Clinical Medicine 2010, Vol 10, No 4: 358–63

A review of grading systems for evidence-based guidelines produced by medical specialties

Adrian Baker, Katharine Young, Jonathan Potter and Ira Madan

ABSTRACT – The development of evidence-based guidelines requires scrupulous attention to the method of critical appraisal. Many critical appraisal systems give 'gold standard' status to randomised controlled trials (RCTs) due to their ability to limit bias. While guidelines with a prominent research base consisting of RCTs have been well served by such systems, specialist societies with research bases consisting of a wide range of study designs have been at a disadvantage, potentially leading to inappropriately low grades being given for recommendations. A review of the Scottish Intercollegiate Guidelines Network, the Grading of Recommendations Assessment, Development and Evaluation, the Graphic Appraisal Tool for Epidemiology and the National Service Framework for Long Term Conditions grading systems was therefore undertaken. A matrix was developed suggesting the optimum grading system for the type of guideline being developed or question being addressed by a specialist society.

KEY WORDS: clinical effectiveness, critical appraisal, evidence-based medicine, grading systems

Introduction

The production of an evidence-based guideline requires a systematic review and critical appraisal of the literature relevant to the scope of the guideline, as guideline recommendations are graded on the strength of evidence on which they are based. The plethora of grading systems available, make it difficult for guideline developers to choose which system to adopt resulting in different guidelines using different systems and confusion among users.

The problem that guideline developers face is that the majority of grading hierarchies are created with randomised controlled trials (RCTs) as the 'gold standard' due to their ability to reduce possible study biases and confounders. However, guidelines developed by specialist societies often pose questions on prognosis or patient's views, rather than on the effectiveness of pharmacological interventions; these questions are best answered by observational studies or qualitative research. This over reliance on RCTs being at the top of the evidence pinnacle often results in specialist society-based guidelines assigning inappropriately low grades to their recommendations and hence reducing their legitimacy.

The aims of this paper are to:

- review the strengths and weaknesses of the current major grading systems in the context of their use by specialist societies
- identify the optimum grading system for the type of guideline being developed or question being addressed by the specialist society.

Method

A small working group was formed from members of the Royal College of Physicians' Clinical Effectiveness Forum.[1] The systems that were chosen for review were the Scottish Intercollegiate Guidelines Network (SIGN), the Grading of Recommendations Assessment, Development and Evaluation (GRADE), the Graphic Appraisal Tool for Epidemiology (GATE) and the National Service Framework for Long Term Conditions (NSF-LTC) grading system. The review of SIGN was chosen due to its established use by societies; GRADE, a relatively new system, was chosen due to its perceived methodological rigour and the extensive resources used to produce its appraisal system. A review of the NSF-LTC system was conducted due to its ability to offer an alternative to SIGN and GRADE through its holistic interpretation of medical research. The GATE system was reviewed due to its simplicity, clarity and ability to be used to critically appraise different types of studies.

The review was undertaken in conjunction with discussions with the system developers and technical advisors from the National Institute for Health and Clinical Excellence (NICE), and was signed off by the Clinical Effectiveness Forum and subgroup.[2] The systems were reviewed in the context of their use by a specialist society; where members of the guideline development groups critically appraise the papers forming the evidence review of the guideline and assign grades for the evidence and recommendations. A matrix was then developed to reflect the strengths and weaknesses of each system in relation to each other and in the context of the characteristics of different fields of research.

Adrian Baker, researcher, NHS Plus; **Katharine Young,** clinical standards facilitator, Clinical Standards Department, Royal College of Physicians; **Jonathan Potter,** clinical director, Clinical Effectiveness and Evaluation Unit, Clinical Standards Department, Royal College of Physicians; **Ira Madan,** consultant occupational physician, Guy's and St Thomas's NHS Foundation Trust, London

Results

Scottish Intercollegiate Guidelines Network

SIGN is a widely used critical appraisal and evidence hierarchy which has the advantage of being simple and clear to use and therefore suitable for small or low-resource guideline development groups. The aim of the SIGN system is to ensure that the extent of the internal and external validity of a study is robustly assessed and leads to the final grade for a recommendation. The methodology behind the system is based on a set of variables that recognise key factors, especially bias and confounding, that can influence the quality of a study or its conclusion.

