arthritis research&therapy

COMMENTARY

The ACR20 and defining a threshold for response in rheumatic diseases: too much of a good thing

David T Felson[1*] and Michael P LaValley[2]

Abstract

In the past 20 years great progress has been made in the development of multidimensional outcome measures (such as the Disease Activity Score and ACR20) to evaluate treatments in rheumatoid arthritis, a process disseminated throughout rheumatic diseases. These outcome measures have standardized the assessment of outcomes in trials, making it possible to evaluate and compare the efficacy of treatments. The methodologic advances have included the selection of pre-existing outcome measures that detected change in a sensitive fashion (in rheumatoid arthritis, this was the Core Set Measures). These measures were then combined into a single multidimensional outcome measure and such outcome measures have been widely adopted in trials and endorsed by the American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR) and regulatory agencies. The secular improvement in treatment for patients with rheumatoid arthritis has been facilitated in part by these major methodologic advancements. The one element of this effort that has not optimized measurement of outcomes nor made it easier to detect the effect of treatments is the dichotomization of continuous measures of response, creating responders and non-responder definitions (for example, ACR20 responders; EULAR good responders). Dichotomizing response sacrifices statistical power and eliminates variability in response. Future methodologic work will need to focus on improving multidimensional outcome measurement without arbitrarily characterizing some patients as responders while labeling others as non-responders.

Prior to 1990 in rheumatology and especially in rheumatoid arthritis (RA), trials tested the efficacy of treatments using outcome measures that varied from trial to trial. One trial might assess 12 outcomes related to symptoms and signs of disease (for example, joint counts, pain, erythrocyte sedimentation rate, morning stiffness), while another might include as many as 15, yet these outcomes might be different from the ones measured in the first trial. Because so many different outcomes were assessed with no primary outcome, the meaning of trial results when one or two of the outcomes showed efficacy for a treatment was unclear. Further, it was not possible to compare the efficacy of treatments across trials because each trial generally used its own set of outcome measures. In trial reports authors could report evidence that a treatment's efficacy was superior to placebo if 1 of 12 outcome measures showed a significant effect of treatment, whereas in another trial report in the same journal, authors could suggest that the same treatment was not efficacious if 2 or 3 of the outcomes showed significant efficacy over placebo. The lack of standardization across trials and the use of multiple comparisons made it impossible to identify which drugs were actually efficacious and how they compared with one another. In addition, many of the outcome measures used in these trials were not sensitive to change and would not have shown efficacy even if the treatment worked terrifically well. Further, the same outcome measures were not always assessed using the same techniques, so that the sensitivity to change of one of the measures might be different in one trial versus another.

With that background, an international group of rheumatologists meeting under the auspices of the American College of Rheumatology (ACR) collected data from randomized trials of second line drugs in RA and carried out a series of analyses that examined, among trials of known effective drugs, which of the outcome measures being used were likely to show efficacy [1]. Among the commonly used outcome measures that were unlikely to

* Correspondence: dfelson@bu.edu
[1]Clinical Epidemiology Research and Training Unit, Boston University School of Medicine, 650 Albany Street, Boston, MA 02118, USA
Full list of author information is available at the end of the article

Box 1. About David T Felson and Michael P LaValley.

David T Felson MD MPH is Professor of Medicine and Epidemiology at Boston University Schools of Medicine and Public Health. He chaired the ACR committee that defined a core set of outcome measures for use in RA trials and that developed the preliminary definition of improvement in RA (also called the ACR20). More recently, he co-chaired an ACR/EULAR effort to define remission in rheumatoid arthritis. Professor Felson also has an active research program in osteoarthritis. He has received the Henry Kunkel Young Investigator Award and the Clinical Research Award from the ACR and the Howley Prize from the Arthritis Foundation for his research.

Dr LaValley has a PhD in Statistics from the Pennsylvania State University and completed a post-doctoral fellowship in Biostatistics at the Harvard University School of Public Health. In 1995 he was hired as a biostatistician for the Boston University Arthritis Center and as an Assistant Professor of Biostatistics at the Boston University School of Public Health. In 2008 he became Professor of Biostatistics, and since 2010 has served as the Research Director of the Center for Enhancing Activity and Participation among Persons with Arthritis (ENACT) at Boston University. His main areas of interest are meta-analysis, clinical trial methods, longitudinal data analysis, logistic regression and survival analysis.

show that effective treatments actually worked were proximal interphalangeal circumference, walk time, functional class (graded 1 through 4), hemoglobin, grip strength and morning stiffness. Morning stiffness was not sensitive to change because it was absent in many patients with RA, making it impossible for them to experience an improvement when treated with an effective drug [1]. Among the outcome measures that were found to be most sensitive to change were the patient global assessment, tender joint count and, in trials of second line drugs, swollen joint count and erythrocyte sedimentation rate.

