

H S A C



HEALTH SERVICES
ASSESSMENT COLLABORATION

Briefing Report

June 2009

What assessment tools are used both in New Zealand and in other countries for grading of evidence?

Wasan Ali

This report should be referenced as follows:

Ali, W. What assessment tools are used both in New Zealand and in other countries for grading of evidence? *HSAC Report 2009*; 2(6)

Health Services Assessment Collaboration (HSAC), University of Canterbury

ISBN 978-0-9864544-5-5 (online)

ISBN 978-0-9864544-6-2 (print)

ISSN 1178-5748 (online)

ISSN 1178-573X (print)

Review Team

This briefing report was commissioned by the New Zealand Ministry of Health. It was undertaken by the Health Services Assessment Collaboration (HSAC). HSAC is a collaboration of the Health Sciences Centre of the University of Canterbury, New Zealand and Health Technology Analysts, Sydney, Australia. It was prepared by Dr Wasan Ali (Researcher) who conducted the literature search, extracted data and prepared the report.

Acknowledgements

Dr Arindam Basu (Senior Researcher) peer reviewed the drafts and Dr Ray Kirk (HSAC Director) reviewed the final draft. Franziska Gallrach (Research Assistant) assisted with retrieval of documents. Cecilia Tolan (Administrator) provided document formatting.

What is a Briefing Report?

This briefing report, prepared by HSAC Reviewers, is an overview of a specific topic area identified from a systematic search strategy of electronic databases and website resources. The material includes lists of abstracts, key full text papers (where readily available from local resources) and website resources.

The report is aimed at giving the client an informed “guided tour” of what the search strategy identifies around the topic area and outlines the contents of the report, highlights information of particular interest and relevance and summarises key articles. Briefing reports do not involve systematic processes for the critical appraisal of identified research, but may present data from full text articles in tables (without appraisal). Another significant limitation is that full text articles of key interest are not retrieved unless freely obtainable from local resources. As a consequence of this, comments and summaries in the brief report may be based on abstracts rather than full text papers.

Copyright Statement & Disclaimer

This report is copyright. Apart from any use as permitted under the Copyright Act 1994, no part may be reproduced by any process without written permission from HSAC. Requests and inquiries concerning reproduction and rights should be directed to the Director, Health Services Assessment Collaboration, Health Sciences Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand.

HSAC takes great care to ensure the accuracy of the information in this report, but neither HSAC, the University of Canterbury, Health Technology Analysts Pty Ltd nor the Ministry of Health make any representations or warranties with respect to the accuracy or quality of the information, or accept responsibility for the accuracy, correctness, completeness or use of this report.

The reader should always consult the original database from which each abstract is derived, along with the original articles, before making decisions based on a document or abstract. All responsibility for action based on any information in this report rests with the reader.

This report is not intended to be used as personal health advice. People seeking individual medical advice should contact their physician or health professional.

The views expressed in this report are those of HSAC and do not necessarily represent those of the University of Canterbury New Zealand, Health Technology Analysts Pty Ltd, Australia or the Ministry of Health.

Contact Details

Health Services Assessment Collaboration (HSAC)
Health Sciences Centre
University of Canterbury
Private Bag 4800
Christchurch 8140
New Zealand
Tel: +64 3 345 8147 Fax: +64 3 345 8191

Email: hsac@canterbury.ac.nz
Web Site: www.healthsac.net

Table of Contents

Review Team	i
Acknowledgements	i
What is a Briefing Report?	i
Copyright Statement & Disclaimer	i
Contact Details	ii
Table of Contents	iii
List of Tables	v
List of Abbreviations and Acronyms	vi
List of Abbreviations and Acronyms	vi
Introduction	1
Background	1
Characteristics of the report	1
Methods	3
Literature search	3
Results	5
What Grading System does New Zealand Use?	7
Grading is a two-tier process	7
Grades of recommendations	8
What are the Other Organisations Using?	11
What are the Most Commonly Used Grading Systems Reported by the Literature?	13
Conclusions	31
List of Systems Most Reported in the Literature	33
The Agency for Health Care Quality and Research (AHRQ)	33
What the grades mean and suggestions for practice	33
Grade definitions	34
Centre of Evidence-Based Medicine (CEBM)	36
Cochrane Group	39
Factors that decrease the quality level of a body of evidence	40
CTFPHC Methods (CTFPHC)	41
Some challenges for evidence-based prevention	42
The Grading of Recommendations Assessment, Development and Evaluation Working Group (GRADE)	43
Medical Services Advisory Committee (MSAC) the National Health and Medical Research Council (NHMRC)	47
National Institute for Health and Clinical Excellence (NICE) from the National Health Services (NHS)	48
The New Zealand Guidelines Group (NZGG)	49
Grades of recommendation	49
Scottish Intercollegiate Guideline Network (SIGN)	51

Strength of Recommendation Taxonomy (SORT)	52
The World Health Organization (WHO)	54
References	55

List of Tables

Table 1:	Different questions need different study designs (taken from NZGG handbook)	7
Table 2:	Quality criteria	8
Table 3:	Grading of recommendations	8
Table 4:	Systems for grading the quality of individual studies	9
Table 5:	Strengths and weaknesses of grading systems in differing research fields	14
Table 6:	Strengths and weaknesses of various strategies for developing recommendations*	17
Table 7:	Comparison of methodologies between NICE and SIGN*	19
Table 8:	Comparison of different systems for grading levels of evidence	20
Table 9:	Comparison of considered judgment to formulate recommendations from various organisations*	27
Table 10:	Pros and cons of using the same system*	29
Table 11:	Grade definitions and suggestions for practice	33
Table 12:	Levels of certainty regarding net benefit	34
Table 13:	Levels of Evidence	36
Table 14:	Levels of quality of a body of evidence in the GRADE approach	39
Table 15:	Factors that may decrease the quality level of a body of evidence	39
Table 16:	Factors that may increase the quality level of a body of evidence	39
Table 17:	Recommendation grades for specific clinical preventive actions	41
Table 18:	Levels of evidence - research design rating	41
Table 19:	Levels of evidence - quality (internal validity) rating (see Harris et al., 2001)	42
Table 20:	GRADE comparison to other grading systems	43
Table 21:	GRADE grid for recording panellists' views in development of guidelines	44
Table 22:	Organisations using GRADE system	45
Table 23:	Types of clinical and public health questions, ideal study types and major appraisal issues (taken from NHMRC 200a, page 10)	47
Table 24:	Classifying the relevance of the evidence	48
Table 25:	The NZGG considered judgment form to grade evidence	50
Table 26:	SIGN grading system	52
Table 27:	Definitions of strength of recommendations	53
Table 28:	Definitions of levels of evidence for each study type	53
Table 29:	Definitions of consistency across studies	54
Table 30:	GRADE quality assessment criteria	54

List of Abbreviations and Acronyms

ACCP	American College of Chest Physicians
AGREE	Appraisal of guidelines research and evaluation
AHCPR	Agency for Health Care Policy and Research
AHRQ	Agency for Health Research and Quality
CADTH	Canadian Agency for Drugs and Technologies in Health
CAM	Complementary and Alternative medicine
CBO	The Dutch Institute for Healthcare Improvement
CCS	Canadian Cardiovascular Society 2000 Consensus
CEBM	Centre for Evidence Based Medicine
CRAG	Clinical Resource and Audit Group
CSP	Chartered Society of Physiotherapy
CTFPHC	Canadian Task Force on Preventive Health Care
EBMC	Evidence Based Medicine Centre
EGSs	Evidence Grading Systems
EPIQ	Effective Practice, Informatics and Quality Improvement
FPIN	Family Practice Inquiries Network
GATE	Graphic Appraisal Tool for Epidemiology
GHC	Group Health Cooperative
GPP	Good Practice Points
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HSAC	Health Services Assessment Collaboration
G-I-N	Guidelines International Network
MSAC	Medical Services Advisory Committee
NHMRC	National Health and Medical Research Council
NHSQIS	National Health Service Quality Improvement Scotland
NICE	National Institute for Health and Clinical Excellence
NSF	National Service Framework (UK)
NSTR	National Service and Technology Review Subcommittee
NZGG	New Zealand Guidelines Group
RCT	Randomised Controlled Trial
SIGN	Scottish Intercollegiate Guidelines Network
SR	Systematic Review
USPSTF	US Preventive Services Task Force

Introduction

Background

The New Zealand Ministry of Health requested HSAC to undertake a *short briefing report* on what assessment tools are used for grading the evidence. The National Service and Technology Review Subcommittee (NSTR) of The Ministry of Health was interested in identifying what rating or grading systems are being used by other countries as well as New Zealand. The specific aim of the request was to provide some helpful information for both the Ministry and other committees (e.g., NSTR) for determining weighting or scoring that should be placed on results of an analysis when making a funding decision. NSTR indicated that they are aware of a number of different methods for assigning levels of evidence from various grading systems and that they are looking at identifying the assessment processes within these systems.

In order to identify the tools that are used worldwide and are considered to be appropriate to evaluate different types of guidelines, HSAC conducted a literature review and sought opinions of the experts and leaders in the areas of evidence assimilation. The following report briefly provides a summary of the findings. For detailed analysis, the reader should review the original texts.

From the perspective of this requirement, this report undertook the following:

- Identified a list of assessment tools used internationally in general and in New Zealand in particular for grading of evidence.
- Annotated commentaries about the assessment processes used. These included the positive and negative points, interesting information, and overall applicability to different scenarios.

Characteristics of the report

This is a briefing report, not a literature review or critical appraisal. This report provides an annotated package of information obtained from available resources on these grading systems. The purpose of this report is to answer the questions from the Ministry of Health and NSTR as to what assessment tools are used in New Zealand and elsewhere for grading health care evidence and what the advantages, disadvantages, and applicability of these systems to different healthcare scenarios are.

Methods

Literature search

HSAC conducted a search of the online literature and other relevant databases to identify and list grading systems used by various guidelines groups. The search initially identified abstracts from handbooks or publications from the Agency for Health Research and Quality (AHRQ), the Oxford Centre for Evidence Based Medicine (CEBM), the Cochrane Collaboration (COCHRANE), the National Institute for Health and Clinical Excellence (NICE), the Medical Services Advisory Committee (MSAC), the New Zealand Guidelines Group (NZGG), and the Scottish Intercollegiate Guidelines Network (SIGN). These groups are based in the USA (AHRQ), the UK (CEBM, COCHRANE, and NICE), Australia (MSAC), New Zealand (NZGG), and Scotland (SIGN). HSAC also identified and listed tools currently in use by member agencies of the Guidelines International Network (G-I-N). This process enabled identification of additional groups through their websites.

HSAC contacted the evidence-based-health list – an electronic discussion group to which many evidence-based medicine and health care experts, academics and researchers subscribe – vetting information and expert advice. HSAC sent an email inquiring after the availability of recent comparative analysis of quality assessment tools in grading evidence. This group involves a wide community of professionals with special interest and expertise in evidence-based-health care, including well-known key specialists and professors in the specialty from around the world.

HSAC contacted authors to obtain further information from work published in conferences or from unpublished work in the literature. The contacts were from the Canadian Agency for Drugs and Technologies in Health (CADTH) and from Evidence Based Medicine Centre in Mexico (details available upon request).

Results

In general, the retrieved guidelines aim to close the gap between research and practice and provide rigorously developed, valid, and applicable recommendations for achieving the best possible outcomes. The process of formulating evidence-based guidelines involves 1) systematic exploration of the body of evidence, 2) critical appraisal of its quality, 3) synthesis of the research findings, and 4) translations into recommendations for best practice. In practice guidelines, quality of evidence indicates the degree of confidence that the evidence is adequate to support recommendations. The quality of evidence can be judged by the following:

Study design: Defines the level of evidence. For example, questions on the efficacy of treatment are best answered by randomised controlled trials (RCTs) whereas questions about diagnostic accuracy are best addressed by properly designed prospective cohort studies.

Internal validity: Refers to the accuracy of the measurement of the effect estimates reported by a study. Specifically, this relates to those aspects of a study design which addresses how well the findings of a study from the sample can be applicable to the study population as a whole. This is dependent on the extent to which study related biases are eliminated in the study design – either at the stage of planning the study, during the process of execution of the study, or during the analysis and interpretation of study results. As an example, in diagnostic accuracy studies, various forms of verification biases, spectrum bias, or review bias can lead to overestimates of diagnostic performance – all of which could invalidate the conclusions and implications of the study results.

Consistency: Refers to the similarity of estimates of effect in comparable studies conducted in different populations addressing the same exposure-outcome relationships.

Directness: The extent to which the study's patients, interventions, and outcomes are similar to those in practice. Diagnostic accuracy is a surrogate for important outcomes for patients and thus is considered to provide indirect evidence.

Precision: Refers to the reliability of an estimate of effect and is influenced by the sample size of the study, the techniques for measuring the analyte, and the variation in analyte values in the populations.