The SIGN methodology includes checklists to critically appraise studies, with one checklist for each of the following study types: systematic reviews and meta-analyses, RCTs, cohort studies, case-control studies, diagnostic studies and economic studies. SIGN undertook an evaluation and iterative adaptation process for each checklist. The raison d'être of systematic appraisal is to reduce study and appraiser bias. SIGN emphasise the aspects of study design which can lead to biased results and, importantly, SIGN also acknowledges the direction of that bias. Though the methodology clearly gives the gold standard to RCTs it is recognised that non-randomised studies can strengthen or question the results of RCTs. Overall assessment of the strength of the evidence within each paper is based on a grading criteria of '++', '+' or '−', as illustrated in Table 1.[3]

The final grade given to the evidence is based on the lowest level of evidence applicable to a key outcome produced through assessing the overall body of evidence. The reason for this is to reduce the overstatement to the risk of benefits. The grades given to the recommendations are based on an 'A, B, C, D' system (Table 2).[3] SIGN include two caveats when grading the overall recommendation. Firstly, strong recommendations should usually not be given if they are based on only a small number of studies. Secondly, the grading level does not relate to the importance of the recommendation, but to the quality or strength of the evidence and studies that support it.[3] In essence, this means that the SIGN grading recommendation indicates the likelihood the outcome of the recommendation can be achieved.

Grading of Recommendations Assessment, Development and Evaluation

The GRADE working group argue that confusion is created by differences and inconsistencies in existing critical appraisal systems.[4] The GRADE system was produced through extensive analysis of other systems and alternatives. The aim was to detect and resolve inherent weaknesses in the other systems while including their strengths, and producing a universal, easily understandable and practical system that can be utilised by a wide variety of practice areas in a number of different contexts. There are four levels in grading the overall quality of evidence (Table 3).[5] It is contingent on the lowest quality of all outcomes that are important for making a decision.

Like SIGN, GRADE places observational studies in lower regard than RCTs, but acknowledges that there may be poor RCTs and strong observational studies. Although the study design initially leads to the hierarchical grading of a study, the

Table 1. Scottish Intercollegiate Guidelines Network (SIGN) grades for evidence.

Levels of evidence	
1++	High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
1+	Well conducted meta-analyses, systematic reviews, or RCTs with a low risk of bias
1−	Meta-analyses, systematic reviews, or RCTs with a high risk of bias
2++	High quality systematic reviews of case control or cohort studies. High quality case control or cohort studies with a very low risk of confounding or bias and a high probability that the relationship is causal
2+	Well conducted case control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal
2−	Case control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal
3	Case reports, case series
4	Expert opinion

RCTs = randomised controlled trials.

Table 2. Scottish Intercollegiate Guidelines Network (SIGN) grades for recommendations.

A	At least one meta-analysis, systematic review, or RCT rated as 1++, and directly applicable to the target population; or A body of evidence consisting principally of studies rated as 1+, directly applicable to the target population, and demonstrating consistency of results
B	A body of evidence including studies rated as 2++, directly applicable to the target population, and demonstrating overall consistency of results; or Extrapolated evidence from studies rated as 1++ or 1+
C	A body of evidence including studies rated as 2+, directly applicable to the target population and demonstrating overall consistency of results; or Extrapolated evidence from studies rated as 2++
D	Evidence level 3 or 4; or Extrapolated evidence from studies rated as 2+

RCT = randomised controlled trial.

study quality may raise or lower that grading.[6] RCTs with some limitations will lead to a 'moderate' categorisation while RCTs with a number of limitations will lead to a 'low' categorisation. Conversely, there may be instances where observational studies are upgraded to 'moderate' – or in very rare cases – 'high quality' categories. Diagnostic studies can be graded as high quality evidence, but to do so they generally have to be RCTs with very few limitations in study design.