Taking into account the sensitivity to change, the desire to eliminate redundant measures (for example, tender joint count and tender joint score) and attempting to select outcome measures that represented the breadth of RA manifestations, the ACR Committee chose a core set of variables to be included in all trials (Table 1), a recommendation that was later endorsed by the International League Against Rheumatism and the World Health Organization [2].

With this list of seven measures, the committee had standardized RA outcome assessment and decreased the number of outcome measures. However, trials still assessed seven measures, often with all as primary outcomes, and there needed to be a single measure that reflected the breadth of RA activity, including both physician-measured assessments and patient-reported outcomes. With this in mind, an international committee again assembled and tested a variety of possible definitions of improvement. Using different thresholds and combinations of core set measures, the committee chose a definition that showed the greatest sensitivity to change. Other factors considered by the committee included ease of use, and accord with rheumatologists' impressions of improvement. The ACR definition of improvement [3] (often called the ACR20 because it requires at least a 20% improvement in the core set measures for a patient to reach improvement) was promulgated and has been widely adopted in RA trials. A little later, the European League Against Rheumatism (EULAR) also developed their own definition of response [4], which broke improvement into three categories and, unlike the ACR definition, required both a low level of disease and a certain degree of improvement for a patient to be characterized as having good improvement. Subsequent work has suggested that the ACR20 and the EULAR definition of improvement perform comparably [5], and many trials have included both, choosing one of the measures as a primary outcome and reporting the other as a secondary outcome. Importantly, the US Food and Drug Administration also recommended the ACR20 as a preferred outcome measure for testing the efficacy of new drugs for RA with respect to signs and symptoms of disease. Since most trials in RA are carried out by industry, this endorsement by the Food and Drug Administration was a critical element to the widespread dissemination and use of the ACR20. Even now [6], the ACR20 is probably the most widely used outcome measure in RA trials.

Table 1 American College of Rheumatology disease activity measures for rheumatoid arthritis clinical trials: Core Set

Disease activity measure	
1	Tender joint count
2	Swollen joint count
3	Patient's assessment of pain
4	Patient's global assessment of disease activity
5	Physician's assessment of physical function
6	Patient's assessment of physical function
7	Acute-phase reactant value
For trial duration ≥1 year and agent being tested as a 'DMARD', also perform:	
8	Radiography or other imaging technique

DMARD disease-modifying anti-rheumatic drug.

With the success and widespread use of the ACR20 came the desire among rheumatologists studying other rheumatic diseases to have similar standardized definitions of response and improvement. In the few years after the ACR20 was published, similar efforts were undertaken for juvenile RA, osteoarthritis, low back pain, psoriatic arthritis, and spondyloarthropathies; more recently, efforts for myositis and vasculitis have paralleled earlier efforts with a focus on developing a uniform set of measures for trial outcomes and sometimes defining a threshold for improvement.

It is not surprising that the promulgation of a rationally selected core set of outcome measures and its consolidation into one multidimensional measure of response has occurred contemporaneously with the improvement of treatments in rheumatic disease. Making uniform and efficient the measurement of response in rheumatic disease has facilitated the comparison of new and conventional treatments. For example, the ACR20 and variations on this measurement tool have been used to assert that anti-tumor necrosis factor inhibitors perform as well or better than conventional treatments in RA [7], an argument that would have been difficult to make with the old chaotic scheme of multiple measurements. Also, meta-analyses have convincingly demonstrated that some new therapies for RA did not work as well as either conventional or new biologic agents [8-10]. These treatments shown to be less efficacious have then lost favor in the marketplace.

Table 2 shows an enumeration of the benefits of defining response from a methods perspective. The

Table 2 Beneficial and detrimental effects of Core Set and ACR20 on trials in rheumatoid arthritis

Beneficial effects	Detrimental effects
Selected outcome measures most likely to change with treatment	Dichotomized a continuous measures of response
Made uniform trial outcome measures across studies, making comparisons possible	
Decreased the number of outcomes from >10 to 7 and then to 1 (ACR20)	

elimination of outcome measures from trials that were insensitive to change improved the likelihood that effective treatments would be found. The widespread agreement to adopt uniformity with respect to trial outcomes made comparisons of treatments possible and even allowed for examination of the consistency of efficacy across trials of the same treatment. The development of core set measures and the ultimate definition of response decreased the multiple comparison problems in RA trials and other rheumatic disease trials. The consolidation of multiple selected outcome measures into one composite measure also served to improve statistical power, providing a single measure that represented multiple elements of disease activity (for example, the RA core set has elements of patient measures, physician measures and blood tests). Analyses of trial data showed that, since the core set measures were all correlated with one another, it was rare for patients to experience extremely discordant outcomes across the measures - generally, if a patient improved, most or all of the measures improved, although often not to the same extent. The core set has served the scientific community well but it is likely that many of the measures for RA and for other diseases will be refined with the development of new patient-reported outcomes such as those produced by the National Institutes of Health's Patient Reported Outcomes Measurement Information System initiative (for example).

