Bringing It All Together: A Novel Approach to the Development of Response Criteria for Chronic Gout Clinical Trials

WILLIAM J. TAYLOR, JASVINDER A. SINGH, KENNETH G. SAAG, NICOLA DALBETH, PATRICIA A. MacDONALD, N. LAWRENCE EDWARDS, LEE S. SIMON, LISA K. STAMP, TUHINA NEOGI, ANGELO L. GAFFO, PUJA P. KHANNA, MICHAEL A. BECKER, and H. RALPH SCHUMACHER Jr

ABSTRACT. ***Objective.*** To review a novel approach for constructing composite response criteria for use in chronic gout clinical trials that implements a method of multicriteria decision-making.

Methods. Preliminary work with paper patient profiles led to a restricted set of core-set domains that were examined using 1000Minds™ by rheumatologists with an interest in gout, and (separately) by OMERACT registrants prior to OMERACT 10. These results and the 1000Minds approach were discussed during OMERACT 10 to help guide next steps in developing composite response criteria.

Results. There were differences in how individual indicators of response were weighted between gout experts and OMERACT registrants. Gout experts placed more weight upon changes in uric acid levels, whereas OMERACT registrants placed more weight upon reducing flares. Discussion highlighted the need for a "pain" domain to be included, for "worsening" to be an additional level within each indicator, for a group process to determine the decision-making within a 1000Minds exercise, and for the value of patient involvement.

Conclusion. Although there was not unanimous support for the 1000Minds approach to inform the construction of composite response criteria, there is sufficient interest to justify ongoing development of this methodology and its application to real clinical trial data. (J Rheumatol 2011;38:1467–70; doi:10.3899/jrheum.110274)

Key Indexing Terms:
GOUT RESPONSE CRITERIA MULTICRITERIA DECISION-MAKING

At OMERACT 10, the focus of the Gout Module was to obtain endorsement of specific instruments that measure each of the 7 core domains identified as required outcomes in chronic gout trials at OMERACT 9[1]. The 3 preceding articles in this series describe results of this process for patient-reported outcomes[2], tophi[3], and serum urate[4]. Flare, which was also discussed at OMERACT 10, forms the basis of an American College of Rheumatology (ACR)-European League Against Rheumatism (EULAR) project that is being reported separately. Although individual indicators of response are very

From the Department of Medicine, University of Otago, Wellington, New Zealand; Birmingham Veterans Affairs (VA) Medical Center and University of Alabama, Birmingham, AL, USA; Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand; Takeda Global Research and Development, Deerfield, IL; Department of Medicine, University of Florida, Gainsville FL; SDG LLC, Cambridge, MA, USA; Department of Medicine, University of Otago, Christchurch, New Zealand; Boston University School of Medicine, Boston, MA; Division of Rheumatology, David Geffen School of Medicine, University of California, Los Angeles, CA; Rheumatology Section, The University of Chicago, Chicago, IL; and University of Pennsylvania and VA Medical Center, Philadelphia, PA, USA.

Partially supported by a University of Otago Matched Funding Summer Scholarship and ACR-EULAR Grant through the University of Pennsylvania; also supported with resources and use of facilities at Birmingham VA Medical Center (Dr. Singh). Dr. Khanna is the recipient of an American College of Rheumatology REF Clinical Investigator Award (Quality of Life and Healthcare Utilization in Chronic Gout) and Co-Investigator on NIH/NIAMS 1AR057936 (Development and Initial Validation of PROMIS GI Distress Scale), Khanna, D; Spiegel B, co-principal investigators. Dr. Neogi is a recipient of NIAMS 5K23AR55127. Dr. Singh has received speaker honoraria from Abbott; research and travel grants from Allergan, Takeda, Savient, Wyeth, and Amgen; and consultant fees from Savient, URL Pharmaceuticals, and Novartis. Dr. Saag has received honoraria from Novartis, Amgen, and Lilly, has conducted clinical trials with Takeda, Novartis, and Regeneron, and serves on the Board of Directors of the Gout and Uric Acid Society. Ms MacDonald is an employee of Takeda Global Research and Development. Dr. Edwards has received consultant fees from Takeda and Savient. Dr. Simon has served on the Board of Directors for Savient Pharmaceuticals, and as a consultant for Takeda. Dr. Becker has been a consultant for Takeda, Savient, BioCryst, URL/Pharma, Ardea, and Regeneron. Dr. Schumacher has received consulting fees from Takeda, Savient, Regeneron, Novartis, and Pfizer.

