Published in final edited form as:
Semin Arthritis Rheum. 2011 October ; 41(2): 95–105. doi:10.1016/j.semarthrit.2010.12.001.

Methods of Formal Consensus in Classification/Diagnostic Criteria and Guideline Development

Raj Nair[1,*], **Rohit Aggarwal**[2,*], and **Dinesh Khanna**[3]

[1]Department of Medicine, Washington Hospital Center, Washington, DC

[2]Department of Medicine, University of Pittsburgh, PA

[3]Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA

Abstract

Guideline or diagnostic criteria in clinical practice assist physicians in their clinical decision-making and improve health outcomes for patients. Diagnostic and classification criteria are based on evidence from rigorously conducted controlled studies. Formal group consensus methods have been developed to organize subjective judgments and to synthesize them with the available evidence. This review discusses four types of formal consensus methods used in the health field and their applications in rheumatology: the Delphi method, nominal group technique, RAND/UCLA Appropriateness Method, and National Institutes of Health consensus development conference.

Physicians are faced with diagnostic and treatment dilemmas on a daily basis in their clinic setting. Well developed diagnostic or treatment guidelines in clinical practice summarize data for physicians to improve their clinical decision-making for an individual patient. (1). Evidence-based-guidelines derived from rigorously conducted controlled studies outperform expert opinion; however, in practice there are instances where sufficient research-based evidence does not exist (2). Various group consensus methods incorporate opinions of a group of experts rather than an individual expert in a formalized manner (3, 4). Given the diversity of opinion that can occur with the diagnosis and management of disease, formal group consensus methods organize subjective judgments and to supplement the available evidence. Formal group consensus methods allow for inclusion of a wide range of knowledge and experience, interaction between members, stimulate constructive debate, and prevent influential behavior of one opinion to formulate suggestions about a specific question where there is insufficient evidence. *Since perfect agreement is seldom reached, the objective of the consensus methodology is to identify a central tendency among the group and grade the level of agreement reached.* Consensus methods are increasingly being used to develop diagnostic, classification, therapeutic guidelines and response criteria sets in various fields of medicine including rheumatology (5, 6).

Correspondence: Dinesh Khanna, MD, MS, Division of Rheumatology, Department of Medicine, David Geffen School of Medicine, 1000 Veteran Avenue, Rm 32-59 Rehabilitation Building, Los Angeles, CA 90095, Phone: (310) 825-3061, Fax: (310) 206-8606, dkhanna@mednet.ucla.edu.

*Both authors contributed equally to the manuscript

Competing Interests

None of the authors have competing interests related to the content of this manuscript.

Our review will focus on details of the consensus methodologies along with the relative advantages and disadvantages of each method. There are other formal quantitative methods to combine information from the literature such as meta-analysis and systematic review and that the purpose of this review is to concentrate on consensus when other quantitative methods are not possible either due to paucity or conflict in the literature.

Defining consensus

Consensus does not need to be defined as full agreement among participants (7). A pre-specified range is decided by the group running or leading the consensus methodology, according to the needs of the task to be performed. There are several methods for defining consensus with options including overall consensus of the group and consensus on each topic being considered by the group (8).

- A final vote determining percent agreement (e.g., 80%) amongst participants
- Rating scale- a specified mean rating must be achieved on each topic for inclusion by group
- A majority of participants must give a topic a certain rating for inclusion of a topic (as detailed later in the RAND-UCLA methodology)

Participants

The composition of the group is important in determining the decision reached. Most agree that a participant should be an expert and have credibility in the appropriate field (8, 9). An expert can be a clinician with vast clinical expertise, a researcher who is well versed with the current literature, or a lay person or patient who has experienced the impact of the disease or intervention or condition in question (4, 10). Ideal characteristics of a group participant would be researchers and clinicians with expertise in the field providing face validity and facilitate dissemination of consensus meeting results (8). In addition, the participants should be enthusiastic about the project and understand the demands and responsibilities of it.

Heterogeneity of the participants (different levels of expertise within a field, different fields of expertise, different geographical areas and population ethnicity or different health care policies in different countries) has advantages and disadvantages to the consensus process. A diverse group may lead to better performance than homogeneity in terms of considering all relevant areas of the topic (11) but may also lead to disagreement among group participants due to different diagnostic and therapeutic approaches based on geographical area. Also, a participant with higher authority may exert a greater degree of influence on the group when consensus group is comprised of people of diverse status, whereas consensus derived by a more uniform group will be more likely to reflect the majority view (12). In our view, heterogeneity of the participants brings more credibility to the consensus methodology and a more robust, creative product is developed.

In addition, larger groups will increase the reliability of the judgment (13), but may be difficult to coordinate and more time consuming. In general, reliability of consensus recommendations declines rapidly when the group size falls below 6, and above 12, improvement in reliability is not substantial (13). An exception is the Delphi and nominal group techniques that can recruit larger number of participants.

Importance of Evidence from Literature when using Group Consensus

Although group consensus participants are generally recruited on the basis that they have superior knowledge of the published literature in the field, it is essential to supplement

participants with up-to-date, evidence based literature. When literature reviews are provided, the evidence is used both in the initial discussion and in later decision making (14). A literature review (preferably a systematic review) also provides a common starting point for the group and increases the perception that the process should be as evidence based as possible (10). The amount of information varies from no information because available evidence is weak to comprehensive synthesis of relevant research when there is abundant information in literature but no general consensus (8). Any level of information, however small is likely to influence the group decision (12), and thus should be provided.

In the development of clinical guidelines for care, incorporating an evidence base has been clearly shown to result in better guidelines. In a retrospective study comparing sets of guidelines, six sets of breast cancer guidelines were classified as evidence based, consensus based, or a combination of both. Instruments to measure quality of guidelines were used to grade the sets of guidelines. The evidence based guidelines scored highest on quality whereas strictly consensus based guidelines scored the lowest (15).