The four main determinants for the strength of recommendation are:

- the balance between the desirable and undesirable consequences when compared to the alternative intervention or management strategy
- the quality of evidence and size of effect
- the value placed by stakeholders on the benefits, risks and inconvenience of the management strategy and its alternative
- high opportunity cost.[7]

Quality is not the only determinant for the strength of recommendations, therefore, the GRADE system (like SIGN) separates the grading of the quality of evidence from the grading of the strength of recommendation.[8] In a move towards simplicity and clarity only two levels of recommendation are used: strong and weak. Strong recommendations mean that they should be adopted by most of the three key stakeholders (patients, physicians and policy makers). A weak recommendation, however, means that while many patients will benefit from the recommendation, clinicians and policy makers should carefully consider circumstances and contexts before abiding by it.[7]

Graphic Appraisal Tool for Epidemiology

The GATE framework is largely a pictorial tool initially aimed at students and people who are not experts in epidemiology.[9] As a result, it is both simple and clear. GATE pictorially depicts the generic design for all epidemiological studies as illustrated in Fig 1.[9] The framework consists of a triangle, circle, square and arrows, which incorporate the PECOT (or PICOT) frame (Participants, Exposure/Intervention, Comparison, Time).

Table 3. Grading of Recommendations Assessment, Development and Evaluation (GRADE) grades for evidence.

High	Further research is very unlikely to change [the] confidence in the estimate of effect
Moderate	Further research is likely to have an important impact on [the] confidence in the estimate of effect and may change the estimate
Low	Further research is very likely to have an important impact on [the] confidence in the estimate of effect and is likely to change the estimate
Very low	Any estimate of effect is very uncertain

Filling in the GATE framework helps appraisers to understand what question is addressed by the study and how the investigators addressed it. This is important because on occasion the title is either obscure or asks a different question to the one answered in the study. Once the GATE frame is filled in, the study is ready to be critically appraised.

The acronym RAMMbo (Represent, Allocation/Adjustment, Maintain, Measured, Blind or objective) is used to guide assessors to ask the key questions about potential bias in a study. The GATE system and RAMMbo facilitates an assessment of the overall impact of a study's limitations. This is done by assessing the likely direction and the degree of impact each limitation has on the study. Once the overall impact of the limitations has been assessed, a judgement can be made about their impact on the study and their impact on the estimate of effect of the intervention. Utilisation of the GATE framework allows for the calculation of occurrence, incidence and size of effect.

To assist with the development of recommendations, a large 'X' is depicted under the GATE frame. The 'X' is used to identify the four quadrants of issues that need to be integrated to develop a meaningful evidence-based recommendation, including the evidence, patient values, clinical considerations and policy issues. Once the evidence is highlighted, experts are better able to consider the other factors already established by the framework to make a final recommendation. Although the GATE tool is an excellent one for teaching critical appraisal of papers, it does not assign a grade to papers or recommendations and therefore its use in guideline development is limited.

National Service Framework for Long Term Conditions grading system

The NSF-LTC typology was created to deal with the challenges of the research base of long-term conditions (LTCs).[10] Typically the

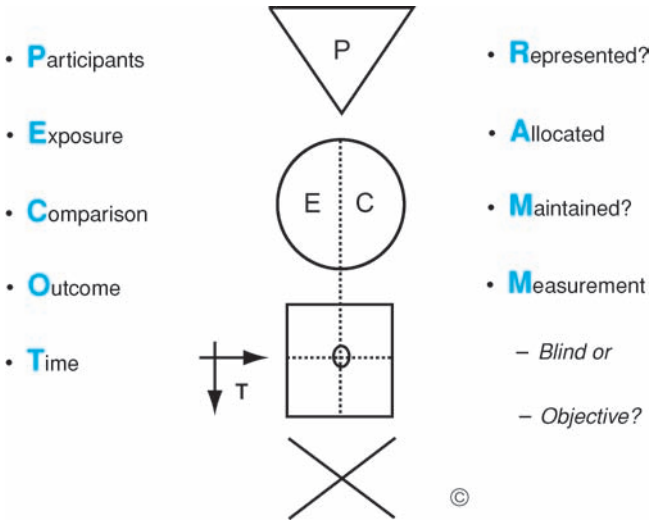

Fig 1. The Graphic Appraisal Tool for Epidemiology (GATE) framework. (Reproduced from *Annals of Internal Medicine* with permission).[9]

research base for LTCs tends to be more varied than traditional intervention studies, and can include longitudinal, case-report and qualitative studies, as well as expert opinion. These types of studies can also be found in the literature base of other specialties, such as occupational and sexual health. Current systems do not adequately address these types of studies and are not geared towards such conditions. The characteristics of life-long conditions pose a number of obstacles that traditional RCT research designs find difficult to cope with, such as the amorphous nature of the condition and the complexity of interventions.