W.J. Taylor, PhD, FRACP, Associate Professor, Department of Medicine, University of Otago; J.A. Singh, MBBS, MPH, Associate Professor; K.G. Saag, MD, MPH, Professor, Birmingham VA Medical Center and University of Alabama; N. Dalbeth, MD, FRACP, Associate Professor, Faculty of Medical and Health Sciences, University of Auckland; P.A. MacDonald, BSN, NP, Takeda Global Research and Development; N.L. Edwards, MD, Professor, Department of Medicine, University of Florida; L.S. Simon, MD, SDG LLC; L.K. Stamp, PhD, FRACP, Associate Professor, Department of Medicine, University of Otago; T. Neogi, MD, PhD, FRCPC, Assistant Professor, Boston University School of Medicine; A.L. Gaffo, MD, MSPH, Assistant Professor, Birmingham VA Medical Center and University of Alabama; P.P. Khanna, MD, MPH, Instructor, Division of Rheumatology, David Geffen School of Medicine, University of California; M.A. Becker, MD, Professor, Rheumatology Section, University of Chicago; H.R. Schumacher Jr, MD, Professor, University of Pennsylvania and VA Medical Center.

Address correspondence to Prof. Taylor. E-mail: will.taylor@otago.ac.nz

important for clinical trials, there is merit in considering whether and how such individual indicators can be summarized as a single, composite measure of response[5]. Such composite measures have been usefully employed in clinical trials for rheumatoid arthritis[6], psoriatic arthritis[7,8], osteoarthritis[9], and ankylosing spondylitis[10,11].

Work towards response criteria for chronic gout studies is being undertaken under the auspices of an ACR-EULAR supported project (principal investigator HRS). Nevertheless, OMERACT 10 offered a useful opportunity to introduce and discuss novel methodology, and this became the focus at the meeting, rather than for seeking endorsement for any particular response criteria.

Composite response measures are examples of multiple criteria decision-making (MCDM), a class of activities characterized as "procedures by which concerns about multiple conflicting criteria can be formally incorporated into the management planning process"[12]. In the case of response criteria for gout clinical trials, multiple indicators of response are incorporated into a decision as to whether the patient has responded or not to the intervention. Other examples of MCDM include classification criteria[13], governmental decision support for transport projects[14], access to publicly funded elective health services[15], and marketing research (this list is far from exhaustive). There are many methods of MCDM, too many to discuss here. However, the general approach of MCDM offers a novel and fruitful way of considering composite criteria, which has not yet been formally employed in the area of rheumatic diseases in relation to development of response criteria. The particular approach we wish to introduce here is PAPRIKA (Potentially All Pairwise RanKings of all possible Alternatives) as implemented in 1000Minds™ software (URL: http://www.1000minds.com/).

To date, typical methods of constructing composite response criteria in rheumatology have involved testing Boolean combinations of key response indicators against a criterion standard of response, which is that the patient received "effective" therapy, rather than placebo. There is some logical inconsistency in this approach, which is that patients do not respond in the same way to drug therapies, and some agents (while generally effective) do not work for some patients. The criterion standard is therefore not without error. This point was highlighted in a study that directly compared the ASAS-Improvement Criteria (ASA-IC, derived from clinical trials) to an expert consensus-derived classification of response (using Delphi methods) in real ankylosing spondylitis patient profiles. In this study, overall agreement was 62%, with only 50% of cases deemed to be responders by expert consensus to have responded according to the ASA-IC[11].

There is also some potential circularity with using allocation to a particular intervention to define response, when it is the response to the intervention that we are trying to determine. Finally, the identification of whether an intervention is "effective" or not must require some prior criterion to determine efficacy. Why, then, shouldn't the prior criterion be taken as the standard for response?

PAPRIKA

The use of a preference-based approach enables the direct weighting of multiple potential indicators of response, based on the direct views of expert physicians or other stakeholders (for example, patients with the disease). PAPRIKA is a mathematical algorithm for constructing relative weights for each response indicator, based upon the results of a series of pairwise comparisons of undominated pairs of all possible alternatives. Each compares different levels of 2 indicators at a time, in order for the respondent to determine which combination of indicators has "responded" the most[16]. An "undominated pair" is characterized by a higher-ranking category for at least one indicator and a lower-ranking category for the other indicator. Further, pairs that are implicitly ranked as corollaries of explicitly ranked pairs are also identified, which leads to efficiency of the algorithm and the requirement for only a portion of all possible combinations to be evaluated by a decision-maker. Sufficient pairwise comparisons are made until the algorithm identifies a series of weights that is consistent with all the decisions that have been made.