Formal Consensus Methods

The focus will be on four types of formal consensus methods used in the health field: Delphi method, Nominal Group Technique, RAND/UCLA Appropriateness Method (RAM), and National Institutes of Health (NIH) consensus development conference methodology (Table 1). In practice a combination of two formal consensus methods or their modifications can be used in a two step process, where one method is used for item generation or some initial consensus and then other method is used for final consensus. When there has not been rigid adherence to one consensus methodology, the description of the procedure in the methods section is referred to as a “modified” consensus method.

Delphi Methods or Technique

The Delphi method, originated in the 1960s, takes its name from the Delphic oracle's skills of interpretation and foresight (16). It was developed at the RAND Corporation to obtain the most reliable consensus of opinion of a group of experts on a subject in a systematic manner (17, 18). It attempts to do this by series of well-defined questionnaires based on surveys and feedback. Usually Delphi is applied when consensus among large number participants are needed who cannot be grouped into a single room for logistic or economic reasons (19).

Process (Figure 1)

The first step is to define a specific problem or question to be solved by the Delphi process (20). The purpose for the Delphi should be made clear. For example, in case of criteria set development it is important to make it explicit that the goal of the Delphi process is to develop diagnostic/ classification criteria set for the specific clinical entity.

Prior to Round 1

The team (or Steering Committee) undertaking the Delphi invites clinical or research experts to provide opinions on a specific matter and participate in subsequent questionnaire rounds. The panel of participants must be adequate (can be hundreds, but at least 10–30) and should represent multi-disciplinary leaders from various geographic areas on the particular subject in consideration. Participants should be motivated and interested in the problem to be solved, which will help ensure adequate response rates to the survey.

Sometimes a thorough literature search is carried out by the team undertaking the Delphi process to study the existing evidence, which may help to define the first set of questions. The results of the literature search and the relevant original publications are sent to the

participant to facilitate the process. Given the privacy issues it has now becoming standard first to send an invitation letter to the group for participation in the Delphi process. This initial mail helps establish relationship, verify e-mail address, and provide the denominator that will be used to calculate the response rate For example, in an **ongoing** effort to develop outcome measures for connective tissue disease-interstitial lung disease clinical trials, appropriate expert panelists were identified via an international investigation of specialist associations, study investigators of multicenter endpoint studies in parenchymal lung disease, members of respective consensus committees and task forces, dedicated to parenchymal and rheumatologic lung disease, members of disease specific groups as well as authors of key publications, in journals and textbooks. Patients with connective tissue disease-interstitial lung disease are also included.

Round 1

The participants are asked broad open-ended questions on the Delphi topic to elicit different opinions. These questions are usually more descriptive to elicit the expert's opinions, novel ideas, and their experience on a specific sub-heading/question. The opinions gathered are grouped together under a limited number of headings and statements drafted for circulation to all participants in a questionnaire form. This initial survey should be as simple as possible and is ideally no longer than a page.

Researchers can send detailed instructions of the process and ask the participants to provide group and items relevant to the project. Alternately, they can propose certain groups (e.g., organ systems and patient reported outcome measures as done recently in scleroderma (21) relevant to the study and have the participants propose additional groups and items.

Round 2

The replies from the participants in round 1 are analyzed with simple descriptive statistics and summarized to generate a series of statements. At this point there may be consensus on some issues. Those issues not reaching consensus are developed into a focused second questionnaire and sent back to the individual participants. Feedback from round 1 is also reported to the participants. Depending upon the question, the participants either answers focused questions in yes/no manner to reach immediate consensus or rank their agreement or disagreement with each statement (or items) in the second round. The questions should be clearly and concisely stated and response elicited should be in a simple and straightforward fashion.

Round 3

In the second survey, the panelists are provided with the overall results and his/her previous reply/scores. Experts then reconsider their previous opinion and re-rate their level of agreement with each statement in the questionnaire, with the opportunity to change their score in view of the group's response. The re-ratings are summarized and assessed for degree of consensus: if an acceptable degree of pre-defined consensus (which should be predefined before initiation of the Delphi) or when sufficient information exchange has been obtained is obtained, the process may cease, with final results fed back to participants; if not, the third round is repeated. Ideally consensus should be reached by round 3 but in theory many more rounds can be done.

Researchers have used a pre-defined cut of the mean/ median values (21) or done cluster analysis to discriminate important from unimportant items (22).

In summary, the process is conducted usually over 2 to 4 "rounds," with the results elicited, tabulated, and reported to the group after each round.

Delphi Technique

Advantages	Disadvantages
• Large number of participants possible (23) • Each participant expresses their opinion freely and impersonally • Personal contact between experts • Limits dominance by eminent, eloquent, or highly opinionated individuals in the field • Less likely that the moderator of the panel may bias the group • Substantial amount of time to express ideas, reflect upon answers and make changes • Cheap, convenient and no geographical constraints • Easy to understand, flexible, and can be applied to broad range of topics • Can be used preceding NGT meeting for initial item generation	• Generalizability of the study findings (external validity) • Dependent on questionnaire design • Vulnerability with respect to who is an "expert" • Obliviousness to reliability measurement and scientific validation of findings (24) • Potential for bias exists in participant selection • Consensus panel judgments influenced by panel composition and by feedback given during the panel process (25) • Coordinating large groups and several rounds can be complicated and costly • Delphi does not allow any personal contact between the experts

Examples in Rheumatology

Delphi strategies have been used to solve an array of problems in health and medicine, from the needs of an individual hospital or department to those of a statewide agency or state (23, 24). Delphi technique has been also widely used in the past in the internal medicine as well as rheumatology literature to develop diagnosis and management guidelines. Specifically, in the rheumatology literature (25–27), Delphi has been used for criteria set development especially for reaching consensus on the outcome and response criteria set in diseases like systemic sclerosis, systemic lupus erythematosus, gout, ankylosing spondylitis and psoriatic arthritis. Use of Delphi exercise in classification criteria set development has been limited and is mostly used in initial phase to gather data and opinion in an open-ended and systematic manner or to make initial consensus on broad categories of variables which are further narrowed down by other consensus methods like NGT in the second part of the consensus development process (5, 6, 28–30).