Unfortunately, one effect of this process has not been beneficial (Table 2). In developing a definition of response, the ACR Committee and other rheumatic disease study groups have used thresholds to define response. Often clinically based, these thresholds initially seemed like a wonderful way of communicating the effect of a new treatment, that a certain number of patients would experience improvement when treated. The problem is that taking a continuous measure and arbitrarily cutting it so as to create a dichotomous response/non-response measure, called 'responder analyses', sacrifices statistical power and inflates the number of patients needed to evaluate the efficacy of treatments. Due mostly to their loss of power, responder analyses are discouraged in the clinical trials literature [11], and in a recent position paper, the Pharmaceutical Research and Manufacturers of America (PhRMA) have advised against use of these analyses [12]. The loss of power in these analyses has been repeatedly shown in simulation studies [13] and has been the subject of prominent editorials in clinical journals [14]. As noted by Altman and Royston [14], responder analyses lead to several problems. First, statistical power is reduced; they estimate that it is equivalent to discarding one-third of the data collected. This is especially inadvisable when only small numbers of patients can be recruited, an especially acute problem in some rare rheumatic diseases like myositis, vasculitis and scleroderma. Generally speaking, the use of a dichotomized response/non-response measure should be discouraged in studies of these diseases and probably in other rheumatic disease trials too. Altman and Royston and the PhRMA position paper also note other problems introduced by responder analysis, including an underestimation of the degree of variation between groups with variation subsumed within each response group and yet made invisible when the response is dichotomized. Individuals close to each other, but on opposite sides of the response cut point, are characterized as being very different rather than similar.

With the enlarging armamentarium of effective treatments in RA, the need to compare the efficacy of treatments will intensify. Small differences would be expected and use of a dichotomous measure of response would demand very large sample sizes to compare treatments. This goal could be accomplished more efficiently with a continuous outcome measure. Further, if only small numbers of patients are needed to test a treatment in a subgroup of persons with RA (or among those with other rheumatic disorders), a continuous outcome measure will facilitate the testing of treatment without demanding impractically large sample sizes. Given these anticipated needs, an ACR committee once again assembled and created a new outcome measure based on the ACR20 called the ACRHybrid. With the ACRHybrid a patient's response is based mostly on their average percentage improvement in the core set measures with the caveat that average improvement is adjusted based on whether it satisfies the ACR20, 50 or 70. While endorsed by the ACR [15], the ACRHybrid has yet to be used as a primary outcome measure in any large-scale RA trial. This measure or another continuous measure would permit the definitive evaluation of the comparative efficacy of RA treatments and would facilitate evaluation of how regimens compare in terms of efficacy. The continued use of dichotomous measures to evaluate these issues has made the evaluation of therapeutic uncertainties more challenging at a time when it is increasingly necessary to determine which of our new agents is more efficacious.

While dichotomous measures sacrifice statistical power and can hide valuable information about treatment response, this does not mean that clinical investigators should avoid defining important dichotomous outcomes like the minimally important clinical improvement or disease activity low enough to be acceptable to patients. It just means, especially for trials of treatments of uncommon rheumatic diseases, comparative RA trials and other similar situations, that these dichotomous measures should not be used as primary outcomes. Recommendations on how to define these dichotomous outcomes can be found elsewhere [16].

Beyond RA, the continued development and use of dichotomous measures of response in rheumatic diseases may be sacrificing our ability to detect whether treatments are efficacious. While core set measures need to be developed for rheumatic disease trials and these ought to follow the process used for RA, the final step of that process should be to identify a single multidimensional outcome on a continuous scale.

Conclusion

The past 20 years have witnessed huge advances not just in the armamentarium of treatments available for RA but in the use of valid and responsive measurement tools to assess their effectiveness. Selecting outcome measures sensitive to change, consolidating these into single measures and adopting standardization of measurement across trials has facilitated the assessment of treatments. The dichotomization of treatment response has unfortunately not produced major benefits and should be jettisoned in favor of a primary assessment of treatment efficacy that utilizes continuous response measures.

Note: This article is part of the collection *Research through the eyes of pioneers*. Other articles in this series can be found at http://arthritis-research.com/series/pioneers.

Abbreviations

ACR: American College of Rheumatology; EULAR: European League Against Rheumatism; PhRMA: Pharmaceutical Research and Manufacturers of America; RA: Rheumatoid arthritis.

Competing interests

The authors declare no competing interests.

Acknowledgements

The authors appreciate the technical assistance of Anne Plunkett.