PAPRIKA is implemented in 1000Minds software, which was used in the 2 exercises described below. Two intrinsic limitations to this approach are that relatively few indicators and levels of each indicator can be used without requiring an overwhelming number of pairwise comparisons to be performed; and that the algorithm generates an additive, linear structure for the response criteria without incorporation of interactions between indicators.

1000Minds Survey

A preliminary exercise among gout experts using paper patient profiles and potential response indicators from the core-set of domains for chronic gout studies from OMERACT 9[1] identified a restricted set of indicators that independently contributed to whether the paper profile was judged (by a small panel of gout experts) to have responded or not. Here follows the list of indicators, with a coefficient from discriminant function analysis given in parentheses: Percentage change (%change) in SUA (0.81), %change in HAQ (0.42), %change in number of tophi (0.36), and %change in flare frequency (0.32) (overall model Wilks' lambda 0.113, chi-square 115, degrees of freedom 4, $p < 0.001$). Each indicator was categorized into levels as shown in Table 1 for pairwise comparisons subsequently performed.

Twenty rheumatologists with gout expertise participated in the first 1000Minds exercise, and 22 OMERACT registrants participated in the second exercise. The median value for the weighting of each indicator level was used to construct the responder index. For gout experts, the order of relative importance appeared to be (most to least): serum uric acid, number of tophi, flare frequency, and Health Assessment Question-

Table 1. Responder index derived from the 1000Minds™ exercises.

Indicators	Weights* Derived from Gout Experts	Weights* Derived from OMERACT Registrants
Serum uric acid		
No change or worsening	0	0
Mild decrease: > 25% reduction, but final value > upper limit of normal range	12	6
Moderate decrease: > 25% reduction and final value within normal range	21	13
Marked decrease: > 50% reduction and final value ≤ 0.36 mmol/l	31	17
Flare frequency		
No change or increased	0	0
Mild decrease: 20% to 50% reduction	7	11
Moderate decrease: 50% to 75% reduction	14	21
Marked improvement: > 75% reduction	19	33
Number of tophi		
No change or increase in number	0	0
Mild improvement: 20% to 50% reduction	16	11
Moderate improvement: 50% to 75% reduction	24	17
Marked improvement: > 75% reduction	27	25
HAQ-DI		
No change or increase	0	0
Mild improvement: 20% to 50% reduction	5	9
Moderate improvement: 50% to 75% reduction	9	13
Marked improvement: > 75% reduction	14	20

* Weights were calculated by the PAPRIKA algorithm, essentially by ordering of all pairwise combinations of indicators.

naire Disability Index (HAQ-DI); whereas for OMERACT registrants, the order of relative importance was flare frequency, HAQ-DI, number of tophi, and serum uric acid.

The performance of each of these responder indices was assessed using patient profiles (for which a panel of rheumatologist gout experts had already determined response to therapy). The area under the receiver operating characteristic (ROC) curve for the gout expert-derived responder index was 0.96 (95% confidence interval 0.92, 1.00) compared to 0.91 (95% CI 0.84, 0.96) for the OMERACT registrant-derived responder index.

OMERACT 10 Discussion and Voting

During OMERACT 10, the discussion revolved around methodological and conceptual considerations rather than endorsement of any particular responder index. It was unclear to a significant proportion of participants that the 1000Minds approach had merit, with 20/84 (24%) participants being uncertain whether 1000Minds was a suitable means of determining weights for different response indicators (Table 2). However, of those who had an opinion, 38/64 (59%) thought that the approach had merit. The fact that this is a novel and unfamiliar methodology should be taken into account when considering the results of the participant voting. That this was an approach that could easily be adapted to determine the weights assigned by patients themselves to different indicators of response was seen to be especially useful.

Some limitations of the approach taken to date were highlighted. First, it was suggested that in view of the importance of the domain "pain" as the principal clinical manifestation of gout, the "pain" domain should be incorporated into the response criteria. This was despite "pain" not being selected by the discriminant function model as being independently associated with response in the paper patient profile exercise [49/79 (62%) participants voted for this]. Second, it was not clear how to interpret the difference in weights obtained by the 2 different surveys. These were obtained by median values across individual respondents, but it was considered that a consensus approach by making each of the decisions involved in the 1000Minds using a group process would be better. This is, in fact, recommended by the developers of 1000Minds (Franz Ombler, personal communication). Third, the construction of different levels within each indicator needed some refinement. For example, "worsening" was incorporated into the "no change" level, whereas perhaps it should be an additional level. Finally, the inclusion of the "tophi" domain rendered difficulties for determining response among those patients without tophi. None of these difficulties are insurmountable, and further work will take each issue into consideration.