Recommendations for use—Delphi exercise is best used for reaching consensus on variables to be used in outcome or response criteria set.

Nominal Group Technique

Nominal Group Technique (NGT) is derived from social-psychological studies of decision conferences and management science studies. It is a face-to-face structured group meeting of experts that is led by an experienced moderator. (8). NGT was historically used in the social services, government, education, and industry since the 1960s.

In NGT, a question is generated for which an expert panel is convened. The expert panel may consist of 5–9 panelists or could be larger. Larger groups can be separated into several groups (or tables) of 5–9 participants which work simultaneously on the same questions (31).

In brief, NGT asks for silent and independent generation of ideas, the round-robin listing the ideas, serial discussion led by a moderator, and independent ranking of the ideas. NGT is preferably done in a room with rectangular table arranged in an open “U” with flip chart or

large screen computer at the open end of the table. Each panelist is asked to generate ideas to specific questions and record privately on a piece of paper for 5–10 minutes. In the second phase, in round robin fashion, participants are asked to provide each of their ideas which are listed on a flip chart or wide screened computer visible to the entire group. No discussion is conducted at this time. In the third phase, a serial brief discussion is led by the moderator; the goal of the discussion is clarification of the ideas or statements (31). During the last phase, each idea is privately ranked or rated on a scale of 1–5 or 1–10. The highest ranking solutions will be kept while the remaining solutions are discarded. Since the NGT can be influenced by strong personalities, the process requires that at each round the first person to speak is different from the previous one; this means that the first round will start with the first person to the left of the moderator, the second round with the second person to the left of the moderator etc. In this way all people will have the possibility to speak first and avoid the unduly influence of strong personalities (19). There are no guidelines regarding the cut-off for what constitutes a high vs. low ranking. Generally 70– 80% consensus is required (32) although researchers may chose to have a lower cut-off but this should be pre-defined (21).

It is important to remember that the purpose of NGT is to establish a prioritization of ideas and issues, and the use of numerical voting can assist in this (33). However, the temptation to attach greater meaning to the numbers, and carry out more sophisticated quantitative analysis should be resisted.

There are several variations of this process. Meetings can occur in several rounds with initial responses collected via mail (Delphi technique) followed by a future face-to-face-meeting. After a discussion period, additional rounds of generating ideas or re-ranking responses can occur. As used with RAND/UCLA method (detailed later), synthesized evidence can be provided to the expert panel prior to submitting solutions for the question that has been generated.

Nominal Group Technique

Advantages	Disadvantages
• Participants meet face-to-face • All participants have an opportunity to voice opinions • Personal contact between experts • Design of NGT does not allow any individual to dominate • Group voting can occur if desired in later rounds	• Certain members of the panel can take over discussion and drive results-experienced moderator required • Limited by time— only a few questions can be discussed and agreed upon • Economic and time costs associated with face-to-face meeting • Limited to providing a solution to a few problems limits its applicability to multiple scenarios

Examples in Rheumatology

While often used in combination with Delphi methodology, there are some classification criteria sets developed with NGT for classification of childhood vasculitis and juvenile systemic sclerosis (5, 6). Other pediatric rheumatologic conditions for which criteria have been developed also include juvenile SLE (34, 35), juvenile idiopathic arthritis (36, 37) and inflammatory myositis (29, 38, 39). Combination of Delphi and NGT have also been used in development of provisional core set items (21). Other examples of NGT used in rheumatology include developing criteria for clinical outcomes in rheumatoid arthritis with the Outcome Measures in Rheumatoid Arthritis Clinical Trials (OMERACT) committee (40,

41) as well as development of the British Isles Lupus Assessment Group's disease activity index (BILAG) (42).

Recommendations for use

1. Primary use is to be give priority to questions/ items to be discussed. It can be adapted to be used for other circumstances such as
    a. Developing classification and response criteria.
    b. Developing consensus practice guidelines.
    c. Developing ideas for exploratory research (43)

This method is often used in combination with the Delphi method as described above. Pure NGT is useful when solutions need to be developed at a face-to-face meeting and expert opinion is needed more than evidence which may be lacking for the particular topic.

RAND-UCLA Appropriateness Method (RAM)

This was a method of group consensus developed in the 1980s by RAND (research and development) Corporation and UCLA (University of California-Los Angeles) (44). Using current scientific evidence in conjunction with expert opinion, this consensus method was initially developed to evaluate the overuse/underuse of medical or surgical procedures. This procedure has been used in the United States for assessing appropriateness of procedures such as coronary angiography, carotid endarterectomy, hysterectomy, and upper gastrointestinal tract endoscopy (45–47).

This process usually involves two interdependent groups: a core panel and an expert panel. The core panel guides the expert panel through the tasks of the RAM and provides synthesized data to the expert panel and the expert panel uses the data provided by the core panel to come to a consensus.

Prior to the consensus process, the core panel will conduct a systematic literature review with evidence synthesis to provide the expert panel with all pertinent information that will guide evidence-based decision-making (48). Then, a list of clinical scenarios (indications) is developed to provide the expert panel. These clinical scenarios will describe a patient with a set of characteristic features and the expert panel will be given a Likert-scale (typically 9-point) that will ask whether a particular intervention is appropriate for the patient.