Author details

[1]Clinical Epidemiology Research and Training Unit, Boston University School of Medicine, 650 Albany Street, Boston, MA 02118, USA. [2]Department of Biostatistics, Boston University School of Public Health, Crosstown Building, 801 Massachusetts Avenue, Boston, MA 02118, USA.

Published: 03 Jan 2014

References

1. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, Furst D, Goldsmith C, Kieszak S, Lightfoot R, Paulus H, Tugwell P, Weinblatt M, Widmark R, Williams HJ, Wolfe F: **The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials.** *Arthritis Rheum* 1993, **36:**729–740.
2. Boers M, Tugwell P, Felson DT, van Riel PL, Kirwan JR, Edmonds JP, Smolen JS, Khaltaev N, Muirden KD: **World Health Organization and International League of Associations for Rheumatology core endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials.** *J Rheumatol Suppl* 1994, **41:**86–89.
3. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, Katz LM, Lightfoot R Jr, Paulus H, Strand V, Tugwell P, Weinblatt M, Williams HJ, Wolfe F, Kieszak S: **American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis.** *Arthritis Rheum* 1995, **38:**727–735.
4. van Gestel AM, Prevoo ML, van 't Hof MA, van Rijswijk MH, van de Putte LB, van Riel PL: **Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis. Comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism Criteria.** *Arthritis Rheum* 1996, **39:**34–40.
5. van Gestel AM, Anderson JJ, van Riel PL, Boers M, Haagsma CJ, Rich B, Wells G, Lange ML, Felson DT: **ACR and EULAR improvement criteria have comparable validity in rheumatoid arthritis trials. American College of Rheumatology European League of Associations for Rheumatology.** *J Rheumatol* 1999, **26:**705–711.
6. Weinblatt ME, Schiff M, Valente R, van der HD, Citera G, Zhao C, Maldonado M, Fleischmann R: **Head-to-head comparison of subcutaneous abatacept versus adalimumab for rheumatoid arthritis: findings of a phase IIIb, multinational, prospective, randomized study.** *Arthritis Rheum* 2013, **65:**28–38.
7. Bathon JM, Martin RW, Fleischmann RM, Tesser JR, Schiff MH, Keystone EC, Genovese MC, Wasko MC, Moreland LW, Weaver AL, Markenson J, Finck BK: **A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis.** *N Engl J Med* 2000, **343:**1586–1593.
8. Felson DT, Anderson JJ, Meenan RF: **Use of short-term efficacy/toxicity tradeoffs to select second-line drugs in rheumatoid arthritis. A metaanalysis of published clinical trials.** *Arthritis Rheum* 1992, **35:**1117–1125.
9. Nam JL, Winthrop KL, van Vollenhoven RF, Pavelka K, Valesini G, Hensor EM, Worthy G, Landewe R, Smolen JS, Emery P, Buch MH: **Current evidence for the management of rheumatoid arthritis with biological disease-modifying antirheumatic drugs: a systematic literature review informing the EULAR recommendations for the management of RA.** *Ann Rheum Dis* 2010, **69:**976–986.
10. Launois R, Avouac B, Berenbaum F, Blin O, Bru I, Fautrel B, Joubert JM, Sibilia J, Combe B: **Comparison of certolizumab pegol with other anticytokine agents for treatment of rheumatoid arthritis: a multiple-treatment Bayesian metaanalysis.** *J Rheumatol* 2011, **38:**835–845.
11. Snapinn SM, Jiang Q: **Responder analyses and the assessment of a clinically relevant treatment effect.** *Trials* 2007, **8:**31.
12. Uryniak T, Chan ISF, Fedorov VV, Jiang Q, Oppenheimer L, Snapinn SM, Teng C-H, Zhang J: **Responder analyses - A PhRMA position paper.** *Stat Biopharmaceut Res* 2011, **3:**476–487.
13. Anderson JJ, Bolognese JA, Felson DT: **Comparison of rheumatoid arthritis clinical trial outcome measures: a simulation study.** *Arthritis Rheum* 2003, **48:**3031–3038.
14. Altman DG, Royston P: **The cost of dichotomising continuous variables.** *BMJ* 2006, **332:**1080.
15. van Vollenhoven RF, Felson D, Strand V, Weinblatt ME, Luijtens K, Keystone EC: **American College of Rheumatology hybrid analysis of certolizumab pegol plus methotrexate in patients with active rheumatoid arthritis: data from a 52-week phase III trial.** *Arthritis Care Res (Hoboken)* 2011, **63:**128–134.
16. Kvien TK, Heiberg T, Hagen KB: **Minimal clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): what do these concepts mean?** *Ann Rheum Dis* 2007, **66:**iii40–iii41.

10.1186/ar4428
Cite this article as: Felson and LaValley: **The ACR20 and defining a threshold for response in rheumatic diseases: too much of a good thing.** *Arthritis Research & Therapy* 2014, 16:101