Other factors: such as reporting bias, can lower the quality of the evidence, whereas a strong association or the presence of a dose-response gradient can increase the quality of the evidence.

Beyond the scientific judgements of the quality and strength of evidence, guideline developers need to make value judgments before formulating final recommendations. In this phase most practice guidelines assess the strength of recommendations which indicate the extent to which one can be confident that adherence to the recommendation will do more good than harm (Horvath, 2009). Value judgments about the strength of recommendation imply that, in addition to evidence, guideline developers have given due consideration to the following practical aspects: balance between benefits and harms; transferability of the evidence to the given population, condition, or outcomes; preferences of the patient; impact on healthcare organisation; and costs. Most grading systems rate separately the quality of a body of evidence and the strength of recommendations, the latter are developed in a more rigorous and transparent fashion which raises confidence in the process and lead, to better medical decisions and improved patient outcomes.

What Grading System does New Zealand Use?

Guideline recommendations need to be based on the best available evidence. There should be explicit links between the strength of the available evidence and the grade of the recommendation (NZGG handbook). This handbook is a compilation of a number of different guideline development processes. The handbooks and manuals of Group Health Cooperative, Puget Sound, Washington, the National Health and Medical Research Council (NHMRC) of Australia and Scottish Intercollegiate Guideline Network (SIGN) have all been invaluable resources. Chapters 2, 3 and 8 have drawn on a previous handbook from GHC. Chapters 7 and 10 have adapted work from NHMRC and research papers from ICSI (Institute for Clinical Systems Improvement), Minnesota, and USPSFC (US Preventive Services Task Force). Chapter 9 is adapted from SIGN and GHC. Chapter 11 was developed from the AGREE tool.

The New Zealand Guidelines Group uses steps in assessing the evidence and developing graded recommendations to assist practice.

Grading is a Two-tier process

Grading is firstly based on an (objective) assessment of the design and quality of each individual study (study quality), secondly based on a judgment (which may be more subjective) on the consistency, relevance and applicability of the whole body of evidence to the questions the guideline seeks to answer (graded recommendations). The process incorporates several approaches in the evaluation system (Harbour & Miller 2001; Greer et al., 2000; Harris et al., 2001). This evaluation process involves assessing the evidence relevant to guideline questions (which is essentially critical appraisal and determination of study design for each study such as randomised controlled trial, cohort, systematic review, etc), assessing the quality scores for each study (+, Ø, or -), and developing graded recommendations from the body of evidence based on the volume of evidence, consistency, clinical relevance and applicability. In assessing the evidence, it is recognized that different study designs require slightly different types of assessment when quality is being evaluated, so separate checklists are provided for systematic reviews and meta-analyses, randomized controlled trials (RCTs), cohort studies, case-control studies and diagnostic studies. The NZGG states that it is unlikely that other types of study design will need to be assessed as evidence in the development of a guideline. The checklists contain three main sections: study validity (steps made to minimize bias), study results (size of effect and precision), and study relevance (containing applicability/generalisability).

Table 1: Different questions need different study designs (taken from NZGG handbook)

Clinical questions	Most appropriate study design	Outcome measures
Diagnosis	Cross-sectional Cohort	Sensitivity, specificity, likelihood ratios, number needed to test Patient expected event rate
Harm	Randomized control trials or Cohort or Case-control	Number needed to harm
Therapy	Systematic reviews or Randomized control trials	Absolute Risk Reduction, Number needed to treat

Once the study design has been determined and the study critically appraised using the relevant check listed, a quality score can be determined for each section. There are three

quality categories that can be assigned based on the extent to which the study design has met the criteria (Table 2).

Table 2: Quality criteria

Quality criteria	Criteria
Plus (+)	Strong study where all or most of the validity criteria are met (i.e. in the shaded boxes of the checklist).
Minus (-)	Weak study where very few of the validity criteria are met and there is a high risk of bias.
Neutral (Ø)	Study where not all of the criteria are met but the results of the study are not likely to be affected.

The New Zealand Guidelines Group, uses the Appraisal of Guidelines for Research and Evaluation (AGREE) instrument, which is a generic tool designed primarily to help guideline developers and users assess the methodological quality of clinical practice guidelines. The AGREE Instrument assesses both the quality of the reporting and the quality of some aspects of the recommendations. It provides an assessment of the predicted validity of a guideline; that is, the likelihood that it will achieve its intended outcome. The New Zealand Guidelines Group uses the following grades of recommendations:

Grades of recommendations

Grades indicate the strength of the supporting evidence rather than the importance of the evidence. Table 3 is taken from the NZGG handbook for the preparation of explicit evidence-based clinical practice guidelines (page 54).

Table 3: Grading of recommendations

Grade of recommendations	Supporting evidence
A	The recommendation is supported by good evidence (based on a number of studies that are valid, consistent, applicable and clinically relevant).
B	The recommendation is supported by fair evidence (based on studies that are valid, but there are some concerns about the volume, consistency, applicability and clinical relevance of the evidence that may cause some uncertainty but are not likely to be overturned by other evidence).
C	The recommendation is supported by international expert opinion.
Good Practice Points (GPP)	Where no evidence is available, best practice recommendations are made based on the experience of the Guideline Development Team, or feedback from consultation within New Zealand.

Below is a comparison of The New Zealand Guidelines Group (including the system used for complementary and alternative medicine (CAM) with other grading systems.

Table 4: Systems for grading the quality of individual studies

NZGG		SIGN	GRADE	USPSTF	Oxford CEBM	NHMRC	CCS 2000 Consensus
CAM	GATE						
Level 1	Good/+	++	High	Good	Level 1 abc	Level I	Level 1
Level 2	Fair/-	+	Moderate	Fair	Level 2 abc	Level II	Level II
Levels 3 and 4	Poor/-	-	Low (very low)	Poor	Level 3ab, and 4	Level III (1, 2, 3) and IV	Level III, IV and V
NZGG New Zealand Guidelines Group, CAM complementary and alternative medicine, GATE Graphic Appraisal Tool for Epidemiology, SIGN Scottish Intercollegiate Guidelines Network, USPSTF US Preventable Services Task Force, Oxford CEBM Centre for Evidence-based Medicine, NHMRC National Health and Medical Research Council 2000, CCS Canadian Cardiovascular Society 2000 Consensus							

What are the Other Organisations Using?

Guideline recommendations are graded depending on the strength of evidence on which they are based. However, the plethora of available grading systems makes it difficult for guideline developers to choose which system to adopt. As a result, guideline developers are often inconsistent in their methods of rating the quality of evidence (sometimes called levels of evidence) and grading the strength of recommendations (Baker et al., 2009; Schünemann, 2006).

The Canadian Task Force on the Periodic Health Examination published in 1979 one of the first efforts to explicitly characterise the level of evidence underlying healthcare recommendations and the strength of recommendations (Canadian Task Force, 1979). The original approach used by the Canadian Task Force was based on study design alone, with randomised controlled trials being classified as good (level I) evidence, cohort and case control studies as fair (level II) evidence, and expert opinion being classified as poor (level III) evidence. The strength of recommendations was based on the level of evidence with direct correspondence between the two. Since then a number of alternative approaches have been proposed and used to classify clinical practice guidelines and different organisations use various grading systems (Atkins et al., 2004).

For example, the US Preventive Services Task Force (USPSTF) uses a grading system that assigns one of three grades of evidence: good, fair, or poor (National Guideline Clearing House).¹ The Task Force uses its assessment of the evidence and magnitude of net benefit to make a recommendation, coded as a letter: from A (strongly recommended) to D (recommend against). The UK National Institute for Health and Clinical Excellence (NICE) has recently begun to use elements of the GRADE approach for questions about interventions in its clinical guidelines (NICE, 2009).² The Scottish Intercollegiate Guideline Network (SIGN) has developed its own grading system for application to SIGN guidelines (Harbour, 2001³; Sign, 2009)⁴. The Australian Medical and Health Research Council is currently using a grading system that includes grading recommendations according to strength of recommendations and quality of evidence (NHMRC, 2008-2009).⁵ This part of the document describes how to grade the 'body of evidence' for each guideline recommendation. The body of evidence considers the evidence dimensions of all the studies relevant to that recommendation. The US Task Force on Community Preventive Services uses a variety of qualitative and quantitative factors to assess the strength of the evidence which is then translated into a Task Force recommendation. This is a system in which the quality of the evidence of effectiveness links directly the strength of the recommendation (Task Force on Community Preventive Services).⁶ The Oxford Centre for Evidence Based Medicine sets out one approach to systematising this process for different question types (CEBM, 2009).⁷

The problem that occupational health and other specialties face is that the majority of grading hierarchies were created in a period when randomised controlled trials (RCTs) were deemed

¹ <http://www.guidelines.gov/>

² http://www.nice.org.uk/media/5F5/22/The_guidelines_manual_2009_-_Chapter_6_Reviewing_the_evidence.pdf

³ <http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=11498496>

⁴ <http://www.sign.ac.uk/guidelines/fulltext/50/annexb.html>

⁵

http://www.nhmrc.gov.au/guidelines/_files/Stage%20%20Consultation%20Levels%20and%20Grades.pdf

⁶ <http://www.thecommunityguide.org/about/strengthofevidence%20assessment.pdf>

⁷ <http://www.cebm.net/index.aspx?o=1025>

the gold standard of medical studies (Baker et al., 2009). Questions posed by specialist societies are often related to answering a clinical problem and may not be answerable by an RCT, for example questions on prognosis or patients' views. This context requires a balance in the grading system between simplicity and clarity, without jeopardising transparency and legitimacy. Therefore, for the purpose of overcoming this difficulty within the specialist societies, SIGN undertook a project to develop a framework of optimum grading systems for grading the evidence in papers for evidence-based guidelines developed by specialist societies (Baker et al., 2009). The review looked at the strengths and weaknesses of the current major grading systems in the context of their use by specialist societies, and identified the optimum grading system for the type of guideline being developed or question being addressed by the specialist society.

Therefore, there exist multiple systems and multiple organisations, but it is recognised that different things need to be graded differently, so which grading system is better?

The following section will first present the systems more commonly reported in the literature, this would be from the latest (at the time of the preparation of this report) information available from the systematic reviews identified in this report and from websites and handbooks of the organisations preparing clinical practice guidelines and guidance. Then these systems will be presented as they appear in the handbooks and will be listed alphabetically.

What are the Most Commonly Used Grading Systems Reported by the Literature?

In an attempt to facilitate movement from evidence to recommendations the Cochrane review published an abstract (Poster presentation at the 16th Cochrane Colloquium-Germany, 2008) of a recent systematic review of evidence grading systems (EGSs) to identify the most appropriate one for grading levels of evidence (Shukla et al., 2008). (Author contacted for further details). This review was built upon the Evidence report/technology assessment from the Agency for Healthcare Research and Quality (AHRQ) that evaluated 34 evidence grading systems developed between 1979 and 2001 (West, 2002). A ten-step approach was used to conduct the same systematic review in 2005 and to continue updating it until September 2007 (Shukla et al., 2008). From 3000 citations, the authors identified 51 existing evidence grading systems, and evaluated 23 systems by the same methodological domains of the AHRQ report. Based on this work, the authors concluded that the GRADE and SIGN 50 systems are the most appropriate evidence grading systems for use in grading evidence for the purpose of making recommendations.

The paper from the Scottish Intercollegiate Guidelines Network (SIGN) (Baker et al., 2009) reviewed systems that were driven by initial search and discussions with members of the Royal College of Physicians Clinical Effectiveness Forum and experts at NICE. They chose the 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) grading system.

The reason for choosing these systems was: SIGN (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), NSF (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), and the GATE system (due to its simplicity and clarity, and its ability to be used to critically appraise different types of studies).

The report has reviewed the strengths and weaknesses of grading systems in differing research fields. These were presented in the **Table 5** taken from the report (Baker et al., 2009).

Table 5: Strengths and weaknesses of grading systems in differing research fields

Field of research	Preferred study design	Suggested appraisal system	Strengths	Weaknesses
Therapy (1)	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 (2)	Cross-sectional survey	GRADE or NSF	GRADE: Allows the assessment of a number of variables. NSF: Easy to use, flexible.	GRADE: Classifies study types by hierarchy. NSF: Fewer variables assessed.
Screening (3)	Cross-sectional or RCT or cohort studies	GRADE or NSF	GRADE: Robust appraisal system; strong on RCTs. NSF: Easy to use; flexible.	GRADE: Classifies study types by hierarchy. NSF: Fewer variables assessed. Does not explicitly take into account confounding and size of effect
Prognosis (4)	Prospective cohort	NSF	Easy to use; allows for flexibility.	Does not explicitly take into account confounding and size of effect.
Causation (5)	Cohort/case-control	GRADE	More robust at appraising observational studies than SIGN; emphasises explicit judgments to increase transparency.	Requires training; weak on case reports.
Psychometric studies (6)	Cross-sectional survey	NSF	Easy to use; little path dependency; acknowledges expert opinion.	Places expert opinion on equal status to other studies.
Qualitative studies (7)	Qualitative studies	NSF	Easy to use; little path dependency; acknowledges qualitative studies more than other studies. Acknowledges expert opinion.	May lead to implicit judgments. Places expert opinion on equal status to other studies.