The expert panel was historically restricted to 9 panelists but can range from 7–15 panelists (44). In general, an odd number is preferred to prevent a “tie.” It is recommended that a multi-disciplinary panel is selected so that bias from a like-minded group with identical agenda is avoided (49).

There are two rounds of rating appropriateness of an intervention. In the first round, the expert panel receives the clinical scenarios by mail (or via. internet as discussed in the Delphi section) and are asked to rate the appropriateness (on a 1 to 9 Likert scale) of an intervention. In RAM, rating between 1–3 is considered “inappropriate” (risks outweigh benefits), 7–9 is considered “appropriate” (benefits outweigh risks), and 4–6 is “uncertain”. They are asked to do this independent of the other panelists. They are allowed to use the synthesized evidence provided by the core panel overseeing the consensus process. The panel is asked not to consider the cost of performing the procedures or intervention in rating of their scenarios. There are other definitions proposed by RAM but this works best for a panel of 7–15 members and most widely used.

The second round is conducted as a face-to-face meeting over 1–2 days and is led by an experienced moderator. Usually 7–11 panelists are sufficient and it is recommended to have an odd number so there is no "tie." All panel experts are given results of other expert's individual ratings for all clinical scenarios. Panel experts are given an opportunity to discuss individual views on the appropriateness of the intervention in each clinical scenario. At the end of the discussion each panelist can reconsider their original rating and re-rate the clinical scenario. The results are then summarized as descriptive statistics. Disagreement is when one-third or more panelists rate scenario in lowest 3 points and one-third or more rate same scenario in highest 3 points of appropriateness scale. In the absence of disagreement median rating in lower 3 points is "inappropriate", median rating in upper 3 points is "appropriate", and rating in the 4–6 range or those with disagreement is "uncertain." These ratings can finally be used to determine whether an intervention has been used inappropriately retrospectively or determine whether an intervention should be performed prospectively. Based on the number of participants included, there can be some variation in the definitions of agreement.

Based on the ratings of the expert panel an indication in each clinical scenario will be considered, appropriate, uncertain, or inappropriate, however it lacks clear rating of the evidence or a ranking of the scientific robustness of the recommendations.

While RAM was initially used for assessing appropriateness of a procedure, it has been applied to the development of practice guidelines and classification criteria. Once indications have been ranked by appropriateness using a Likert scale as described above, summary statements reflecting the output from the consensus exercise are used to make a set of classification criteria or practice guidelines (50). In making summary statements, both the perceived appropriateness of a certain intervention or guideline as well as the agreement among participants is critical.

RAND-UCLA Appropriateness Method

Advantages	Disadvantages
• Synthesis of published literature prior to consensus techniques incorporated • Allows for both confidential ratings as well as group discussion • Multi-disciplinary panel encourage consensus from a wider group • Reproducibility of RAM ranges from moderate to excellent as determined by different panelists for "appropriate" and "inappropriate" care (53). • Acceptable predictive validity for a recommendation supported by RCTs(54).	• Misclassification is expected (49) • Takes great deal of time from gathering of the evidence to multiple rounds of consensus. • Face-to-face which can add cost/time delay and lead to highly opinionated individuals in the field dominating the discussion • Requires third party (core panel) to construct clinical indications for an intervention and analyze/ interpret the results from the expert panel meeting • 9-point Likert scale can be cumbersome • Requires voting on multiple case scenarios (sometimes > 1,000)

Examples in Rheumatology

This methodology has been used for development of quality indicators in rheumatology (51, 52). Recommendations have been developed using this methodology including recommendations for clinical agent evaluation for treatment in osteoporosis which was endorsed by the American Society for Bone and Mineral Research (ASBMR), the International Society for Clinical Densitometry (ISCD), and the National Osteoporosis Foundation (NOF) (53). Recently, quality indicators for systemic lupus erythematosus (54)

have been published as well as recommendations for evaluating and treating rheumatoid arthritis (10, 55).

Recommendations for use—Any project that needs a formal method for combining scientific evidence from the literature and expert consensus, might consider RAM. Specifically, it can be used in the:

1. Developing clinical practice guidelines where evidence is not sufficient.
2. Developing quality of care indicators.
3. Determining appropriate use criteria for a medical intervention.
4. Determining overuse or underuse of a procedure.

This rigorous method can be used when resources and time are available and can be used in any circumstance where thorough combination of scientific evidence and expert opinion is required.

Consensus Development Conference (CDC)

In 1997, the National Institute of Health (NIH) in the United States introduced the consensus development conference (CDC) (8). The CDC brings together selected experts in the field and concerned individuals (from expert to public) to reach general agreement about the safety, efficacy, and appropriateness of using various medical procedures, drugs, and devices. The Office for Medical Applications of Research (OMAR) of the NIH is charged with responsibility of this conference (56). OMAR's primary goal is to help bring the results of biomedical research into direct use in the practice of medicine. To this end, it acts to coordinate consensus development activities at the NIH for evaluating current evidence and promulgating consensus of opinions on a particular topic. Each conference is jointly sponsored and administered by one or more Institutes or Centers (ICs) of NIH and by the OMAR. They have run more than 100 conferences on a variety of topics (57). The development of consensus conferences is a hybrid of methods used by judicial decision-making, scientific conferences and the town hall meeting (58).

