Published in final edited form as:
Arthritis Rheumatol. 2019 January ; 71(1): 2–4. doi:10.1002/art.40725.

The Evolving Art and Science of ACR Guidelines

Jinoos Yazdany, M.D., M.P.H.[1], **Liron Caplan, M.D., Ph.D.**[2], **John Fitzgerald, M.D., Ph.D**[3], **Gabriela Schmajuk, M.D., M.Sc.**[4]

[1] Division of Rheumatology, Zuckerberg San Francisco General Hospital, University of California, San Francisco

[2] Section of Rheumatology, Rocky Mountain Region Veterans Affairs Medical Center; Chair, ACR Practice Guidelines Subcommittee

[3] Division of Rheumatology, David Geffen School of Medicine, University of California, Los Angeles

[4] Division of Rheumatology, Department of Veterans Affairs, University of California, San Francisco

Frequently, guidelines generate controversy. Earlier this year, in a research letter to *JAMA Internal Medicine*, authors critiqued ACR guidelines for not being adequately evidence-based (1). When such criticisms arise, there is an important opportunity for reflection. Are ACR guidelines appropriately rigorous? Are there ways to improve upon current processes for guideline development? How can we improve the value and integrity of ACR guidelines moving forward?

The ACR guideline development process has evolved significantly over the last two decades. The early ACR guideline statements in the 1990s on topics such as lupus or osteoarthritis were primarily consensus statements generated by a group of experts to address key clinical management issues. Formal systematic literature reviews were generally not performed, there was no validated rating system for evaluating evidence, patients were not involved, and there was an overall lack of transparency about group processes, including conflict of interest policies. Still, these statements had advantages, including that they were written by respected leaders in the field and were simple to interpret.

In the 2000's, with calls for a more rigorous process and a growing body of evidence in rheumatology, this "eminence-based" process evolved into one that used RAND-UCLA appropriateness methodology, combined with components of the American College of Cardiology/American Heart Association (ACC/AHA) evidence rating process (2). The RAND-UCLA method relied on the development of clinical scenarios and was applied to convert systematic literature review data into clinically relevant recommendations using expert opinion. Although the ACR completed a number of guidelines using this method, there were shortcomings. Perhaps most importantly, the manner by which expert panels used the evidence ratings to derive their recommendations continued to be ambiguous. The

Correspondence: Jinoos Yazdany M.D., M.P.H., Division of Rheumatology, University of California, San Francisco, 1001 Potrero Avenue, Building 30, San Francisco, California 94110, jinoos.yazdany@ucsf.edu, Phone: (415) 206-8618.

method also strongly weighed evidence derived from randomized controlled trials (RCTs) over observational data without substantial consideration of a study's rigor. Systematic reviews were sometimes conducted, but were not routinely published, limiting critical re-appraisal of the methods. In addition, patients were minimally involved in the process, raising concerns that their interests were not adequately prioritized.

In 2012, after an evaluation of the 2011 National Academies standards for guideline development (3), the ACR overhauled its guideline development processes and began to incorporate the use of GRADE methodology (www.gradeworkinggroup.org) and also significantly increased patient involvement. All recent ACR guidelines, including the psoriatic arthritis guideline published in this issue of the journal, now use this methodology.

GRADE methodology offered advantages over previous approaches. The method consists of two components: 1) a systematic literature review to assess the certainty of effect estimates for each intervention considered, and 2) a recommendation strength that takes into account not only the quality of the available evidence, but also variability in patient values and preferences. GRADE standardizes the process of moving from evidence to recommendations and enables greater transparency about the judgments made during that process. A recommendation may be strong in the face of low quality evidence in situations where convincing observational evidence exists and/or potential benefits greatly outweigh risks (or vice versa). Similarly, GRADE allows for a conditional recommendation even in the face of high quality evidence; for example, an expensive new therapy supported by clinical trials but producing only a minimal added benefit might warrant a conditional recommendation. Importantly, the scientific evidence and the judgments of the voting panels are published to allow readers to review the specific factors underlying each recommendation.

GRADE ratings are reproducible when people with extensive experience use the method, although interrater reliability diminishes when evidence is more complex (4, 5). The overall benefits of GRADE, perhaps augmented by the extensive resources facilitating its implementation by the GRADE working group, led to rapid adoption of the method among guideline developers such as the ACR; it is now used by over 80 groups (6).

But GRADE also has limitations. One limitation is that GRADE methods are not always easily adaptable to the clinical questions posed by guideline developers. GRADE is simplest to apply when high-quality evidence allows for generation of quantitative estimates of effect for each outcome. If data are qualitative or if there is significant heterogeneity between studies precluding pooling of estimates, user judgment is required to assign levels of certainty and to develop recommendations. Guideline development using GRADE is also significantly more expensive and results in a product that is more complex to read given the detailed analyses required. Finally, GRADE assigns observational studies a default rating of "low quality," and users can then modify these evidence ratings per GRADE guidance. However, some have argued that the general approach to rating of observational data is inadequate, and that GRADE should further specify the manner by which confidence assessments should be modified (7).

This latter point is relevant in understanding why ACR guidelines include many recommendations that rely on low quality evidence. Essentially, recommendations that do not harvest evidence from RCTs are less likely to be designated as having a high or moderate evidence rating. This poses challenges in rheumatology where disease heterogeneity and low prevalence make RCTs challenging for many aspects of care. In addition, rheumatologists understandably seek the most guidance in areas with the lowest quality evidence. Increasing calls for evidence-based medicine through reports like the National Academies *Clinical Practice Guidelines We Can Trust* have led some groups to restrict their recommendations to only those with high-level evidence (3). However, recent criticism of American College of Physicians guidelines on topics such as gout and osteoporosis, which limited recommendations to areas with high quality evidence despite significant observational evidence and expert consensus in other areas, suggest that a more nuanced approach is needed (8, 9). Ultimately, the clinical utility of a guideline may be more in the ability to synthesize the *best available* literature with expert consensus relevant to challenging, real-world clinical situations for which little evidence exists. Limiting the scope of guidelines to only the few areas and populations where RCTs are available ignores the needs of clinicians who seek more comprehensive advice from experts in the field.