Table 5: Strengths and weaknesses of grading systems in differing research field (*continued*)

<p>(1) Testing the efficacy of drug treatments, surgical procedures, alternative methods of service delivery or other interventions. Preferred study design is the randomised controlled trial (RCT).</p> <p>(2) Demonstrating whether a new diagnostic test is valid (can we trust it?) and reliable (would we get the same result, every time?) preferred study design is the cross-sectional survey.</p> <p>(3) Demonstrating the value of tests which can be applied to large populations and which pick up disease at a pre-symptomatic stage, preferred study design is cross-sectional survey. However, if the question is whether screening improves the outcome, this is best tested by an RCT.</p> <p>(4) Determining what is likely to happen to someone whose disease is picked up at an early stage. Preferred design is the longitudinal survey.</p> <p>(5) Determining whether a putative harmful agent is related to the development of illness. Preferred study design is cohort or case-control study; depending on how rare the disease, the case reports may also provide crucial information.</p> <p>(6) Measuring attitudes, beliefs or preferences, often about the nature of the illness or its treatment. Cross sectional studies are usually sufficient.</p> <p>(7) Measures attitudes about healthcare intervention or provision.</p>
--

The authors also reported on the Graphic Appraisal Tool for Epidemiology (GATE) designed by Professor Rod Jackson and colleagues in the EPIQ (Effective Practice, Informatics & Quality Improvement) group⁸ at the University of Auckland and it was later analysed and further discussed with two analysts from NICE Public Health section: Nicole Taske and Chris Carmona. The GATE approach is pictorial and depicts the generic design for all epidemiological studies. The framework consists of a triangle, circle, square, and arrows, which incorporate the PICOT (or PECOT) frame (Population; Exposure/Intervention; Comparison; Study Time).

GATE uses the following grading of recommendations approach:

There is a large 'X' depicted under the GATE frame which 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 using the GATE frame, experts are better able to consider the other factors already established by the framework to make a final recommendation. The authors mentioned that the NICE Public Health section has used an adaptation of the GATE system alongside other critical appraisal models to critically appraise their literature base.

The report by Baker et al. (2009) discussed the GATE approach, and noted that it takes a linear approach to assessment like other systems but differentiates itself by attempting to focus on the study as a whole rather than assessing multiple variables within a study and producing an overall grade based on the sum of the grades given to those variables. Whilst the GATE system includes checklists which examine individual components, they are less comprehensive than SIGN or GRADE. The authors also stated that with GATE, the simplicity of the process overarches the clarity, though the clarity is not overly affected. The final part of the GATE frame, represented by the 'X', highlights the 'non-scientific'

⁸ <http://www.fmhs.auckland.ac.nz/soph/depts/epi/epiq/default.aspx>

components of recommendations that often lead to implicit judgment due to the lack of evidence or the failure to identify ‘evidence’ questions that need to be answered to demonstrate how an opinion/judgment was derived. The interlinkage of these attributes means that the GATE system minimises path dependency and increases user flexibility. Like any framework, the extent to which these balances are played out results in its overall characteristic. As such, the GATE system is a tool which can be easily utilised on any epidemiological study and provides the flexibility needed for a universal critical appraisal tool, yet may not have the robustness needed to be the stand-alone tool for institutional utilisation.

The authors of the report from SIGN concluded that “The decision on which grading system should be used for specialist society guidelines depends on the research area to which the guideline questions pertain. Further work is being done to assess the ease of use and reliability of the grading systems reviewed in this report. The use of one grading system is only recommended if the research base for a guideline consists of one research field or predominantly one study type. If the research base is heterogenous, then more than one system maybe considered depending on the fields or types. However, the final decision should be made by the specialities and guideline groups based on the relevant context” (Baker et al., 2009).

Another grading system was reported in the literature and that is the Strength of Recommendation Taxonomy (SORT). SORT addresses the quality, quantity, and consistency of evidence and allows authors to rate individual studies or bodies of evidence. Ebell et al. (2004) stated that journals, including the *American Family Physician* and *Journal of Family Practice*, have adopted evidence-grading scales that are used in some of the articles published in those journals. Other organisations and publications have also developed evidence-grading scales. The diversity of these scales can be confusing for readers. More than 100 grading scales are in use by various medical publications. Therefore, the editors of the US family medicine and primary care journals (i.e., *American Family Physician*, *Family Medicine*, *Journal of Family Practice*, *Journal of the American Board of Family Practice*, and *MBJ-USA*) and the Family Practice Inquiries Network (FPIN) came together to develop a unified taxonomy for the strength of recommendations based on a body of evidence. The taxonomy is designed to emphasize the use of patient-oriented outcomes that measure changes in morbidity or mortality. An A-level recommendation is based on consistent and good quality patient-oriented evidence; a B-level recommendation is based on inconsistent or limited quality patient-oriented evidence; and a C-level recommendation is based on consensus, usual practice, opinion, disease-oriented evidence, or case series for studies of diagnosis, treatment, prevention, or screening. Levels of evidence from I to III for individual studies also are defined (Ebell et al., 2004).

Strengths and weaknesses of SORT were compared with SIGN and GRADE by Palda et al. (2007) (**Table 6**).

Table 6: Strengths and weaknesses of various strategies for developing recommendations*

Strategy	Strengths	Weaknesses
GRADE	<p>Working group is an international collaboration interested in developing a common grading system to address limitations and draw on strengths of existing systems.</p> <p>System sequentially assesses quality of evidence, balance between risks and benefits, and judgment about the strength of recommendations.</p> <p>Weak: recommendations reflect evidence that the benefits, risks and burdens are finely balanced, or there is appreciable uncertainty about the balance; furthermore, the recommendation is classified as “weak” if, across the range of patient values, fully informed patients are liable to make different choices.</p>	<p>Application is complicated.</p> <p>Often difficult for recommendation.</p> <p>Developers use formulaic approaches to global judgments about evidence.</p>
SIGN Method	<p>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.</p> <p>Levels of evidence (1++, 1+, 1-, 2++, 2+, 2-, 3 or 4) depend on type and quality of study design; grade of recommendation (A, B, C or D) reflects assigned level of evidence.</p> <p>“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.</p>	<p>System lacks transparency; no rationale provided to clarify which factors are weighted more heavily for any particular recommendation.</p> <p>Use of numbers and letters may not be intuitive.</p>

Table 6: Strengths and weaknesses of various strategies for developing recommendations* (continued)

Strategy	Strengths	Weaknesses
SORT Taxonomy	<p>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.</p> <p>Emphasizes patient-oriented outcomes (i.e. "outcomes that matter to patients and help them live longer or better lives, including reduced morbidity, mortality or symptoms, improved quality of life or lower cost" (Ebell et al., 2004)).</p> <p>Rates quality of individual studies as follows; 1= good-quality patient-oriented evidence, 2= limited-quality patient-oriented evidence, 3= other evidence.</p> <p>Grades strength of recommendations by letters: A recommendations based on consistent, good-quality patient-oriented evidence; B on inconsistent or limited-quality patient-oriented evidence; C on consensus, usual practice, opinion, disease-oriented evidence or case series for studies of diagnosis, treatment, prevention or screening .</p>	<p>Limited guidance for developers on how to classify studies within numeric categories (1, 2 or 3).</p> <p>Use of numbers and letters may not be intuitive.</p>
<p>*taken from Palda et al., 2007; Note: GRADE = Grading of Recommendations, Assessment, Development and Evaluation system (www.gradeworkinggroup.org), SIGN = Scottish Intercollegiate Guidelines Network method (www.sign.ac.uk), (SIGN 50), SORT = Strength of Recommendation Taxonomy (Ebell et al., 2004)</p>		

In another instance, the SIGN compared methodologies between NICE and SIGN by senior staff from the NICE guideline directorate and SIGN. The staff meet quarterly to discuss items of common interest and to identify areas where sharing of information is mutually beneficial. During 2004 they were asked by the Chairmen and Chief Executives of NICE and NHS Quality Improvement Scotland (NHS QIS) to produce a short paper to identify where variations occur in the two approaches to guideline development (**Table 7**).

Table 7: Comparison of methodologies between NICE and SIGN*

NICE	SIGN
Topic selection processes	
NICE formally referred topics by the Department of Health and the Welsh Assembly, but NICE is involved in various stages of topic identification and selection.	Healthcare professionals and members of the public suggest topics to SIGN. SIGN Council and its subgroups recommend the proposed programme to NHS QIS, who give final approval.
Evaluation of evidence	
Trained systematic reviewers at the National Coordinating Centres commissioned by NICE to evaluate the evidence in discussion with clinical experts.	Guideline development group (GDG) members are given training by SIGN in critical appraisal and assess the evidence themselves.
Economics	
NICE aim to ensure health economists are core members of the technical team in the GDG; include cost effectiveness analysis in all recommendations and undertake a cost impact analysis for each guideline.	SIGN do not undertake economic analysis, but include any relevant high quality published economic evaluation in the evidence base. SIGN guidelines include commentary on the resource implications of recommendations if these are significant.
Stakeholder consultation	
NICE has three stakeholder consultations, review by independent experts when appropriate and independent review by the Guideline Review Panel.	SIGN has one national meeting which anyone can attend, peer review by independent experts and lay reviewers, and a final independent review by the SIGN Editorial Group.

*National Institute for Clinical Excellence (NICE), Scottish Intercollegiate Guidelines Network (SIGN)
<http://www.sign.ac.uk/methodology/comparison.html>

In 2004, the GRADE (Grading of Recommendations Assessment, Development and Evaluation) working group published a critical appraisal of the six most prominent systems for grading levels of evidence and strength of recommendations (Atkin et al., 2004).⁹ The systems were those adopted by The American College of Chest Physicians (ACCP), the Australian National Health and Medical Research Council (NHMRC), the Oxford Centre of Evidence Based Medicine (OCEBM), the Scottish Intercollegiate Guidelines Network (SIGN), US Preventive Service Task Force (USPSTF), and the US Task Force on Community Preventive Services (USTFCPS).

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. See **Table 8** (data extracted from the paper).

⁹ <http://www.biomedcentral.com/content/pdf/1472-6963-4-38.pdf>

Table 8: Comparison of different systems for grading levels of evidence

Grading system	Strength of recommendations	Strengths	Weaknesses	Target audiences
ACCP ¹⁰ (American College for Chest Physicians)	<p>Quality of evidence contributes directly to grade of recommendations.</p> <p>If experts are very certain that benefits do, or do not, outweigh risks, they will make a strong recommendation (in GRADE formulation as Grade 1). If less certain of the magnitude of benefits and risks, and thus their relative impact, they must make a weaker, Grade 2, recommendation. The approach expresses the primacy of the risk/benefit judgment in determining the recommendation and its strength by placing it first. The grades generated are 1A, 1B, 1C+, 1C, 2A, 2B and 2C.</p>	<p>Has evolved over 15 years, methodologies and sophisticated expert clinicians have subjected the approach to intense scrutiny which has resulted in repeated improvements to the formulation.</p> <p>Relatively simple, clinicians can focus on the numeric grade, and see either a strong or weak recommendation. This two category approach has a clear clinical correlate: the clinician can apply strong clinical recommendations to most patients without hesitation, while careful thought and discussion with the patient are likely to be required for weak recommendations.</p> <p>Linkage of methodological strength with the grade or recommendation reminds clinicians of the importance of considering the strength of evidence in formulation recommendations, and in making clinical decisions.</p> <p>Clinicians have become familiar with this approach because the widespread distribution of the ACCP Antithrombotic Therapy Guidelines.</p>	<p>Since clinicians do not make recommendations for prognosis, the assessment of the quality of evidence for studies evaluating disease prognosis is not practicable with this approach.</p> <p>Guidelines in areas of health care and public health that lack evidence from clinical trials would reveal uniformly Grade C or Grade C+ recommendations and generating the latter grade could include subjective decisions. Although there is little reason to believe that this approach could not be applied to guidelines and recommendations in other areas of health care, previous versions of this approach have been used little outside the antithrombotic therapy use.</p>	Clinicians providing therapy, including trainees in internal medicine, general practitioners, specialists and sub-specialists.