Process

A topic for the CDC, usually of public health importance, is decided by agreement between the sponsoring IC and OMAR. A set of predetermined questions on which a consensus is warranted are selected from the topic by experts appointed by sponsoring IC and OMAR, which defines the scope of the conference. A different selected group of experts (around 10 experts) is invited and brought together to form an expert panel or decision making group for the conference. The panel is an independent, broad-based, non-Department of Health and Human Services (DHHS), non-advocacy group with appropriate expertise. The panel members are highly regarded in their own fields but are not closely aligned with the subject. The evidence is then presented to the conference by various experts appointed by sponsoring IC and OMAR, who are not a member of the decision making group. Usually, the Agency for Healthcare Research and Quality (AHRQ) provides a systematic review of literature on the conference topic through one of its Evidence-Based Practice Centers. The group or expert panel hears the scientific data presented by invited experts as well as the comments from the general public in a public (open) session followed by discussion. The format allows the participation of the audience (members of the public) in an open meeting, who can ask questions to the group. The panel then meets in a private (executive) session to further deliberate on the evidence and discussion to reach a consensus. The chairperson is the moderator responsible for guiding and controlling the proceedings of both the open part of the conference and the private group discussion as well as helping reaching consensus on

topic/question where 2 or more experts differ in their opinion. The panel weighs the information and reaches a consensus statement that addresses a set of predetermined questions. The consensus statement draft is then presented in a plenary session and is subject to review and comment by conference attendees. Following the discussion, the panel may then modify the statement if appropriate in the final executive session. The final consensus statement is then released and disseminated widely to achieve maximum impact on health care practice and medical research. This is an independent report and is not considered a policy of NIH or the Federal Government. A major contribution of the NIH consensus panels has been to describe current levels of agreement on important topics like coronary artery bypass surgery, intraocular lens implantation, cesarean section, Reye' syndrome, and the treatment of breast cancer.

Consensus Development Conference

Advantages	Disadvantages
• Mix of practicing physicians, researchers, consumers, and others to come together and jointly evaluate an existing technology • Wide circulation through both lay and medical media • Unbiased panel	• Interaction is not structured • The aggregation methodology used is implicit- a formal feedback system is lacking • CDC has not been used for making new criteria sets

Examples in Rheumatology

Use of CDC in the field of rheumatology has been limited to systemic review of diagnosis and therapeutic options on the topics like osteoporosis, total knee and hip replacement as well as recommendations for IVIG therapy (68–71). There have been no criteria set developed by NIH CDC meeting in any field of medicine including rheumatology.

Recommendations for use

1. CDC conferences are particularly useful for providing guidance when a controversy exists over preventive, therapeutic, or diagnostic options for *public policy*
2. When the issue is of high degree of public interest or has impact on health care cost.

It is intended as a statement that reflects the views of expert panel that have examined and discussed the scientific data, rather than a legal document or a practice guideline. The statement may reflect uncertainties, options, or minority viewpoints.

Use in Rheumatology—Use of CDC in the field of rheumatology has been limited to systemic review of diagnosis and therapeutic options on the topics like osteoporosis, total knee and hip replacement as well as recommendations for IVIG therapy (59–62). There have been no criteria set developed by NIH CDC meeting in any field of medicine including rheumatology.

Grading of Recommendations Assessment, Development and Evaluation (GRADE)

This review also discusses GRADE as it has been adopted by scientific world for the development of recommendations. *GRADE is not a consensus methodology per se but uses consensus methods discussed above (especially Delphi and NGT) to assess quality and strength of a recommendation.* GRADE methodology provides a systematic and transparent approach to rate the quality of evidence and grade the strength of recommendations for

patient important outcomes. GRADE was developed by experts with a goal to answer some drawbacks of previous methods, which include the lack of separation between quality of evidence and strength of recommendation and the lack of explicit acknowledgment of values and preferences (63). Several organizations and guideline developers, including the World Health Organization, the American College of Chest Physicians the American Endocrine Society, and UpToDate, have adopted the GRADE system. GRADE Quality of evidence is graded into 1 of the 4 levels— high, moderate, low, and very low. RCTs usually get a higher grading compared to observational studies. The system allows the quality of evidence derived from observational data to be upgraded from low to moderate or high categories and the quality of evidence coming from randomized trials to be downgraded depending on the design and execution of such studies. The strength of recommendation is divided into strong (benefits far outweigh risks or that risks far outweigh benefits and virtually all patients will make same decisions) and weak (where as tradeoffs between risks and benefits is less clear and patient's values and preferences are incorporated before making a treatment plan). Factors influencing strength of recommendations may include quality of evidence, balance between desirable and undesirable effects, individual's values and preferences for a treatment, and resource utilization (63, 64).

Examples in Rheumatology

American College of Rheumatology guidelines in osteoarthritis (unpublished). Other examples in the literature include use in reaching decisions on practice guidelines (65) and grading recommendations for the use of thrombolytic agents and prevention of coronary artery disease (64, 66).

Recommendations for use

1. To develop clinical guidelines.

Limitations of Review

Although we present a comprehensive review of formal consensus methods, we didn't conduct a systematic review. In our comprehensive review, the majority of the articles on formal consensus methodologies are implementations in diagnostic criteria, guidelines, and response indices development.

Conclusion

Formal consensus methodologies have various applications in the medical literature from guideline development to development of criteria sets. These methods are being increasingly used in rheumatology. Each methodology has it unique attributes and utilization of a particular methodology or combination of methodologies depends on: 1. clinical question, 2: audience, and 3: available resources.

When closely tied to evidence-based literature, these summarized statements can provide physicians with guidance in classifying and treating disease. Dangers with group consensus occur when there is not rigor in the methods used and when the ensuing recommendations are non-specific (67). Also, consensus methodologies provide areas/ strength of agreement (and disagreement) and should not replace evidence-based literature.

Acknowledgments

We would like to acknowledge Drs. Carol Wallace and Gillian Hawker for providing constructive feedback during the writing of this manuscript.

Funding Source

No funding was obtained for the study design, collection, analysis and interpretation of data; in the writing of the manuscript; and in the decision to submit the manuscript for publication.