These characteristics in part led to the identification of three criteria required for the new NSF-LTC typology. Firstly, the viewpoint and experience of professionals, service users, their families and carers must be taken into account as valid evidence. Secondly, emphasis should be placed on the quality of the study design and its generalisability, with the acknowledgement that qualitative, quantitative and mixed studies as well as expert opinion could be equally valid depending on the context and quality of the design. Thirdly, the typology should have a framework that can be applied to all types of research design and be practical, simple and quick.[10]

The type of evidence reviewed is differentiated by an 'E' (signifying 'expert' evidence – be it user, carer or professional) or 'R' (signifying research-based evidence) grade. Expert evidence is undertaken through consultation or consensus processes while research-based evidence is assessed in three categories (design, quality and applicability) and then awarded a grade. The design category is split into three, with each subsection containing a further subsection (Table 4).[10]

Quality is assessed through five questions, each being scored by a 0, 1 or 2. 'No' is denoted by a '0', 'In part' is denoted by a '1', and 'Yes' is denoted by a '2'. The five questions are as follows:

1 'Are the research question/aims and design clearly stated?'
2 'Is the research design appropriate for the aims and objectives of the research?'
3 'Are the methods clearly described?'
4 'Is the data adequate to support the authors' interpretations/conclusions?'
5 'Are the results generalisable?'

A poor quality study will score three or less. A medium quality study will score between four and six, while a high quality study will score seven or above. Applicability is then rated based on 'direct' and 'indirect' subcategories and the overall rating for a research-based evidence is an amalgamation of the scores for all the categories. For example, 'P2 high direct' would signify that the research-based evidence is a qualitative study, which is of high quality and directly applicable to the context. Each recommendation is given a grading (A, B or C) based on the overall quality of evidence (Table 5).[10]

Table 4. National Service Framework categories used to classify design.

Primary research-based evidence	
P1	Primary research using quantitative approaches
P2	Primary research using qualitative approaches
P3	Primary research using mixed methods (qualitative and quantitative)
Secondary research-based evidence	
S1	Meta-analysis of existing data analysis
S2	Secondary analysis of existing data
Review-based evidence	
R1	Systematic reviews of existing research
R2	Descriptive of summary reviews of existing research

Summary

The final categorisation of the appropriate critical appraisal and grading system(s) as shown in Table 6 identifies the strengths and weaknesses in relation to the field of research. This is to acknowledge that the optimal system depends on the nature of the research question posed.

Conclusion

The research base of specialist societies tends to consist of a wide range of research fields and study types and to be disadvantaged by traditional grading systems. The gold standard status of RCTs within these systems means that graded recommendations in evidence-based guidelines, with a research base mostly consisting of non-RCTs, are often low. Furthermore,

Table 5. National Service Framework grades for evidence.

Grade	Criteria
Research grade A	More than one study of high quality score (⩾7/10); and
	At least one of these has direct applicability
Research grade B	One high quality study; or
	More than one medium quality study (4–6/10); and
	At least one of these has direct applicability
	Or
	More than one study of high quality score (⩾7/10) of indirect applicability
Research grade C	One medium quality study (4–6/10)
	Or
	Lower quality (2–03/10) studies; or
	Indirect studies only

rigid grading systems may be misinterpreted in that users may assume that an absence of evidence means that there *is* evidence against a recommendation, when in reality it means that there is no evidence available for or against a clinical action. Another unintended consequence of grading systems is that those which have the highest grading of evidence of effectiveness may be given clinical priority over clinically more important recommendations which have been given a lower grading simply because they are backed by weaker evidence.

Appraisal systems have to balance simplicity with clarity while providing scope for flexibility and explicit judgements. Such a balance is difficult to maintain. 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 need 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 which will most effectively

Table 6. Summary of the strengths and weaknesses of the grading systems reviewed in relation to the type of study being appraised.