¹⁰ <http://www.chestnet.org/education/hsp/guidelinesProducts.php>

Table 8: Comparison of six systems for grading levels of evidence (continued)

Grading system	Strength of recommendations	Strengths	Weaknesses	Target audiences
CEBM ¹¹ (Centre of Evidence-Based Medicine)	<i>Grade of recommendation</i> is a compression of the 10 'levels' into four 'grades', without any added deliberation or assessment. Level 1a to 1c studies give grade A recommendations; 2a to 3b map to grade B; level 4 studies are grade C and level 5 are imprecise ('minus' level) studies give a grade D recommendation.	Detailed development of the levels of evidence. The different axes allow for questions related to diagnosis, aetiology and prognosis to be considered as "evidence-based" as well as traditionally intervention-oriented recommendations. Partial incorporation of aspects of heterogeneity into the grade of recommendation. The detailed description of the study levels and their objectivity make reproducibility likely to be high. However, this detail may introduce problems for inexperienced users. A study estimating inter-tester reliability has been performed in the Oxford CEMB, and is under analysis (at the time of that report).	The simplistic translation of level of evidence into grade of recommendation. No assessment is made of the clinical importance of the outcomes under consideration. There is no way of balancing of benefits or harms, or assessment of applicability of the studies. There is no clear way of compiling the body of evidence (often of separate levels) into a single grade of recommendation, or differentiation of direct or indirect evidence.	Intended to be used by clinicians in practice. This approach is not intended for use by consumers or policy makers.

¹¹ <http://www.cebm.net/>

Table 8: Comparison of six systems for grading levels of evidence (continued)

Grading system	Strength of recommendations	Strengths	Weaknesses	Target audiences
NHMRC ¹² (National Health and Medical Research Council)	<p>Defines 'strength of evidence' based on level, quality and statistical precision. Guidelines argue against basing 'strength of recommendation' on this alone. Guidelines also recommend against reducing evidence to a single metric that represents 'strength of recommendation'. Instead it is argued that the various dimensions of evidence (strength of evidence, size of effect, and relevance of evidence) should be considered in disaggregated form. It is recommended that relative importance of dimensions be considered in the context of the clinical problem being addressed (e.g. evidence from a good quality RCT may be of limited relevance due to sub-optimal outcomes measured). In that case, the most important basis for a recommendation may be a study from a lower <i>level</i> of evidence that provides a precise estimate of a sizeable effect measured as a change in a highly <i>relevant</i> outcome measure.</p> <p>A checklist that summarises the data and classifies it according to its <i>level, quality, statistical precision, relevance</i> and <i>the size</i> of the treatment effect should accompany each major recommendation. This checklist should reflect the results, where possible, from a formal synthesis of the available evidence. If there is no systematic review of the relevant studies, the data from the best available studies should be rated. A single strength of recommendation rating using A, B, C etc is not advocated in this process.</p>	<p>The advantage of assessing and presenting the evidence in this approach is that decision makers can make up their own minds about the intervention based on the dimensions that appear important to the relevant constituencies. Decision-makers can apply specific weights to a particular dimension that reflects the context in which a decision is being made, ranging from a clinical practice guideline for individual patient care through to a policy decision regarding subsidisation of a medical intervention that may involve expenditure of hundreds of millions of dollars. Combining the dimensions into a single "strength of recommendation" cannot address all viewpoints and preferences.</p>	<p>The main weakness of this method (to some) is that it does not provide a single classification signifying the 'strength of recommendation'.</p> <p>It does not consider fully issues of applicability of results to individual patients. These are covered in a separate guide.</p> <p>The approach does not integrate benefits harms, and costs. The dimensions of evidence on each have to be assessed before they are brought together prior to a decision being made.</p>	<p>This approach was developed for multidisciplinary groups that are preparing clinical practice guidelines under the auspices of the NHMRC. It is also of value to policy makers who require a summary of a body of data.</p>

¹² <http://www.nhmrc.gov.au/>

Table 8: Comparison of six systems for grading levels of evidence (continued)

Grading system	Strength of recommendations	Strengths	Weaknesses	Target audiences
(SIGN) ¹³ Scottish Intercollegiate Guidelines Network	Scale from A to D. 'Grade of recommendation' drawn from <i>level</i> of evidence and 'considered clinical judgment'. This includes size and consistency of body of evidence, its applicability, clinical impact (including economic factors) and generalisability. The recommendation is then assigned a <i>grade</i> , but additionally the wording of the recommendation will reflect the strength of recommendation.	Structural simplicity and potential to discriminate between different study design requirement for different clinical questions. The levels of evidence are likely to be reproducible.	Unstructured formation of grades of recommendation. The definition of 'considered judgment' outlines many areas to be considered. There is a clear explanation of how study quality may limit the grade of recommendation. Assimilation of the other factors is not well described. There is no way of assessing or challenging these considerations, and the method is unlikely to be reproducible.	Intended for use by a wide audience of healthcare providers including doctors, nurses, and managers.
(USPSTF) ¹⁴ US Preventive Services Task Force	Strength of recommendation is rated on a scale of A to D (based on estimate of net benefit (benefit minus harm)) where A is substantial, B is moderate, C is small, and D is zero or negative. Ratings reflect two dimensions: Quality of evidence, and Assessment of balance of harms and benefits.	Analytic frameworks make explicit key questions and can be adopted to other treatment questions. Clear and direct linkage between quality of evidence and strength of recommendation. Recommendations using A-D letter grades. Communicates clearly to clinicians. Considers other elements of evidence beyond study design. Weighs benefits and harms.	The approach with different levels of analysis, can seem complex. Subjective judgments required in integrating different dimensions into an overall assessment of strength of evidence. May not adopt as easily to prognostic/ diagnostic questions. Assessments don't always adjust for individual patient values. Difficult to make recommendations in the absence of good evidence.	Primary target audience is primary care clinicians. Professional organisations, health plans, insurers, quality organisations, purchasers and policy makers have also used the USPSTF recommendations.

¹³ <http://www.sign.ac.uk/>¹⁴ <http://www.ahrq.gov/CLINIC/uspstfix.htm>

Table 8: Comparison of six systems for grading levels of evidence (continued)

Grading system	Strength of recommendations	Strengths	Weaknesses	Target audiences
(USTFCPS) ¹⁵ U.S. Task Force on Community Preventive Services	Strength of evidence of effectiveness links directly to strength of recommendation. Evidence other than effectiveness rarely may be incorporated in Task Force recommendations. For example, an intervention with harms thought by the Task Force to be out of proportion to its benefits would not be recommended even if effective in improving some outcomes.	<p>Contributions of people with a broad range of backgrounds and perspectives limit institutional and individual biases.</p> <p>Many kinds of evidence included (effectiveness, economic evaluations, etc.) provide important information to support decision making.</p> <p>Assessment of effectiveness that incorporates many different factors (e.g. study design, study execution, numbers of studies, etc) allows a broad range of public health interventions to be evaluated in ways that incorporate both scientific rigor, feasibility, and appropriateness of the evaluation.</p> <p>Feasible, evidence-based approach to public health has been argued to be a positive development, bringing 'public health to the same level of scientific scrutiny'.</p>	Complex and costly in terms of time, resources, and expertise required, the possibility that some parts of the process could be seen as arbitrary (e.g. numbers of studies required), that aspects of the process that depend on Task Force opinion might not result in identical conclusions given a different group of decision makers, and that the recommendations unavoidably do not incorporate all of the information that will be important to policy makers.	Primary audience includes persons involved in planning, funding, and implementing population-based services and policies to improve health at the state and local levels in the US. These include federal agencies, state and local health departments, legislators, managed care, and purchasers of health care and public health services.

¹⁵ <http://www.thecommunityguide.org/about/task-force-members.html>

In 2002 the US Agency for Healthcare Research and Quality (AHRQ) evaluated schemes for grading the strength of evidence underlying recommendations. Of the 121 critically evaluated schemes, only 19 for assessing study quality and seven for rating strength of evidence met the AHRQ criteria (West, 2002). The report identified 20 systems for rating the quality of systematic reviews, 49 for RCTs, 19 for observational studies, and 18 for diagnostic test studies. It also identified 40 scales that graded the strength of a body of evidence consisting of one or more studies. The authors of the AHRQ report proposed that any system for grading the strength of evidence should consider quality, quantity, and consistency of the studies. Quality refers to the extent to which the identified studies minimise bias (concept of validity), whereas, quantity is judged by the number of studies and subjects included in those studies. Consistency refers to which findings are similar between different studies on the same topic. Only seven of the 40 systems identified and addressed all three of these key elements (Ebell et al., 2004). As mentioned earlier, this was updated by The Canadian Optimal Medication Prescribing and Utilization Service and recommended only two grading systems for use (the GRADE approach and SIGN). The AHRQ report was a comprehensive review and provided information on the majority of the grading systems being used (West, 2002).

The following are annotated information, from the discussion section of the report that might be of importance when considering analysing results using a grading system. For further detailed information it is advised that the readers consult the original document.

The report by the AHRQ identified factors that are important when developing and using rating systems:

- Distinctions among types of studies, evaluation criteria, and systems
- Numbers of quality rating systems
- Challenges of rating observational studies
- Instrument length
- Reporting guidelines
- Conflicting findings when bodies of evidence contain different types of studies

Overall, many systems covered most of the domains that are considered generally informative for assessing study quality. From this set, the authors identified 19 generic systems that fully address the key quality domains (with the exception of funding or sponsorship for several systems). Three systems were used for both RCTs and observational studies.

In the authors' judgment, those who plan to incorporate study quality into a systematic review, evidence report, or technology assessment can use one or more of these 19 systems as a starting point, being sure to take into account the types of study designs occurring in the articles under review.

The authors identified seven systems that fully addressed all three domains for grading the strength of a body of evidence. The earliest system was published in 1994 (Gyorkos et al., 1994); the remaining systems were published in 1999 (Clarke and Oxman 1999) and 2000 (Briss et al., 2000; Greer et al., 2000; Guyatt et al., 2000; NHS, 2001; Harris et al., 2001), indicating that this is a rapidly evolving field.

Systems for grading the strength of a body of evidence are much less uniform than those for rating study quality. This variability complicates the job of selecting one or more systems that might be put into use today.

The literature has also showed some comparisons of considered judgment forms/processes to formulate recommendations from various organisations that use various grading systems (van der Wees et al., 2007). These comparisons are presented in **Table 9**.

Table 9: Comparison of considered judgment to formulate recommendations from various organisations*

Considered judgment to formulate recommendations		
Organisation	Criteria for considered judgment	Process
CBO ¹⁶	Clinical relevance Safety Patient perspective Professional perspective Available resources Cost-effectiveness Organisation of care Legal consequences Ethical considerations Commercial interest	Considered judgment is described after description of the evidence. Formulation of the recommendation is based on the evidence and the considered judgment. A checklist is available for the development group with detailed criteria within ten domains as listed in this table.
CSP ¹⁷	Strength of evidence Clinical relevance and applicability of evidence Acceptability to patients Benefits and risks Costs	Development group should discuss considerations. When possible, quantitative analysis should be made to estimate relative risks and benefits. Guideline document should describe some of the discussion and clearly describe the link between evidence review and recommendations.
NHMRC ¹⁸	Applicability of the evidence Probable outcome of intervention Balance of benefits against risks Alternative interventions Economic appraisal	A balance sheet is described to balance benefits and harms.
NZGG ¹⁹	Volume of evidence Consistency of the evidence Applicability of the evidence Clinical impact of the intervention	A considered judgment form is used to link clinical questions, evidence and recommendation. The development group needs to make a decision at the beginning of the process about how to resolve differences.
SIGN ²⁰	Quantity, quality and consistency of evidence Generalisability of study findings Directness of application to target population Clinical impact Implementability	Development group summarises view of considered judgment using a form to record their main points. The level of evidence is assigned to the judgment and a graded recommendation is formulated.
USPSTF ²¹	Quality of studies, linkage to key question using three criteria (internal validity, external validity, consistency), linkage to entire preventive service Magnitude and weighing of benefits and harms Extrapolation and generalization Other issues such as cost effectiveness, resource prioritization, logistical factors, ethical and legal concerns, patient and societal expectations should be considered, but recommendations reflect primarily the state of the evidence.	Guideline topic team assesses criteria using systematic methods and rating systems. Recommendations reflect primarily the state of evidence. Making recommendations is done with the understanding that clinicians and policymakers must still consider additional factors in making their own decisions. Setting priorities in clinical practice (e.g. based on resource requirements) are beyond the scope of the review.

* taken from van der Weiss et al., 2007

¹⁶ http://www.cbo.nl/english/default_view¹⁷ <http://www.csp.org.uk/>¹⁸ <http://www.nhmrc.gov.au/>¹⁹ <http://www.nzgg.org.nz/index.cfm?>²⁰ <http://www.sign.ac.uk/>

In a pilot study of the The GRADE Working Group 2005,²² Aitken et al., (2005) approach to grading the quality of evidence and strength of recommendations the study helped to identify and addressed problems with the proposed approach. The pilot study found that it was possible to resolve most of the disagreements when making judgements independently and there was agreement that this approach warrants further development and evaluation. The pilot study identified the criteria for assessing sensibility and understandability for grading evidence and recommendations:

1. To what extent is the approach applicable to different types of questions - effectiveness, harm, diagnosis and prognosis?
2. To what extent can the system be used with different audiences - patients, professionals and policy makers?
3. How clear and simple is the system?
4. How often will information not usually available be necessary?
5. To what extent are subjective decisions needed?
6. Are dimensions included that are not within the construct (level of evidence or strength of recommendation)?
7. Are there important dimensions that should have been included and are not?
8. Is the way in which the included dimensions are aggregated clear and simple?
9. Is the way in which the included dimensions are aggregated appropriate?
10. Are the categories sufficient to discriminate between different levels of evidence and strengths of recommendations?
11. How likely is the system to be successful in discriminating between high and low levels of evidence or strong and weak recommendations?
12. Are assessments reproducible?