References

1. Lugtenberg M, Burgers JS, Westert GP. Effects of evidence-based clinical practice guidelines on quality of care: A systematic review. Br Med J. 2009; 18(5):385.
2. Chassin, M. Appropriate investigation and treatment in clinical practice. London: Royal College of Physicians; 1989. How do we decide whether an investigation or procedure is appropriate; p. 21-29.
3. Agency for Healthcare Policy and Research. AHCPR clinical practice guideline program. report to congress. 1995.
4. Black N, Murphy M, Lamping D, McKee M, Sanderson C, Askham J, et al. Consensus development methods: A review of best practice in creating clinical guidelines. J Health Serv Res Policy. 1999 Oct; 4(4):236–248. [PubMed: 10623041]
5. Ozen S, Ruperto N, Dillon MJ, Bagga A, Barron K, Davin JC, et al. EULAR/PReS endorsed consensus criteria for the classification of childhood vasculitides. Ann Rheum Dis. 2006 Jul; 65(7): 936–941. [PubMed: 16322081]
6. Zulian F, Woo P, Athreya BH, Laxer RM, Medsger TA Jr, Lehman TJ, et al. The pediatric rheumatology european Society/American college of Rheumatology/European league against rheumatism provisional classification criteria for juvenile systemic sclerosis. Arthritis Rheum. 2007 Mar 15; 57(2):203–212. [PubMed: 17330294]
7. Linstone, HA.; Turoff, M. The delphi method : Techniques and applications. Reading, Mass: Addison-Wesley Pub. Co.; 1975. Advanced Book Program
8. Fink A, Kosecoff J, Chassin M, Brook RH. Consensus methods: Characteristics and guidelines for use. Am J Public Health. 1984 Sep; 74(9):979–983. [PubMed: 6380323]
9. Jones J, Hunter D. Consensus methods for medical and health services research. BMJ. 1995 Aug 5; 311(7001):376–380. [PubMed: 7640549]
10. Saag KG, Teng GG, Patkar NM, Anuntiyo J, Finney C, Curtis JR, et al. American college of rheumatology 2008 recommendations for the use of nonbiologic and biologic disease-modifying antirheumatic drugs in rheumatoid arthritis. Arthritis Rheum. 2008 Jun 15; 59(6):762–784. [PubMed: 18512708]
11. Jackson, S. Team composition in organizational settings: Issues in managing an increasingly diverse work force. In: Worchel, S.; Wood, W.; Simpson, J., editors. Group Process and Productivity. Sage; 1992. p. 138-173.
12. Vinokur A, Burnstein E, Sechrest L, Wortman PM. Group decision making by experts: Field study of panels evaluating medical technologies. J Pers Soc Psychol. 1985 Jul; 49(1):70–84. [PubMed: 4020617]
13. Kahan JP, Park RE, Leape LL, Bernstein SJ, Hilborne LH, Parker L, et al. Variations by specialty in physician ratings of the appropriateness and necessity of indications for procedures. Med Care. 1996 Jun; 34(6):512–523. [PubMed: 8656718]
14. Jacoby I. Evidence and consensus. JAMA. 1988 May 27.259(20):3039. [PubMed: 3367478]
15. Cruse H, Winiarek M, Marshburn J, Clark O, Djulbegovic B. Quality and methods of developing practice guidelines. BMC Health Serv Res. 2002; 2(1):1. [PubMed: 11825346]
16. Dalkey, NC. Delphi. Santa Monica, Calif.: Rand; 1967.
17. Helmer-Hirschberg, O. The use of the delphi technique in problems of educational innovations. Santa Monica, Calif.: Rand; 1966.
18. Dalkey, NC.; Helmer-Hirschberg, O. An experimental application of the delphi method to the use of experts. Santa Monica, Calif.: Rand Corp.; 1962.
19. Delbecq, AL.; Van de Ven, AH.; Gustafson, DH. Group techniques for program planning: A guide to nominal group and delphi processes. Middleton, WI: Green Briar Press; 1975.
20. Murphy MK, Black NA, Lamping DL, McKee CM, Sanderson CF, Askham J, et al. Consensus development methods, and their use in clinical guideline development. Health Technol Assess. 1998; 2:i–iv. 1–88. [PubMed: 9561895]