Field of research	Preferred study design	Suggested appraisal system	Strengths	Weaknesses
Therapy	RCTs	SIGN or GRADE	Both are established systems; appraisal focus is on RCTs	Training is required for both
				GRADE: classifies study types by hierarchy
Diagnosis	Cross-sectional survey	GRADE or NSF	GRADE: allows the assessment of a number of variables	GRADE: classifies study types by hierarchy
			NSF: easy to use, flexible	NSF: fewer variables assessed
Screening	Cross-sectional survey or RCT or Cohort studies	GRADE or NSF	GRADE: robust appraisal system; strong on RCTs	GRADE: classifies study types by hierarchy
			NSF: easy to use; flexible	NSF: fewer variables assessed
				Does not explicitly take into account confounding and size of effect
Prognosis	Prospective cohort	NSF	Easy to use; flexible	Does not explicitly take into account confounding and size of effect
Causation	Cohort/case-control	GRADE	More robust at appraising observational studies than SIGN; emphasises explicit judgements to increase transparency	Requires training; weak on case reports
Psychometric studies	Cross-sectional survey	NSF	Easy to use; little path dependency[a]; acknowledges expert opinion	Places expert opinion on equal status to other studies
Qualitative studies	Qualitative studies	NSF	Easy to use; little path dependency; acknowledges qualitative studies more than other studies; acknowledges expert opinion	May lead to implicit judgements
				Places expert opinion on equal status to other studies

[a]Path dependency occurs when a system, framework or set of questions leads the user towards a preconceived outcome in terms of the maximum grade that can be awarded or the likely range of grades that will be awarded due to a study design; GRADE = Grading of Recommendations Assessment, Development and Evaluation; NSF = National Service Framework; RCTs = randomised controlled trials; SIGN = Scottish Intercollegiate Guidelines Network.

address the predominant type of study design being appraised. Further work is being done to assess the ease of use and inter-assessor reliability of the grading systems reviewed in this paper.

Acknowledgements

We are grateful to Françoise Cluzeau, technical adviser, NICE; Chris Carmona, research analyst, NICE; Nichole Taske, analyst, NICE; Robin Harbour, information director, SIGN; Rod Jackson, head of epidemiology and biostatstics, School of Population Health, University of Auckland; and Kristina Pedersen for their advice on the review. We thank Barbara Smiley for her administrative assistance.

Competing interests

IM was former director of clinical standards for NHS Plus and commissioned evidence-based guidelines through the Clinical Standards Department, Royal College of Physicians. The project was funded by NHS Plus.

References

1 Royal College of Physicians' Clinical Effectiveness Forum. www.rcplondon.ac.uk/clinical-standards/organisation/partnership/Pages/joint-specialty-clinical-effectiveness-forum.aspx

2 Baker A, Young K, Potter J, Madan I. *A review of grading systems and critical appraisal tools for use by specialist medical societies developing evidence-based guidelines.* NHS Plus, 2009.

3 Scottish Intercollegiate Guidelines Network. *SIGN 50: A guideline developer's handbook.* Edinburgh: SIGN, 2008.

4 Atkins D, Best D, Briss PA *et al.* Grading quality of evidence and strengths of recommendations. *BMJ* 2004;328:1490–7.

5 Guyatt G, Vist G, Ytter-Falck Y *et al.* An emerging consensus on grading recommendations? *ACP J Club* 2006;144:A8–9.

6 Guyatt G, Oxman A, Vist G *et al.* What is 'quality of evidence' and why is it important to clinicians? *BMJ* 2008;336:995–8.

7 Guyatt G, Oxman A, Kunz R *et al.* Going from evidence to recommendations. *BMJ* 2008;336:1049–51.

8 Guyatt G, Oxman A, Vist G *et al.* GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924–6.

9 Jackson R, Ameratunga S, Broad J *et al.* The GATE frame: critical appraisal with pictures. *ACP J Club* 2006;144:A8–11.

10 Turner-Stokes L, Harding R, Sergeant J, Lupton C, McPherson K. Generating the evidence base for the National Service Framework for Long Term Conditions: a new research typology. *Clin Med* 2006;6:91–7.

Address for correspondence: Dr I Madan, Occupational Health Department, St Thomas' Hospital, Lambeth Palace Road, London SE1 7EH.
Email: ira.madan@kcl.ac.uk