Others have argued about using one grading system for all situations, but what are the pros and cons of this approach? **Table 10** presents pros and cons of using the same system for grading evidence and formulating recommendations for a wide range of health care interventions, including clinical and non-clinical interventions (Shunemann et al., 2006).

²¹ (USPSTF) <http://www.ahrq.gov/CLINIC/uspstfix.htm>

²² <http://www.biomedcentral.com/1472-6963/5/25>

Table 10: Pros and cons of using the same system*

Arguments for having a common approach	Arguments against having a common approach
<p>Having less demanding systems for some kinds of questions might result in false positive conclusions.</p> <p>People with vested interests in particular interventions could choose the system that makes their intervention look best.</p> <p>People with vested interests in particular evaluation approaches could choose the system that makes their preferred evaluation approach look best.</p> <p>Having different systems for different types of interventions might be confusing.</p> <p>It is intellectually honest to recognise the limits of evidence where this is appropriate</p> <p>Admitting the limitations of evidence, if this is appropriate, might promote more and better research.</p>	<p>Having an infeasible system for some kinds of questions might result in false negative conclusions.</p> <p>False negative conclusions due to inappropriate evaluation requirements may have negative political and health consequences; for example, effective programs that cannot be studied with randomised trials might experience funding cuts.</p> <p>Interventions that cannot be studied with randomised trials might not be evaluated.</p> <p>A single system might not discriminate adequately within the range of evidence that is appropriate to consider for clinical and non-clinical interventions.</p> <p>A system that can adequately address evidence across a wide range of interventions and contexts may be overly complex.</p>
<p>* Table from Schunemann H, Fretheim A, and Oxman A.. Improving the use of research evidence in guideline development: 9 Grading evidence and recommendations.. Health Research Policy and Systems 2006, 4:21. http://www.health-policy-systems.com/content/4/1/21</p>	

Conclusions

The key results are summarised below:

1) The majority of G-I-N member organisations used the AGREE tool (or the adapted version of the AGREE tool) as the quality system for guideline development programmes. Systematic reviews on the use of evaluation tools for assessment of guidelines also found that AGREE was an optimum tool (supplementary document available upon request).

2) Several international organisations based in the United States, Canada, as well as the World Health Organisation use GRADE as a preferred system for evaluation of evidence. GRADE provides a systematic process to identify, analyse, and present a large body of evidence and a transparent methodological approach for the development of evidence-based optimal therapy recommendations (Shukla et al., 2008).²³ According to an evaluation made by the US Agency for Health Care Research Quality (AHRQ), GRADE and SIGN were among the best evaluated by a group of experts. This finding was also supported by a very recent systematic review by the CADTH (Canada) which adopted the same methodological approach to the AHRQ systematic review of rating systems for grading evidence supported the results above. In personal communication with the authors of this review, they indicated that the CADTH is adopting the GRADE system for new projects. A similar finding was found in a previous work on this topic by WHO on preliminary results from a non-systematic review of the literature on grading evidence and recommendations in guidelines. They have found a large body of work on the development and evaluation of various grading systems, and that GRADE was grading system to grade the quality of evidence and strength of recommendations that is sensible and is being widely used.

3) Personal communication with experts on the evidence-based health electronic discussion group showed that GRADE and SIGN were among the best evaluated, followed by NICE and CEBM (information from the email is available upon request).

4) The AHRQ conducted a comprehensive review on the majority of the grading systems being used (West, 2002). Information from the discussion section of the report that might be of importance when considering analysing results using a grading system is annotated here for reference. For further detailed information, it is advised that the readers consult the original document.

The report by the AHRQ identified factors that are important when developing and using rating systems:

- Distinctions among types of studies, evaluation criteria, and systems
- Numbers of quality rating systems
- Challenges of rating observational studies
- Instrument length
- Reporting guidelines
- Conflicting findings when bodies of evidence contain different types of studies

Overall, many systems covered most of the domains that are considered generally informative for assessing study quality. From this set, the authors identified 19 generic systems that fully address the key quality domains (with the exception of funding or

²³ http://cochrane.org/colloquium/2008/virtual_posters/?s=reviews

sponsorship for several systems). Three systems were used for both randomised controlled trials (RCTs) and observational studies.

In the authors' judgment, those who plan to incorporate study quality into a systematic review, evidence report, or technology assessment can use one or more of these 19 systems as a starting point, being sure to take into account the types of study designs occurring in the articles under review.

The authors identified seven systems that fully addressed all three domains for grading the strength of a body of evidence. The earliest system was published in 1994 (Gyorkos et al., 1994). The remaining systems were published in 1999 (Clarke and Oxman 1999) and 2000 (Briss et al., 2000; Greer et al., 2000; Guyatt et al., 2000; NHS 2001; Harris et al., 2001) indicating that this is a rapidly evolving field.

Systems for grading the strength of a body of evidence are much less uniform than those for rating study quality. This variability complicates the job of selecting one or more systems that might be put into use today.

In conclusion, this brief survey of the existing literature has identified several desirable attributes of a grading system, including ease of use, perceived quality or validity of the grading system, and clarity of the output or time taken. Based on these considerations, HSAC identified the five tools that are more frequently used and highly rated worldwide. In alphabetical order, these are AGREE, GRADE, NICE, OCEBM, 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.

The following sections are the grading systems reported in this document, listed alphabetically and taken from the handbooks or websites of these organisations. A supplementary document with a table listing all G-I-N members and relevant information is available upon request.

List of Systems Most Reported in the Literature

The Agency for Health Care Quality and Research (AHRQ)²⁴

Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment Number 47 (West et al., 2002).

The Agency for Health Care Quality and Research (AHRQ) has proposed that any system assigning levels of evidence should incorporate quality, quantity, and consistency of the evidence (West et al., 2002).

<http://www.ahrq.gov/clinic/uspstf/gradespost.htm>

What the grades mean and suggestions for practice

The U.S. Preventive Services Task Force (USPSTF) has updated its definitions of the grades it assigns to recommendations and now includes "suggestions for practice" associated with each grade. The USPSTF has also defined levels of certainty regarding net benefit. These definitions apply to USPSTF recommendations voted on after May 2007 (*U.S. Preventive Services Task Force Grade Definitions After May 2007*).

Table 11: Grade definitions and suggestions for practice

Grade	Definition	Suggestions for Practice
A	The USPSTF recommends the service. There is high certainty that the net benefit is substantial.	Offer or provide this service.
B	The USPSTF recommends the service. There is high certainty that the net benefit is moderate or there is moderate certainty that the net benefit is moderate to substantial.	Offer or provide this service.
C	The USPSTF recommends against routinely providing the service. There may be considerations that support providing the service in an individual patient. There is at least moderate certainty that the net benefit is small.	Offer or provide this service only if other considerations support the offering or providing the service in an individual patient.
D	The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits.	Discourage the use of this service.
I Statement	The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined.	Read the clinical considerations section of USPSTF Recommendation Statement. If the service is offered, patients should understand the uncertainty about the balance of benefits and harms.

²⁴ <http://www.ahrq.gov/clinic/uspstf/gradespost.htm>

Table 12: Levels of certainty regarding net benefit

Level of Certainty*	Description
High	The available evidence usually includes consistent results from well-designed, well-conducted studies in representative primary care populations. These studies assess the effects of the preventive service on health outcomes. This conclusion is therefore unlikely to be strongly affected by the results of future studies.
Moderate	The available evidence is sufficient to determine the effects of the preventive service on health outcomes, but confidence in the estimate is constrained by such factors as: The number, size, or quality of individual studies' Inconsistency of findings across individual studies, Limited generalisability of findings to routine primary care practice, Lack of coherence in the chain of evidence. As more information becomes available, the magnitude or direction of the observed effect could change, and this change may be large enough to alter the conclusion.
Low	The available evidence is insufficient to assess effects on health outcomes. Evidence is insufficient because of: The limited number or size of studies, Important flaws in study design or methods, Inconsistency of findings across individual studies, Gaps in the chain of evidence, Findings not generalisable to routine primary care practice, Lack of information on important health outcomes, More information may allow estimation of effects on health outcomes.

* The USPSTF defines certainty as "likelihood that the USPSTF assessment of the net benefit of a preventive service is correct." The net benefit is defined as benefit minus harm of the preventive service as implemented in a general, primary care population. The USPSTF assigns a certainty level based on the nature of the overall evidence available to assess the net benefit of a preventive service. *Current as of May 2008*

Grade definitions

Strength of Recommendations

The USPSTF grades its recommendations according to one of five classifications (A, B, C, D, I) reflecting the strength of evidence and magnitude of net benefit (benefits minus harms).

A.— The USPSTF strongly recommends that clinicians provide [the service] to eligible patients. The USPSTF found good evidence that [the service] improves important health outcomes and concludes that benefits substantially outweigh harms.

B.— The USPSTF recommends that clinicians provide [this service] to eligible patients. The USPSTF found at least fair evidence that [the service] improves important health outcomes and concludes that benefits outweigh harms.

C.— The USPSTF makes no recommendation for or against routine provision of [the service]. The USPSTF found at least fair evidence that [the service] can improve health outcomes but concludes that the balance of benefits and harms is too close to justify a general recommendation.

D.— The USPSTF recommends against routinely providing [the service] to asymptomatic patients. The USPSTF found at least fair evidence that [the service] is ineffective or that harms outweigh benefits.

I.— The USPSTF concludes that the evidence is insufficient to recommend for or against routinely providing [the service]. Evidence that the [service] is effective is lacking, of poor quality, or conflicting and the balance of benefits and harms cannot be determined.

Quality of Evidence

The USPSTF grades the quality of the overall evidence for a service on a 3-point scale (good, fair, poor):

Good: Evidence includes consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes.

Fair: Evidence is sufficient to determine effects on health outcomes, but the strength of the evidence is limited by the number, quality, or consistency of the individual studies, generalisability to routine practice, or indirect nature of the evidence on health outcomes.

Poor: Evidence is insufficient to assess the effects on health outcomes because of limited number or power of studies, important flaws in their design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes.

In a recently published paper (Barton et al., 2007) it was noted that the approaches of the GRADE working group and the USPSTF have many elements in common. Both place separate attention on assessing the evidence and making a recommendation on the basis of the evidence. The GRADE approach assigns evidence “quality” at one of four levels: very low, low, moderate, and high, on the basis of specific criteria. The USPSTF assigns evidence “certainty” at one of three levels: high, moderate, and low, on the basis of six critical appraisal questions. The GRADE criteria are similar to the USPSTF’s six questions. The recommendation phase for both GRADE (Aitkins et al., 2004; Aitkens et al., 2005) and the USPSTF rely on a judgment of net benefits (benefits minus harms), including whether net benefits are positive, negative, or uncertain.

GRADE’s process more directly includes costs than the USPSTF approach, although the USPSTF does consider the time and effort of patients and providers. The GRADE working group is developing a system that will apply to many areas, including public health, diagnostic, treatment, and prevention issues, whereas the USPSTF is more narrowly focused on prevention.

Centre of Evidence-Based Medicine (CEBM)²⁵

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

(For definitions of terms used see glossary at <http://www.cebm.net/?o=1116>)

<http://www.cebm.net/index.aspx?o=1025>

Table 13: Levels of Evidence

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-series	Absolute SpPins and SnNouts††	All or none case-series	Absolute better-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

²⁵ <http://www.cebm.net/>

Table 13: Levels of Evidence (continued)

Level	Therapy/ prevention, aetiology/harm	Prognosis	Diagnosis	Differential diagnosis/symp tom prevalence study	Economic and decision 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"

(For definitions of terms used see glossary at <http://www.cebm.net/?o=1116>)

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. <http://www.cebm.net/index.aspx?o=1025>

Notes

Users can add a minus-sign "-" to denote the level 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 one-six months acute, one-five 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.

Cochrane Group²⁶

The *Cochrane Handbook for Systematic Reviews of Interventions* (the *Handbook*) provides guidance to authors for the preparation of Cochrane Intervention reviews (including Cochrane Overviews of reviews). This is Version 5.0.1 of the *Handbook*, last edited 30 September 2008.