21. Khanna D, Lovell DJ, Giannini E, Clements PJ, Merkel PA, Seibold JR, et al. Development of a provisional core set of response measures for clinical trials of systemic sclerosis. Ann Rheum Dis. 2008 May; 67(5):703–709. [PubMed: 17893248]
22. Distler O, Behrens F, Pittrow D, Huscher D, Denton CP, Foeldvari I, et al. Defining appropriate outcome measures in pulmonary arterial hypertension related to systemic sclerosis: A delphi consensus study with cluster analysis. Arthritis Rheum. 2008 Jun 15; 59(6):867–875. [PubMed: 18512721]
23. Loughlin KG, Moore LF. Using delphi to achieve congruent objectives and activities in a pediatrics department. J Med Educ. 1979 Feb; 54(2):101–106. [PubMed: 762686]
24. Thomson WA, Ponder LD. Use of delphi methodology to generate a survey instrument to identify priorities for state allied health associations. Allied Health Behav Sci. 1979; 2(4):383–399. [PubMed: 10297887]
25. Bertsias G, Ioannidis JP, Boletis J, Bombardieri S, Cervera R, Dostal C, et al. EULAR recommendations for the management of systemic lupus erythematosus. report of a task force of the EULAR standing committee for international clinical studies including therapeutics. Ann Rheum Dis. 2008 Feb; 67(2):195–205. [PubMed: 17504841]
26. Gazi H, Pope JE, Clements P, Medsger TA, Martin RW, Merkel PA, et al. Outcome measurements in scleroderma: Results from a delphi exercise. J Rheumatol. 2007 Mar; 34(3):501–509. [PubMed: 17299843]
27. Dernis E, Lavie F, Pavy S, Wendling D, Flipo RM, Saraux A, et al. Clinical and laboratory follow-up for treating and monitoring patients with ankylosing spondylitis: Development of recommendations for clinical practice based on published evidence and expert opinion. Joint Bone Spine. 2007 Jul; 74(4):330–337. [PubMed: 17590366]
28. Wallace CA, Ruperto N, Giannini E. Childhood Arthritis and Rheumatology Research Alliance, Pediatric Rheumatology International Trials Organization, Pediatric Rheumatology Collaborative Study Group. Preliminary criteria for clinical remission for select categories of juvenile idiopathic arthritis. J Rheumatol. 2004 Nov; 31(11):2290–2294. [PubMed: 15517647]
29. Rider LG, Giannini EH, Brunner HI, Ruperto N, James-Newton L, Reed AM, et al. International consensus on preliminary definitions of improvement in adult and juvenile myositis. Arthritis Rheum. 2004 Jul; 50(7):2281–2290. [PubMed: 15248228]
30. Oddis CV, Rider LG, Reed AM, Ruperto N, Brunner HI, Koneru B, et al. International consensus guidelines for trials of therapies in the idiopathic inflammatory myopathies. Arthritis Rheum. 2005 Sep; 52(9):2607–2615. [PubMed: 16142757]
31. Ruperto N, Meiorin S, Iusan SM, Ravelli A, Pistorio A, Martini A. Consensus procedures and their role in pediatric rheumatology. Curr Rheumatol Rep. 2008 Apr; 10(2):142–146. [PubMed: 18460270]
32. Giannini EH, Ruperto N, Ravelli A, Lovell DJ, Felson DT, Martini A. Preliminary definition of improvement in juvenile arthritis. Arthritis Rheum. 1997 Jul; 40(7):1202–1209. [PubMed: 9214419]
33. Van de Ven AH, Delbecq AL. The nominal group as a research instrument for exploratory health studies. Am J Public Health. 1972 Mar; 62(3):337–342. [PubMed: 5011164]
34. Ruperto N, Ravelli A, Cuttica R, Espada G, Ozen S, Porras O, et al. The pediatric rheumatology international trials organization criteria for the evaluation of response to therapy in juvenile systemic lupus erythematosus: Prospective validation of the disease activity core set. Arthritis Rheum. 2005 Sep; 52(9):2854–2864. [PubMed: 16142708]
35. Ruperto N, Ravelli A, Oliveira S, Alessio M, Mihaylova D, Pasic S, et al. The pediatric rheumatology international trials Organization/American college of rheumatology provisional criteria for the evaluation of response to therapy in juvenile systemic lupus erythematosus: Prospective validation of the definition of improvement. Arthritis Rheum. 2006 Jun 15; 55(3):355–363. [PubMed: 16739203]
36. Wallace CA, Ravelli A, Huang B, Giannini EH. Preliminary validation of clinical remission criteria using the OMERACT filter for select categories of juvenile idiopathic arthritis. J Rheumatol. 2006 Apr; 33(4):789–795. [PubMed: 16482643]

37. Wallace CA, Huang B, Bandeira M, Ravelli A, Giannini EH. Patterns of clinical remission in select categories of juvenile idiopathic arthritis. Arthritis Rheum. 2005 Nov; 52(11):3554–3562. [PubMed: 16255044]
38. Rider LG, Giannini EH, Harris-Love M, Joe G, Isenberg D, Pilkington C, et al. Defining clinical improvement in adult and juvenile myositis. J Rheumatol. 2003 Mar; 30(3):603–617. [PubMed: 12610824]
39. Ruperto N, Ravelli A, Murray KJ, Lovell DJ, Andersson-Gare B, Feldman BM, et al. Preliminary core sets of measures for disease activity and damage assessment in juvenile systemic lupus erythematosus and juvenile dermatomyositis. Rheumatology (Oxford). 2003 Dec; 42(12):1452–1459. [PubMed: 12832713]
40. Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: Development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. OMERACT committee. J Rheumatol. 1993 Mar; 20(3):561–565. [PubMed: 8478874]
41. Tugwell P, Boers M. Developing consensus on preliminary core efficacy endpoints for rheumatoid arthritis clinical trials. OMERACT committee. J Rheumatol. 1993 Mar; 20(3):555–556. [PubMed: 8478872]
42. Isenberg DA, Rahman A, Allen E, Farewell V, Akil M, Bruce IN, et al. BILAG 2004. development and initial validation of an updated version of the british isles lupus assessment group's disease activity index for patients with systemic lupus erythematosus. Rheumatology (Oxford). 2005 Jul; 44(7):902–906. [PubMed: 15814577]
43. Gallagher M, Hares T, Spencer J, Bradshaw C, Webb I. The nominal group technique: A research tool for general practice? Fam Pract. 1993 Mar; 10(1):76–81. [PubMed: 8477899]
44. Fitch, K. The RandUCLA appropriateness method user's manual. Santa Monica: Rand; 2001.
45. Chassin MR, Kosecoff J, Solomon DH, Brook RH. How coronary angiography is used. clinical determinants of appropriateness. JAMA. 1987 Nov 13; 258(18):2543–2547. [PubMed: 3312657]
46. Kravitz RL, Park RE, Kahan JP. Measuring the clinical consistency of panelists' appropriateness ratings: The case of coronary artery bypass surgery. Health Policy. 1997 Nov; 42(2):135–143. [PubMed: 10175621]
47. Shekelle PG, Park RE, Kahan JP, Leape LL, Kamberg CJ, Bernstein SJ. Sensitivity and specificity of the RAND/UCLA appropriateness method to identify the overuse and underuse of coronary revascularization and hysterectomy. J Clin Epidemiol. 2001 Oct; 54(10):1004–1010. [PubMed: 11576811]
48. Lopez-Olivo MA, Suarez-Almazor ME. Developing guidelines in musculoskeletal disorders. Clin Exp Rheumatol. 2007 Nov–Dec; 25(6 Suppl 47):28–36. [PubMed: 18021504]
49. Fraser GM, Pilpel D, Kosecoff J, Brook RH. Effect of panel composition on appropriateness ratings. Int J Qual Health Care. 1994 Sep; 6(3):251–255. [PubMed: 7795961]
50. Shekelle PG, Schriger DL. Evaluating the use of the appropriateness method in the agency for health care policy and research clinical practice guideline development process. Health Serv Res. 1996 Oct; 31(4):453–468. [PubMed: 8885858]
51. MacLean CH. Quality indicators for the management of osteoarthritis in vulnerable elders. Ann Intern Med. 2001 Oct 16; 135(8 Pt 2):711–721. [PubMed: 11601954]
52. Mikuls TR, MacLean CH, Olivieri J, Patino F, Allison JJ, Farrar JT, et al. Quality of care indicators for gout management. Arthritis Rheum. 2004 Mar; 50(3):937–943. [PubMed: 15022337]
53. Silverman SL, Cummings SR, Watts NB. Consensus Panel of the ASBMR, ISCD, and NOF. Recommendations for the clinical evaluation of agents for treatment of osteoporosis: Consensus of an expert panel representing the american society for bone and mineral research (ASBMR), the international society for clinical densitometry (ISCD), and the national osteoporosis foundation (NOF). J Bone Miner Res. 2008 Jan; 23(1):159–165. [PubMed: 17892379]
54. Yazdany J, Panopalis P, Gillis JZ, Schmajuk G, MacLean CH, Wofsy D, et al. A quality indicator set for systemic lupus erythematosus. Arthritis Rheum. 2009 Mar 15; 61(3):370–377. [PubMed: 19248127]

