of results between individual studies. Not all systematic reviews with statistically significant heterogeneity need be worrisome, and not all

	worrisome heterogeneity need be statistically significant. As noted above, studies displaying worrisome heterogeneity should be tagged with a "-" at the end of their designated level.
"	Clinical Decision Rule. (These are algorithms or scoring systems that lead to a prognostic estimation or a diagnostic category.)
"¡	See note above for advice on how to understand, rate and use trials or other studies with wide confidence intervals.
§	Met when all patients died before the Rx became available, but some now survive on it; or when some patients died before the Rx became available, but none now die on it.
§§	By poor quality cohort study we mean one that failed to clearly define comparison groups and/or failed to measure exposures and outcomes in the same (preferably blinded), objective way in both exposed and non-exposed individuals and/or failed to identify or appropriately control known confounders and/or failed to carry out a sufficiently long and complete follow-up of patients. By poor quality case-control study we mean one that failed to clearly define comparison groups and/or failed to measure exposures and outcomes in the same (preferably blinded), objective way in both cases and controls and/or failed to identify or appropriately control known confounders.
§§§	Split-sample validation is achieved by collecting all the information in a single tranche, then artificially dividing this into "derivation" and "validation" samples.
" "	An "Absolute SpPin" is a diagnostic finding whose Specificity is so high that a Positive result rules-in the diagnosis. An "Absolute SnNout" is a diagnostic finding whose Sensitivity is so high that a Negative result rules-out the diagnosis.
"¡"¡	Good, better, bad and worse refer to the comparisons between treatments in terms of their clinical risks and benefits.
" " "	Good reference standards are independent of the test, and applied blindly or objectively to applied to all patients. Poor reference standards are haphazardly applied, but still independent of the test. Use of a non-independent reference standard (where the 'test' is included in the 'reference', or where the 'testing' affects the 'reference') implies a level 4 study.
" " " "	Better-value treatments are clearly as good but cheaper, or better at the same or reduced cost. Worse-value treatments are as good and more expensive, or worse and the equally or more expensive.
**	Validating studies test the quality of a specific diagnostic test, based on prior evidence. An exploratory study collects information and trawls the data (e.g. using a regression analysis) to find which factors are 'significant'.
***	By poor quality prognostic cohort study we mean one in which sampling was biased in favour

	of patients who already had the target outcome, or the measurement of outcomes was accomplished in <80% of study patients, or outcomes were determined in an unblinded, non-objective way, or there was no correction for confounding factors.
****	Good follow-up in a differential diagnosis study is >80%, with adequate time for alternative diagnoses to emerge (for example 1-6 months acute, 1 – 5 years chronic)

Grades of Recommendation

A	consistent level 1 studies
B	consistent level 2 or 3 studies ***or*** extrapolations from level 1 studies
C	level 4 studies ***or*** extrapolations from level 2 or 3 studies
D	level 5 evidence ***or*** troublingly inconsistent or inconclusive studies of any level

"Extrapolations" are where data is used in a situation that has potentially clinically important differences than the original study situation.

U.S. Preventative Services Task Force (USPSTF)

4.2.1 Assessing Internal Validity (Quality) of Individual Studies

The Task Force recognizes that research design is an important component of the validity of the information in a study for the purpose of answering a key question. Although RCTs cannot answer all key questions, they are ideal for questions regarding benefits or harms of various interventions. Thus, for the key questions of benefits and harms, the Task Force currently uses the following hierarchy of research design:

I. Properly powered and conducted RCT; well-conducted systematic review or meta-analysis of homogeneous RCTs
II-1. Well-designed controlled trial without randomization
II-2. Well-designed cohort or case-control analysis study
II-3. Multiple time-series, with or without the intervention; results from uncontrolled studies that yield results of large magnitude
III. Opinions of respected authorities, based on clinical experience; descriptive studies or case reports; reports of expert committees

Although research design is an important determinant of the quality of information provided by an individual study, the Task Force also recognizes that not all studies with the same research design have equal internal validity (quality).

To assess more carefully the internal validity of individual studies within research designs, the Task Force has developed design-specific criteria for assessing the internal validity of individual studies. The EPC may supplement these with the use of newer methods of assessing quality of individual studies as appropriate.