As indicated by the Cochrane the quality of the evidence reviewed is assessed using the GRADE approach. Authors will comment on the quality of the body of evidence as ‘High’, ‘Moderate’, ‘Low’, or ‘Very Low’. This is a matter of judgement, but the judgement process operates within a transparent structure (displayed below) and is described in Chapter 12 (Section 12.2)²⁷ of the handbook. As an example, the quality would be ‘High’ if the summary is of several randomized trials with low risk of bias, but the rating of quality becomes lower if there are concerns about design or implementation, imprecision, inconsistency, indirectness, or reporting bias. Authors should use the specific evidence grading system developed by the GRADE collaboration (GRADE Working Group, 2004)²⁸,

The Cochrane uses the GRADE approach as shown below.

Table 14: Levels of quality of a body of evidence in the GRADE approach

Underlying methodology	Quality rating
Randomized trials; or double-upgraded observational studies.	High
Downgraded randomized trials; or upgraded observational studies.	Moderate
Double-downgraded randomized trials; or observational studies.	Low
Triple-downgraded randomized trials; or downgraded observational studies; or case series/case reports.	Very low

Table 15: Factors that may decrease the quality level of a body of evidence

1. Limitations in the design and implementation of available studies suggesting high likelihood of bias.
2. Indirectness of evidence (indirect population, intervention, control, outcomes).
3. Unexplained heterogeneity or inconsistency of results (including problems with subgroup analyses).
4. Imprecision of results (wide confidence intervals).
5. High probability of publication bias.

Table 16: Factors that may increase the quality level of a body of evidence

1. Large magnitude of effect.
2. All plausible confounding would reduce a demonstrated effect or suggest a spurious effect when results show no effect.
3. Dose-response gradient.

²⁶ <http://www.cochrane.org/>

²⁷ <http://www.mrc->

[bsu.cam.ac.uk/cochrane/handbook/chapter_12/12_2_assessing_the_quality_of_a_body_of_evidence.htm](http://www.mrc-bsu.cam.ac.uk/cochrane/handbook/chapter_12/12_2_assessing_the_quality_of_a_body_of_evidence.htm)

²⁸ <http://www.gradeworkinggroup.org/>

Factors that decrease the quality level of a body of evidence

Limitations in the design and implementation: Every study addressing a particular outcome will differ, to some degree, in the risk of bias. Review authors must make an overall judgement on whether the quality of evidence for an outcome warrants downgrading on the basis of study limitations.

Indirectness of evidence: Two types of indirectness are relevant., For example in a review comparing the effectiveness of an intervention for secondary prevention of coronary heart disease, the majority of identified studies happened to be in people who also had diabetes; the evidence may be regarded as indirect in relation to the broader question of interest because the population is restricted to people with diabetes. The opposite can apply in which a review addressing the effect of a preventative strategy for coronary heart disease in people with diabetes may consider trials in people without diabetes to provide relevant, albeit indirect, evidence. Other sources of indirectness may arise from interventions studied, comparators used, and outcomes assessed.

Unexplained heterogeneity or inconsistency of results: this occurs when studies yield widely differing estimates of effect; investigators should look for robust explanations for that heterogeneity. When heterogeneity exists and affects the interpretation of results, but authors fail to identify a plausible explanation, the quality of evidence decreases.

Imprecision of results: occurs when studies include few participants and few events and thus have wide confidence intervals; authors can lower their rating of the quality of the evidence.

High probability of publication bias: The quality of evidence level may be downgraded if investigators fail to report studies (typically those that show no effect: publication bias) or outcomes (typically those that may be harmful or for which no effect was observed: selective outcome reporting bias) on the basis of results.

A particular body of evidence can suffer from problems associated with more than one of the five factors above, and the greater the problems, the lower the quality of evidence rating that should result. One could imagine a situation in which randomised trials were available, but all or virtually all of these limitations would be present, and in serious form. A very low quality of evidence rating would result.

CTFPHC Methods (CTFPHC)²⁹

The Canadian Task Force for Preventive Health Care strives to provide a bridge between research findings and clinical preventive practice. When research does not provide clear guidance, this lack of evidence is articulated. A major objective is to help physicians choose tests, immunizations, counselling strategies and other preventive interventions of proven utility and avoid those that lack demonstrated value.

The CTFPHC uses a standardized methodology, employing explicit analytic criteria, for evaluating the effectiveness of preventive health care interventions. Key features are to:

- Make recommendations of graded strength, based on the quality of published medical evidence for a discussion of these recommendation grades, please link to the 2003 article in the Canadian Medical Association Journal³⁰.
- Place greatest weight on the features of study design and analysis that tend to eliminate or minimize biased results tables 17, 18 and 19 (adapted from CTFPHC)³¹ provide a summary of the CTFPHC's grades of recommendations, quality of evidence, and analytic criteria.

Table 17: Recommendation grades for specific clinical preventive actions

A	The CTF concludes that there is good evidence to recommend the clinical preventive action.
B	The CTF concludes that there is fair evidence to recommend the clinical preventive action.
C	The CTF concludes that the existing evidence is conflicting and does not allow making a recommendation for or against use of the clinical preventive action, however other factors may influence decision-making.
D	The CTF concludes that there is fair evidence to recommend against the clinical preventive action.
E	The CTF concludes that there is good evidence to recommend against the clinical preventive action.
I	The CTF concludes that there is insufficient evidence (in quantity and/or quality) to make a recommendation, however other factors may influence decision-making.
The CTF recognizes that in many cases patient specific factors need to be considered and discussed, such as the value the patient places on the clinical preventive action; its possible positive and negative outcomes; and the context and /or personal circumstances of the patient (medical and other). In certain circumstances where the evidence is complex, conflicting or insufficient, a more detailed discussion may be required.	

Table 18: Levels of evidence - research design rating

I	Evidence from randomized controlled trial(s)
II-1	Evidence from controlled trial(s) without randomization
II-2	Evidence from cohort or case-control analytic studies, preferably from more than one centre or research group
II-3	Evidence from comparisons between times or places with or without the intervention; dramatic results in uncontrolled experiments could be included here
III	Opinions of respected authorities, based on clinical experience; descriptive studies or reports of expert committees

²⁹ <http://www.ctfphc.org/>

³⁰ <http://www.cmaj.ca/cgi/content/full/169/3/207>

³¹ <http://www.ctfphc.org/ctfphc&methods.htm>

**Table 19: Levels of evidence - quality (internal validity) rating
(see Harris et al., 2001)**

Good	A study (including meta-analyses or systematic reviews) that meets all design-specific criteria* well.
Fair	A study (including meta-analyses or systematic reviews) that does not meet (or it is not clear that it meets) at least one design-specific criterion* but has no known "fatal flaw".
Poor	A study (including meta-analyses or systematic reviews) that has at least one design-specific* "fatal flaw", or an accumulation of lesser flaws to the extent that the results of the study are not deemed able to inform recommendations.
*General design-specific criteria are outlined in Harris et al., 2001 ³² .	

Some challenges for evidence-based prevention

The CTFPHC has identified that a number of important issues arise during and following the development of clinical preventive guidelines:

There are a relatively large number of "C" and "I" Recommendations (due to insufficient, inconclusive, or conflicting evidence), leaving clinicians to make decisions on other grounds.

There is a need to consider **all** varieties of benefit and harm associated with any preventive manoeuvre, including improved quality or length of life, anxiety relieved or money saved, cost, "labelling" and anxiety, including that induced by earlier diagnosis

A "C" or "I" Recommendation can serve as a caution to those who have to decide which preventive measures justify public funding

Many preventive interventions that have the potential to improve health lie outside the context of the clinician-patient encounter - the prevention of poverty, of violence and of pollution are striking examples

While cost of care is an inescapable and serious consideration, economic analysis of clinical preventive actions is complex and not yet fully developed. It is not yet a major focus of Task Force evaluations. Choices may have to be made, on both monetary and ethical grounds between preventive interventions for unrelated conditions.

³² <http://www.ahrq.gov/clinic/ajpmsuppl/review.pdf>

The Grading of Recommendations Assessment, Development and Evaluation Working Group (GRADE)³³

The latest conclusion from the literature indicated that the GRADE provides a systematic process to identify, analyse, and present a large body of evidence and a transparent methodological approach for the development of evidence-based optimal therapy recommendations (Cochrane Colloquium 2008). Table 20 shows a comparison of GRADE to other grading systems, followed by Table 21 which shows the grid for recording panellists' views in development of guidelines, and finally Table 22 displays the list of the organisations using the GRADE system.

Table 20: GRADE comparison to other grading systems

Factor	Other systems	GRADE	Advantages of GRADE system*
Definitions	Implicit definitions of quality (level) of evidence and strength of recommendation	Explicit definitions	Makes clear what grades indicate and what should be considered in making these judgments
Judgments	Implicit judgments regarding which outcomes are important, quality of evidence for each important outcome, overall quality of evidence, balance between benefits and harms, and value of incremental benefits	Sequential, explicit judgments	Clarifies each of these judgments and reduces risks of introducing errors or bias that can arise when they are made implicitly
Key components of quality of evidence	Not considered for each important outcome. Judgments about quality of evidence are often based on study design alone	Systematic and explicit consideration of study design, study quality, consistency, and directness of evidence in judgments about quality of evidence	Ensures these factors are considered appropriately
Other factors that can affect quality of evidence	Not explicitly taken into account	Explicit consideration of imprecise or sparse data, reporting bias, strength of association, evidence of a dose-response gradient, and plausible confounding	Ensures consideration of other factors
Overall quality of evidence	Implicitly based on the quality of evidence for benefits	Based on the lowest quality of evidence for any of the outcomes that are critical to making a decision	Reduces likelihood of mislabeling overall quality of evidence when evidence for a critical outcome is lacking
Relative importance of outcomes	Considered implicitly	Explicit judgments about which outcomes are critical, which ones are important but not critical, and which ones are unimportant and can be ignored	Ensures appropriate consideration of each outcome when grading overall quality of evidence and strength of recommendations

³³ <http://www.gradeworkinggroup.org/index.htm>

Table 20: GRADE comparison to other grading systems (continued)

Factor	Other systems	GRADE	Advantages of GRADE system*
Balance between health benefits and harms	Not explicitly considered	Explicit consideration of trade-offs between important benefits and harms, the quality of evidence for these, translation of evidence into specific circumstances, and certainty of baseline risks	Clarifies and improves transparency of judgments on harms and benefits
Whether incremental health benefits are worth the costs	Not explicitly considered	Explicit consideration after first considering whether there are net health benefits	Ensures that judgments about value of net health benefits are transparent
Summaries of evidence and findings	Inconsistent presentation	Consistent GRADE evidence profiles, including quality assessment and summary of findings	Ensures that all panel members base their judgments on same information and that this information is available to others
Extent of use	Seldom used by more than one organisation and little, if any empirical evaluation	International collaboration across wide range of organisations in development and evaluation	Builds on previous experience to achieve a system that is more sensible, reliable, and widely applicable

*Most other approaches do not include any of these advantages, although some may incorporate some of these advantages.

Table 21: GRADE grid for recording panellists' views in development of guidelines

GRADE SCORE					
	1	2	0	2	1
Balance between desirable and undesirable consequences of intervention	Desirable clearly outweigh undesirable	Desirable probably outweigh undesirable	Trade-offs equally balanced or uncertain	Undesirable probably outweigh desirable	Undesirable clearly outweigh desirable
Recommendation	Strong: definitely do it	Weak: probably do it	No specific recommendation	Weak: probably don't do it	Strong: "definitely don't do it"

Box 1 Factors that influence the strength of recommendation

Balance between desirable and undesirable effects—The larger the difference between the desirable and undesirable effects, the more likely a strong recommendation is warranted. The narrower the gradient, the more likely a weak recommendation is warranted.

Quality of evidence—The higher the quality of evidence, the more likely a strong recommendation is warranted.

Values and preferences—The more variability in values and preferences, or more uncertainty in values and preferences, the more likely a weak recommendation is warranted.

Costs (resource allocation)—The higher the costs of an intervention (that is, the more resources consumed) the less likely a strong recommendation is warranted.

The following organisations have endorsed or are using GRADE* [GRADE] system.