55. Furst DE, Halbert RJ, Bingham CO 3rd, Fukudome S, Anderson L, Bonafede P, et al. Evaluating the adequacy of disease control in patients with rheumatoid arthritis: A RAND appropriateness panel. Rheumatology (Oxford). 2008 Feb; 47(2):194–199. [PubMed: 18178593]
56. Perry S, Kalberer JT Jr. The NIH consensus-development program and the assessment of health-care technologies: The first two years. N Engl J Med. 1980 Jul 17; 303(3):169–172. [PubMed: 6104294]
57. Ferguson JH. The NIH consensus development program. the evolution of guidelines. Int J Technol Assess Health Care. 1996 Summer;12(3):460–474. [PubMed: 8840666]
58. Lomas J. Words without action? the production, dissemination, and impact of consensus recommendations. Annu Rev Public Health. 1991; 12:41–65. [PubMed: 2049143]
59. NIH consensus statement on total knee replacement. NIH Consens State Sci Statements. 2003 Dec 8–10; 20(1):1–34. [PubMed: 17308549]
60. Osteoporosis prevention, diagnosis, and therapy. NIH Consens Statement. 2000 Mar 27–29; 17(1): 1–45.
61. Total hip replacement. NIH Consens Statement. 1994 Sep 12–14; 12(5):1–31.
62. Intravenous immunoglobulin. Consens Statement. 1990 May 21–23; 8(5):1–23.
63. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008 Apr 26; 336(7650):924–926. [PubMed: 18436948]
64. Guyatt GH, Cook DJ, Jaeschke R, Pauker SG, Schunemann HJ. Grades of recommendation for antithrombotic agents: American college of chest physicians evidence-based clinical practice guidelines (8th edition). Chest. 2008 Jun; 133(6 Suppl):123S–131S. [PubMed: 18574262]
65. Jaeschke R, Guyatt GH, Dellinger P, Schunemann H, Levy MM, Kunz R, et al. Use of GRADE grid to reach decisions on clinical practice guidelines when consensus is elusive. BMJ. 2008 Jul 31.337:a744. [PubMed: 18669566]
66. Becker RC, Meade TW, Berger PB, Ezekowitz M, O'Connor CM, Vorchheimer DA, et al. The primary and secondary prevention of coronary artery disease: American college of chest physicians evidence-based clinical practice guidelines (8th edition). Chest. 2008 Jun; 133(6 Suppl):776S–814S. [PubMed: 18574278]
67. Shekelle PG, Kravitz RL, Beart J, Marger M, Wang M, Lee M. Are nonspecific practice guidelines potentially harmful? A randomized comparison of the effect of nonspecific versus specific guidelines on physician decision making. Health Serv Res. 2000 Mar; 34(7):1429–1448. [PubMed: 10737446]

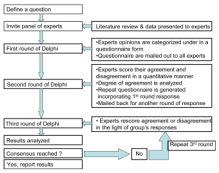

Figure 1.
Schematic diagram of Delphi process step by step.

Table 1

Characteristics of informal and formal consensus development methods (20).

Consensus development method	Mailed* questionnaires/ Survey	Private decisions elicited prior to group discussion?	Formal feedback of group choice	Face-to-face contact	Interaction structured	Incorporates evidence	Aggregation method
Delphi	Yes	Yes	Yes	No	Yes	+	Explicit
NGT	No	Yes	Yes	Yes	Yes	+	Explicit
RAND	Yes	Yes	Yes	Yes	Yes	+++	Explicit
Consensus dev. Conf.	No	No	No	Yes	No	+	Implicit

* postal survey, internet, fax