In their critique of ACR guidelines, Duarte-Garcia et al. found that more than 50% of the ACR's 403 recommendations were classified as level C, while only one-fourth were level A. In addition, they noted cases of discordance between the evidence level and strength of recommendation, with the rheumatoid arthritis guidelines having the greatest number of strong recommendations based on weak (level C) evidence (16%) (1). This brings up a key question: do clinicians want recommendations, regardless of the quality of the evidence? This question was recently studied in a multi-center RCT involving almost 500 clinicians (10). Investigators randomized participants to receive summaries in areas with either low or high quality evidence, with or without recommendations that followed GRADE. Not surprisingly, in all scenarios, the vast majority of clinicians (>80%) preferred having recommendations in addition to just evidence summaries. The fact that ACR guidelines provide recommendations, and sometimes strong recommendations, even in the face of weak evidence agrees with this study's findings of what clinicians hope for in guidelines and with what the ACR's membership has requested. An additional question is whether ACR is an outlier with regard to the evidence base of its guidelines, with more recommendations based on low quality evidence compared to other specialties. In a recent systematic query of UpToDate, the most widely used reference by clinicians worldwide, 49.7% of 9,451 GRADE recommendations had low evidence/certainty, 39.8% had moderate certainty and 10.5% had high certainty. Uncertainty or low-quality evidence is therefore prevalent, and in fact, the means for rheumatology are similar to those across all fields of medicine (11).

The ACR has remained on the forefront of evolving guideline methodology in order to maintain credibility with guideline users and disseminators. However, no methodology is perfect, and GRADE has its limitations. The ACR should continue to work with GRADE developers to address shortcomings and to contribute to the science around guideline development in general. Moreover, the ACR should continue to begin each guideline project by asking which questions are most clinically relevant and posting that list for public comment, encouraging input from patients and providers alike. Including all relevant questions, regardless of the level of evidence for the eventual recommendations, makes the final guideline more useful to all stakeholders, including patients.

While it may be argued that the ACR's current approach results in many recommendations that are not supported by high quality evidence using current GRADE definitions, the counterargument is that these same recommendations are precisely the ones that rheumatologists and patients find the most useful. This approach is justified as long as there is: 1) transparency about the quality of the evidence in these areas, 2) clarity regarding other factors that played into the guideline development group's decision making (including clinical experience and expertise, observational studies, as well as patient values and preferences), and 3) an acknowledgment that clinicians should have leeway to deviate from recommendations in areas of uncertainty. Finally, it is important to recognize that there are gaps in evidence for much of what we do as rheumatologists, and for many areas with low quality evidence, RCTs may not be feasible. Accelerating the collection of high quality observational data that will facilitate closure of knowledge gaps and improve the evidence base for guidelines through efforts such as the ACR's RISE registry is therefore an important priority.

Acknowledgements:

The authors would like to acknowledge the assistance of Amy Turner at the American College of Rheumatology in reviewing the manuscript.

References

1. Duarte-Garcia A, Zamore R, Wong JB. The Evidence Basis for the American College of Rheumatology Practice Guidelines. JAMA Intern Med. 2018;178(1):146–8. [PubMed: 29181496]
2. Fitch K, Bernstein SJ, Aguilar MD, et al.; The RAND/UCLA Appropriateness Method User's Manual. Santa Barbara, CA: RAND Corporation, 2001.
3. Graham R, Mancher M, Wolman DM, et al.; Committee on Standards for Developing Trustworthy Clinical Practice Guidelines; Board on Health Care Services Clinical Practice Guidelines We Can Trust. Washington, DC: National Academies Press; 2011.
4. Mustafa RA, Santesso N, Brozek J, Akl EA, Walter SD, Norman G, et al. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. J Clin Epidemiol. 2013;66(7):736–42; quiz 42 e1–5. [PubMed: 23623694]
5. Kumar A, Miladinovic B, Guyatt GH, Schunemann HJ, Djulbegovic B. GRADE guidelines system is reproducible when instructions are clearly operationalized even among the guidelines panel members with limited experience with GRADE. J Clin Epidemiol. 2016;75:115–8. [PubMed: 26845745]
6. Norris SL, Bero L. GRADE Methods for Guideline Development: Time to Evolve? Ann Intern Med. 2016;165(11):810–1. [PubMed: 27654340]
7. Malmivaara A. Methodological considerations of the GRADE method. Ann Med. 2015;47(1):1–5. [PubMed: 25356772]
8. Neogi T, Mikuls TR. To Treat or Not to Treat (to Target) in Gout. Ann Intern Med. 2017;166(1):71–2. [PubMed: 27802507]
9. Caplan L, Hansen KE, Saag KG. Response to the American College of Physicians Osteoporosis Guideline, 2017 Update. Arthritis Rheumatol. 2017;69(11):2097–101. [PubMed: 28881479]
10. Neumann I, Alonso-Coello P, Vandvik PO, Agoritsas T, Mas G, Akl EA, et al. Do clinicians want recommendations? A multi-center study comparing evidence summaries with and without GRADE recommendations. J Clin Epidemiol 2018.

11. Agoritsas T, Merglen A, Heen AF, Kristiansen A, Neumann I, Brito JP, et al. UpToDate adherence to GRADE criteria for strong recommendations: an analytical survey. BMJ Open. 2017;7(11):e018593.