Table 22: Organisations using GRADE system

Country	Organisation/website
International	<p>European Society of Thoracic Surgeons http://www.ests.org/</p> <p>JIDC Journal of Infection in Developing Countries http://www.oloep.org/content.asp?id=687</p> <p>Kidney Disease: Improving Global Outcome http://www.kdigo.org/</p> <p>The Cochrane Collaboration http://www.cochrane.org/</p> <p>WHO http://whqlibdoc.who.int/hq/2003/EIP_GPE_EQC_2003_1.pdf</p> <p>Surviving Sepsis http://www.survivingsepsis.org/</p>
USA	<p>Endocrine Society Clinical Guidelines http://www.endo-society.org/</p> <p>American College of Chest Physicians Guidelines http://www.chestnet.org/education/hsp/gradingSystem.php</p> <p>UpToDate - Putting Clinical Information Into Practice http://www.uptodate.com/home/about/policies/editorial_policy.html</p> <p>American Thoracic Society http://www.thoracic.org/sections/publications/statements/docs-committee/index.html</p> <p>American College of Physicians http://www.acponline.org/</p> <p>Agency for Healthcare Research and Quality (AHRQ) http://www.ahrq.gov/</p> <p>Society of Critical Care Medicine (SCCM) http://www.sccm.org/</p> <p>The University of Pennsylvania Health System Center for Evidence-based Practice http://www.ups.upenn.edu/cep/</p> <p>Society for Vascular Surgery http://www.vascularweb.org/</p> <p>Infectious Diseases Society of America http://www.idsociety.org/</p> <p>Emergency Medical Services for Children National Resource Center http://www.childrensnational.org/EMSC/</p>
Italy	<p>Agenzia sanitaria regionale, Bologna http://asr.regione.emilia-romagna.it/wcm/asr/eventi/2006/20060504_sem_gradeprior.htm</p> <p>Evidence-based Nursing Südtirol, Alto Adige http://www.provincia.bz.it/</p>
Canada	<p>Ministry of Health and Long-Term Care, Ontario http://www.health.gov.on.ca/english/providers/program/mas/mas_mn.html</p> <p>COMPUS at The Canadian Agency for Drugs and Technologies in Health (CADTH) http://cadth.ca/index.php/en/compus</p>
Germany	<p>Ärztliches Zentrum für Qualität in der Medizin http://www.aezq.de/?set_language=en&cl=en</p> <p>German Center for Evidence-based Nursing "sapere aude" http://www.medizin.uni-halle.de/pflegewissenschaft/index.php?id=346</p>

What assessment tools are used both in New Zealand and in other countries for grading of evidence?

Table 22: Organisations using GRADE system (continued)

Country	Organisation/website
UK**	British Medical Journal http://resources.bmj.com/bmj/authors/article-submission/article-requirements BMJ Clinical Evidence http://www.clinicalevidence.com/ National Institute for Clinical Excellence (NICE) http://www.nice.org.uk/
Norway	Norwegian Knowledge Centre for the Health Services http://www.kunnskapssenteret.no/Prosjekter/1599.cms
Finland/International	EBM Guidelines http://ebmg.wiley.com/ebmg/ltk.koti
Poland	Polish Institute for EBM http://ebm.org.pl/show.php?aid=15258&_tc=491AFBFBBE2645DAA2DE93CA9C423714
Europe	European Respiratory Society (ERS) http://ersnet.org/
Japan	Japanese Society for Temporomandibular Joint http://wwwsoc.nii.ac.jp/jstmj/
Sweden	National Board of Health and Welfare http://www.socialstyrelsen.se/en/
Spain	Spanish Society for Family and Community Medicine http://www.semfyec.es/es/
*some endorsements have included minor modifications, most commonly collapsing low and very low quality evidence into a single category, ** Groups submitting guidelines to the BMJ are encouraged to use GRADE	

Medical Services Advisory Committee (MSAC) the National Health and Medical Research Council (NHMRC³⁴)

In 2000 the Australian National Health and Medical Research Council (NHMRC) adopted a handbook on the review of the evidence to systematically identify and review the scientific literature (NHMRC, 2000a). This handbook describes how systematically to identify scientific literature relevant to a particular question, select and review the most important (highest quality) studies, and summarise and present the results for further consideration by the committee that will develop the clinical practice guidelines. The handbook states that it is important to note which study types are the most appropriate to answer different types of questions. While RCTs (or systematic reviews of RCTs) are the most appropriate study types for intervention questions, other study types are more appropriate for answering non-intervention questions. Table 23 (taken from NHMRC 2000a, page 10) summarises the most appropriate study types for specific questions and the major appraisal issues associated with each study type.

Table 23: Types of clinical and public health questions, ideal study types and major appraisal issues (taken from NHMRC 200a, page 10)

Question	Study types	Major appraisal issues
Intervention	Systematic review, RCTs, Cohort study, Case-control study	Randomisation, follow-up complete, blinding of patients and clinicians
Frequency/rate (burden of illness)	Systematic review, Cohort study, Cross-sectional study	Sample frame, case ascertainment, adequate response/follow-up achieved
Diagnostic test performance	Systematic review, cross-sectional study (random or consecutive sample)	Independent, blind comparison with 'gold standard', appropriate selection of patients
Aetiology and risk factors	Systematic review, cohort study, case-control study	Groups only differ in exposure, outcomes measurement, reasonable evidence for causation
Prediction and prognosis	Systematic review, cohort/survival study	Inception cohort, sufficient follow-up

The strength of the evidence is determined by the level and quality of the evidence in combination with the statistical precision. While the level of evidence is based on the study design from which the evidence was obtained, the quality is based on the study's type-specific questions. The magnitude of the effect is based on the size of the p value, whereas precision is considered by the width of the confidence interval (CI).

Another important issue is whether the evidence is relevant to the original study questions. A revised version of the Handbook states that the relevance includes the appropriateness of outcome measures and whether the intervention has been applied to the appropriate population. The Handbook has classified the relevance of the evidence based on a system which is most relevant for intervention questions (see **Table 24** from NHMRC (2000) p28) and (NHMRC, 2000^b).

³⁴ <http://www.nhmrc.gov.au/>

Table 24: Classifying the relevance of the evidence

Ranking	Relevance of the evidence
1	Evidence of an effect on patient-relevant clinical outcomes, including benefit and harms, and quality of life and survival
2	Evidence of an effect on a surrogate outcome that has been shown to be predictive of patient-relevant outcomes for the same intervention
3	Evidence of an effect on proven surrogate outcomes but for a different intervention
4	Evidence of an effect on proven surrogate outcomes but for a different intervention and population
5	Evidence confined to unproven surrogate outcomes

National Institute for Health and Clinical Excellence (NICE)³⁵ from the National Health Services (NHS)

NICE issues high quality guidelines based on a systematic review of the evidence and have extensive consultation not only with clinicians but also with patients and, where relevant, industry. Professional associations do not have the resources to carry out this type of consultation, but they can follow the principles set out in the AGREE protocol, which helps guideline writers minimise bias, meet the needs of all stakeholders, and maximise clarity (NHS Evidence 2009).

The guidelines manual has been updated after public consultation. The 2009 edition of ‘The guidelines manual’ describes the process and methods used for all clinical guidelines starting scoping after 5 January 2009. Guidelines already in development at this date will switch to the methods and processes described in the 2009 edition when the draft documents are being prepared for consultation. The draft and published full guidelines will specify which edition of ‘The guidelines manual’ was used for each stage of development (National Institute for Health and Clinical Excellence, 2009) (Guidelines Manual 2009).

The following is an extract from the manual:

The manual describes how to review the evidence after identification of studies during a comprehensive literature search. The studies need to be reviewed to identify the most appropriate data to help address the review question, and to ensure that the guideline recommendations are based on the best available evidence. The process of reviewing the evidence involves selecting relevant studies, assessing their quality, synthesizing and interpreting the results. The study selection process for clinical studies and economic evaluations includes documentation and giving details of inclusion criteria that were applied. The process for sifting and selecting economic evaluations for assessment is the same as for clinical studies.

In assessing the quality of studies (which is defined as the degree of confidence about the estimate of a treatment effect), the quality criteria and ways of summarising the data are slightly different between the clinical effectiveness and cost-effectiveness.

From chapter six section 2(1):

Assessing study quality for clinical effectiveness: Study quality can be defined as the degree of confidence about the estimate of a treatment effect.

³⁵

<http://www.nice.org.uk/aboutnice/howwework/developingniceclinicalguidelines/clinicalguidelinedevelopmentmethods/GuidelinesManual2009.jsp>

The first stage is to determine the study design so that the appropriate criteria can be applied in the assessment. Because it is sometimes difficult to identify the exact design used in a study, a checklist is provided to help the systematic reviewer to classify study design for answering questions of effectiveness (see Appendix B from NICE Guidelines Manual 2009).

Once a study has been classified, it should be assessed using the methodology checklist for that type of study (see appendices C–F). To minimise errors and any potential bias in the assessment, two reviewers should independently assess a random selection of studies. Any differences arising from this should be discussed fully at a GDG meeting.

The quality of a study can vary depending on which of its measured outcomes is being considered. Well-conducted randomised controlled trials are more likely than non-randomised studies to produce similar comparison groups, and are therefore particularly suited to estimating the effects of interventions. However, short-term outcomes may be less susceptible to bias than long-term outcomes because of greater loss to follow-up with the latter.

The New Zealand Guidelines Group (NZGG)³⁶

The New Zealand Guidelines Group, uses the AGREE instrument which is a generic tool designed primarily to help guideline developers and users assess the methodological quality of clinical practice guidelines. The AGREE Instrument assesses both the quality of the reporting, and the quality of some aspects of the recommendations. It provides an assessment of the predicted validity of a guideline; that is, the likelihood that it will achieve its intended outcome. The New Zealand Guidelines Group uses the following grades of recommendations:

Grades of recommendation

Grades indicate the strength of the supporting evidence rather than the importance of the evidence.

A - The recommendation is supported by good evidence (based on a number of studies that are valid, consistent, applicable and clinically relevant).


B - The recommendation is supported by fair evidence (based on studies that are valid, but there are some concerns about the volume, consistency, applicability and clinical relevance of the evidence that may cause some uncertainty but are not likely to be overturned by other evidence).

C - The recommendation is supported by international expert opinion.

Good Practice Points (GPP) - Where no evidence is available, best practice recommendations are made based on the experience of the Guideline Development Team, or feedback from consultation within New Zealand.

³⁶ <http://www.nzgg.org.nz/index.cfm?>

Table 25: The NZGG considered judgment form to grade evidence

 S I G N	Considered judgement on quality of evidence	
Key question:	Evidence table ref:	
1. Volume of evidence Comment here on any issues concerning the quantity of evidence available on this topic and its methodological quality.		
2. Applicability Comment here on the extent to which the evidence is directly applicable to the NHS in Scotland.		
3. Generalisability Comment here on how reasonable it is to generalise from the results of the studies used as evidence to the target population for this guideline.		
4. Consistency Comment here on the degree of consistency demonstrated by the available of evidence. Where there are conflicting results, indicate how the group formed a judgement as to the overall direction of the evidence		
5. Clinical impact Comment here on the potential clinical impact that the intervention in question might have – e.g. size of patient population; magnitude of effect; relative benefit over other management options; resource implications; balance of risk and benefit.		
6. Other factors Indicate here any other factors that you took into account when assessing the evidence base.		
7. Evidence statement Please summarise the development group's synthesis of the evidence relating to this key question, taking all the above factors into account, and indicate the evidence level which applies.	Evidence level	
8. Recommendation What recommendation(s) does the guideline development group draw from this evidence? Please indicate the grade of recommendation(s) and any dissenting opinion within the group.	Grade of recommendation	

Scottish Intercollegiate Guideline Network (SIGN)

The Scottish Intercollegiate Guidelines Network (SIGN) was established in 1993 by the Conference (later, the Academy) of Royal Colleges and their Faculties in Scotland, to develop evidence-based clinical guidelines for the National Health Service (NHS) in Scotland (SIGN 50 handbook 2008). This followed the publication of a report by the Clinical Resource and Audit Group (CRAG) which highlighted the need for national, evidence based clinical guidelines to be developed by “the Royal Colleges, the specialist associations of the healthcare professionals and relevant educational bodies”. SIGN has evolved significantly since then but remains a collaborative initiative a network of clinicians, patients’ representatives and other healthcare professionals, including all the medical specialties, nursing, pharmacy, dentistry, professions allied to medicine, and NHS management. Patients are represented on SIGN by Voluntary Health Scotland and lay representation. The current membership of SIGN Council is noted on the website: www.sign.ac.uk

This is the third revision of SIGN 50, previous versions having been issued in 2002 and 2004. SIGN 50 is structured to follow the guideline development process from beginning to end, taking each step in turn it starts with the context of guideline development in Scotland, and progresses from first proposal of a new topic to final publication and implementation of the guideline.

The SIGN methodology complies with the criteria used by the AGREE (Appraisal of Guidelines for Research and Evaluation in Europe) to identify good quality guidelines.

Chapter 7 of the handbook describes the process of forming guideline recommendations in which levels of evidence and grades of recommendations are explained. SIGN formerly used the levels of evidence developed by the US Agency for Health Care Policy and Research (AHCPR, now the US Agency for Health Research and Quality, AHRQ). As a number of limitations were becoming apparent in that system, a review was carried out and new levels of evidence and associated grades of recommendation were developed. Following extensive consultation and international peer review, the new grading system was introduced in autumn 2000. The current grading system is shown below (taken from the handbook Annex B).

The grading system is an improvement on the previous system, but still has weaknesses that need to be addressed. SIGN has been participating in the international GRADE project aimed at developing a methodologically sound system that can be applied across countries and cultures.

Whether and to what extent the GRADE should be adopted by SIGN is under discussion, but whatever is decided there remains a problem in dealing with different types of evidence. GRADE addresses evidence of effectiveness where it is possible to clearly quantify benefits and harms. In other questions addressed by guidelines evidence is more likely to be presented in narrative form. As the grading system develops, means of dealing with both types of evidence in a rigorous manner will be required. Whatever changes are made are likely to be evolutionary rather than revolutionary in nature (SIGN, 2008).

Table 26: SIGN grading system

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 of RCTs or RCTs with a low risk of bias
1-	Meta-analyses, systematic reviews of RCTs 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, bias or chance and with a high probability that the relationship is causal
2+	Well-conducted case-control or cohort studies with a low risk of confounding, bias or chance and with a moderate probability that the relationship is causal
2-	Case-control or cohort studies with a high risk of confounding, bias or chance and with a significant risk that the relationship is not causal
2	Non-analytic studies, such as case reports and case series
4	Expert opinion
RCT: randomized controlled trial.	
GRADES OF RECOMMENDATION	
A	At least one meta-analysis, systematic review, or RCT rated as 1++ and directly applicable to the target population OR A systematic review of RCTs or a body of evidence consisting principally of studies rated as 1+ directly applicable to the target population and demonstrating overall 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+
<i>Source:</i> Scottish Intercollegiate Guidelines Network	

Strength of Recommendation Taxonomy (SORT)

The SORT Addresses the quality, quantity, and consistency of evidence and allows authors to rate individual studies or bodies of evidence. The taxonomy is built around the information mastery framework, which emphasizes the use of patient-oriented outcomes that measure changes in morbidity or mortality. An A-level recommendation is based on consistent and good quality patient-oriented evidence; a B-level recommendation is based on inconsistent or limited quality patient-oriented evidence; and a C-level recommendation is based on consensus, usual practice, opinion, disease-oriented evidence, or case series for studies of diagnosis, treatment, prevention, or screening. Levels of evidence from 1 to 3 for individual studies also are defined. (Ebell MH, Siwek J, Weiss BD, Woolf SH, Susman J, Ewigman B, Bowman M. Strength of Recommendation Taxonomy (SORT): A Patient-Centered Approach to Grading Evidence in the Medical Literature. *J Am Board Fam Pract* 2004; 17:59-67)

Table 27: Definitions of strength of recommendations

Strength of Recommendation	Definition
A	Recommendation based on consistent and good quality patient-oriented evidence*
B	Recommendation based on inconsistent or limited quality patient-oriented evidence*
C	Recommendation based on consensus, usual practice, opinion, disease-oriented evidence,* and case series for studies of diagnosis, treatment, prevention, or screening.

Table 28: Definitions of levels of evidence for each study type

	Type of study		
	Diagnosis	Treatment/ Prevention/ Screening	Prognosis
Study Quality			
Level 1 Good quality patient-oriented evidence	-Validated clinical decision rule -Systematic review (SR)/meta-analysis of high quality studies -High quality diagnostic cohort study**	-SR/meta-analysis of RCTs with consistent findings -High quality individual RCT+ -All or none study ++	-SR/meta-analysis of good quality cohort studies -Prospective cohort study with good follow-up
Level 2 Limited quality patient-oriented evidence	-Unvalidated clinical decision rule -SR/meta-analysis of lower quality studies or studies with inconsistent findings -Lower quality diagnostic cohort study or diagnostic case control study**	-SR/meta-analysis of lower quality clinical trials or of studies with inconsistent findings -Lower quality clinical trial+ -Cohort study -Case-control study	-SR/meta-analysis of lower quality cohort studies or with inconsistent results -Retrospective cohort study or prospective cohort study with poor follow-up -Case-control study -Case series
Level 3 Other evidence	Consensus guidelines, extrapolations from bench research, usual practice, opinion, disease-oriented evidence (intermediate or physiologic outcomes only), and case series for studies of diagnosis, treatment, prevention, or screening.		

Notes for Table 27 and Table 28

*Patient-oriented evidence measures outcomes that matter to patient: morbidity, mortality, symptom improvement, cost reduction, equality of life. Disease-oriented evidence measures intermediate, physiologic, or surrogate endpoints that may or may not reflect improvements in patient outcomes (i.e. blood pressure, blood chemistry, physiological function, and pathological findings).

**High quality diagnostic cohort study: cohort design, adequate size, adequate spectrum of patients, blinding, and a consistent, well-defined reference standard.

+High quality RCT: allocation concealed, blinding if possible, intention-to-treat analysis, adequate statistical power, adequate follow-up (>80%)

++An all-or-none study is one where the treatment causes a dramatic change in outcomes, such as antibiotics for meningitis or surgery for appendicitis, which precludes study in a controlled trial.

Table 29: Definitions of consistency across studies

	Consistency across studies
Consistent	-Most studies found similar or at least coherent conclusions (coherence means that differences are explainable) OR -If high quality and up-to-date systematic reviews or meta-analyses exist, they support the recommendation
Inconsistent	-Considerable variation among study findings and lack of coherence OR -If high quality and up-to-date systematic reviews or meta-analyses exist, they do not find consistent evidence in favour of the recommendation

The World Health Organization (WHO)

The method used to assess the quality of the evidence by WHO is through weighting according to the GRADE rating scheme (see **Table 30**).

Table 30: GRADE quality assessment criteria

Quality of evidence	Study design	Lower If*	Higher If*
High	Randomised trial	Study quality: -1 serious limitations -2 very serious limitations -1 important inconsistency Directness: -1 some uncertainty -2 Major uncertainty -1 Sparse data -1 High probability of Reporting Bias	Strong association: +1 Strong, no plausible confounders, consistent and direct evidence** +2 very strong, no major threats to validity and direct evidence*** +1 Evidence of a Dose response gradient +1 All plausible confounders would have reduced the effect
Moderate			
Low	Observational study		

*1 = move up or down one grade (for example, from high to intermediate); 2 = move up or down two grades (for example, from high to low)
**A statistically significant relative risk of >2 (<0.5), based on consistent evidence from two or more observational studies, with no plausible confounders.
***A statistically significant relative risk of >5(<0.2) based on direct evidence with no major threats to validity.

METHODS USED TO ANALYSE THE EVIDENCE

Systematic Review

Systematic Review with Evidence Tables

References

- American College of Chest Physicians (n.d). ACCP Guidelines and Evidence-Based Products. Retrieved from <http://www.chestnet.org/index.php>
- Atkins, D., Briss, P. A., Eccles, M., Flottorp, S., Guyatt, G. H., Harbour, R. T., et al. (2005). Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system. *BMC Health Services Research*, 5(1), 25.
- Atkins, D., Eccles, M., Flottorp, S., Guyatt, G., Henry, D., Hill, S., et al. (2004). Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches The GRADE Working Group. *BMC Health Services Research*, 4(1), 38.
- Baker A, Young K, Potter J, Madan A. (2009) A review of grading systems and critical appraisal tools for use by specialist medical societies developing evidence-based guidelines. London: NHS Plus. Available from: <http://www.nhsplus.nhs.uk/public/images/library/files/ReviewofGradingSystemsandCriticalAppraisalTools>
- Barton, M. B., Miller, T., Wolff, T., Petitti, D., LeFevre, M., Sawaya, G., et al. (2007). How to Read the New Recommendation Statement: Methods Update from the U.S. Preventive Services Task Force. *Ann Intern Med*, 0000605-200707170-200700171.
- Briss, P. A., Zaza, S., Pappaioanou, M., Fielding, J., Wright-De Agüero, L., Truman, B. I., et al. (2000). Developing an evidence-based guide to community preventive services: methods. *American Journal of Preventive Medicine*, 18(1, Supplement 1), 35-43.
- Canadian Task Force on the Periodic Health Examination (1979). The periodic health examination. *CMAJ* 121:1193-1254.
- CEBM (2009). Oxford Centre for Evidence-based Medicine. Levels of Evidence. Retrieved from <http://www.cebm.net/index.aspx?o=1025>
- Clarke, M. and Oxman, A.D. (editors), The Cochrane Collaboration (1999). *Cochrane Reviewers' Handbook* Version 4.0.
- Ebell, M. H., Siwek, J., Weiss, B. D., Woolf, S. H., Susman, J., Ewigman, B., et al. (2004). Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *American family physician*, 69(3), 548-556.
- EPIQ Group – Critical Appraisal and Evidence-based practice. Auckland School of Population Health. Retrieved June 2009 from <http://www.fmhs.auckland.ac.nz/soph/depts/epi/epiq/ebp.aspx>
- GRADE Working Group (2004). Grading quality of evidence and strength of recommendations. *BMJ*, 328:1490.
- Greer, N., Mosser, G., Logan, G., & Halaas, G. W. (2000). A practical approach to evidence grading. *Joint Commission Journal on Quality and Patient Safety*, 26, 700-712.
- Guyatt, G. H., Haynes, R. B., Jaeschke, R. Z., Cook, D. J., Green, L., Naylor, C. D., et al. (2000). Users' guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care. *Jama*, 284(10), 1290-1296.
- Gyorkos, T. W., Tannenbaum, T. N., Abrahamowicz, M., Oxman, A. D., Scott, E. A., Millson, M. E., et al. (1994). An approach to the development of practice guidelines for community health interventions. *Can J Public Health*, 85 Suppl 1, S8-13.

- Harbour, R., & Miller, J. (2001). A new system for grading recommendations in evidence based guidelines. *BMJ (Clinical research ed)*, 323(7308), 334-336.
- Harris, R. P., Helfand, M., Woolf, S. H., Lohr, K. N., Mulrow, C. D., Teutsch, S. M., et al. (2001). Current methods of the US Preventive Services Task Force: a review of the process. *Am J Prev Med*, 20(3 Suppl), 21-35.
- Higgins, J. P. T., Green, S. (editors), The Cochrane Collaboration (2008). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.0.1 Retrieved from www.cochrane-handbook.org.
- Horvath, A. R. (2009). Grading Quality of Evidence and Strength of Recommendations for Diagnostic Tests and Strategies. *Clin Chem*, 55(5), 853-855.
- National Institute for Health and Clinical Excellence (2009) The guidelines manual. London: National Institute for Health and Clinical Excellence. Available from: www.nice.org.uk
- New Zealand Guidelines Group (2003). Handbook for the preparation of explicit evidence-based clinical practice guidelines. Retrieved June 2009, from http://www.nzgg.org.nz/download/files/nzgg_guideline_handbook.pdf
- NHS Research and Development Centre of Evidence-Based Medicine (n.d) Levels of Evidence. Accessed January 12, 2001, from <http://cebm.jr2.ox.ac.uk>.
- Oxman, A., Fretheim, A., Schunemann, H., & Sure (2006). Improving the use of research evidence in guideline development: introduction. *Health Research Policy and Systems*, 4(1), 12.
- Palda, V. A. M. D. M., Davis, D. M. D., & Goldman, J. M. (2007). A guide to the Canadian Medical Association Handbook on Clinical Practice Guidelines. *Cmaj*, 177(10), 1221-1226.
- NHMRC (2008 -2009). Additional levels of evidence and grades for recommendations for developers of guidelines. Stage 2: consultation from NHMRC: http://www.nhmrc.gov.au/guidelines/_files/Stage%202%20Consultation%20Levels%20and%20Grades.pdf
- Schunemann, H. J., Fretheim, A., & Oxman, A. D. (2006). Improving the use of research evidence in guideline development: 9. Grading evidence and recommendations. *Health research policy and systems / BioMed Central*, 4, 21.
- Shukla, V.K., Bai, A., Milne, S., Wells, G. (2008) Systematic Review of Evidence Grading Systems for Grading Levels of Evidence. Cochrane Colloquium, poster 21. Retrieved from http://cochrane.org/colloquium/2008/virtual_posters/?poster=21
- SIGN. (2008). *SIGN 50. A guideline developers' handbook*. Edinburgh: Scottish Intercollegiate Guidelines Network. Retrieved from <http://www.sign.ac.uk/pdf/sign50.pdf>
- U.S. Preventive Services Task Force (n.d). *Grade definitions. Guide to clinical preventive services, third edition: periodic updates, 2000-2003*. Agency for Healthcare Research and Quality, Rockville, MD. Retrieved from <http://www.ahrq.gov/clinic/3rduspstf/ratings.htm>
- U.S. Preventive Services Task Force (2008). *Grade definitions after May 2007*. Agency for Healthcare Research and Quality, Rockville, MD. Retrieved from <http://www.ahrq.gov/clinic/uspstf/gradespost.htm>

- Van der Wees, P., Hendriks, E., Custers, J., Burgers, J., Dekker, J., & de Bie, R. (2007). Comparison of international guideline programs to evaluate and update the Dutch program for clinical guideline development in physical therapy. *BMC health services research*, 7(1), 191.
- West, S., King, V., Carey, T. S., Lohr, K. N., McKoy, N., Sutton, S. F., et al. (2002). Systems to rate the strength of scientific evidence. *Evid Rep Technol Assess (Summ)*(47), 1-11.