Conclusion

The 2 exercises using a MCDM approach have shown that highly discriminative composite responder indices can be constructed for chronic gout trials. Nevertheless, additional work is to be undertaken through the ACR-EULAR initiative. The OMERACT 10 discussions helped clarify some of this agenda. Future activities will include further work to determine how to use the new flare definition as an outcome, refinement of the structure of response criteria to include

Table 2. Results of voting for questions relating to response criteria.

Questions	Responses, n (%)
Which additional domains should be included in a composite indicator of response, irrespective of their independent contribution to a statistical model for gout response? (Select one or more)	
Pain	49 (62)
Patient global	41 (52)
Health-related quality of life	24 (30)
None of the above	8 (10)
Do you agree that more patient involvement in determining the components of a responder index is the most important next step?	
Yes	51 (65)
No	17 (22)
Don't know	11 (14)
Do you agree that weights derived from statistical models should be incorporated into composite response criteria?	
Yes	49 (64)
No	7 (9)
Don't know	21 (27)
Do you agree that the 1000Minds™ approach is suitable for determining weights for different indicators of response?	
Yes	38 (45)
No	26 (31)
Don't know	20 (24)

levels for worsening and a domain for "pain," utilizing the 1000Minds approach through group consensus (patients and physicians), rather than averaging individual views, and comparing the composite measure against standard approaches in real clinical trial data.

ACKNOWLEDGMENT

The work of Sena Han and Nicole Lynch in collecting data required for the paper patient profile exercise and 1000Minds surveys; and the assistance of Franz Ombler from 1000Minds are gratefully acknowledged. We also thank all the participants who responded to these surveys. Rheumatologists with expertise in gout who participated in at least one of the preliminary surveys were: Michael Doherty, Maarten Boers, John Sundy, Lan Chen, Peter Chapman, Janitzia Vazquez-Mellado, Herb Baraf, Eswar Krishnan, Peter Gow, Xuejun Zeng, Ted Mikuls, Victoria G. Barskova, Sergio Kowalski, Francisca Sivera, Claudia Goldenstein Schainberg, Naomi Schlesinger, Cesar Diaz-Torne, Eliseo Pascual, Hisashi Yamanaka, Rebecca Grainger, Dan Furst, and Dinesh Khanna.

REFERENCES

1. Schumacher HR Jr, Taylor W, Edwards NL, Grainger R, Schlesinger N, Dalbeth N, et al. Outcome domains for studies of acute and chronic gout. J Rheumatol 2009;36:2342-5.
2. Singh JA, Taylor WJ, Simon LS, Khanna PP, Stamp LK, McQueen FM, et al. Patient reported outcomes in chronic gout: a report from OMERACT 10. J Rheumatol 2011;38:1452-7.
3. Dalbeth ND, McQueen FM, Singh JA, MacDonald PA, Edwards NL, Schumacher HR Jr, et al. Tophus measurement as an outcome measure for clinical trials of chronic gout: progress and research priorities. J Rheumatol 2011;38:1458-61.
4. Stamp LK, Khanna PP, Dalbeth ND, Boers M, Maksymowych WP, Schumacher HR Jr, et al. Serum urate in chronic gout — Will it be the first validated soluble biomarker in rheumatology? J Rheumatol 2011;38:1462-6.
5. Anderson JJ. Mean changes versus dichotomous definitions of improvement. Stat Methods Med Res 2007;16:7-12.
6. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727-35.
7. Gladman DD, Landewe R, McHugh NJ, Fitzgerald O, Thaci D, Coates L, et al. Composite measures in psoriatic arthritis: GRAPPA 2008. J Rheumatol 2010;37:453-61.
8. Gladman DD, Tom BDM, Mease PJ, Farewell VT. Informing response criteria for psoriatic arthritis. I: Discrimination models based on data from 3 anti-tumor necrosis factor randomized studies. J Rheumatol 2010;37:1892-7.
9. Pham T, van der Heijde D, Lassere M, Altman RD, Anderson JJ, Bellamy N, et al. Outcome variables for osteoarthritis clinical trials: The OMERACT-OARSI set of responder criteria. J Rheumatol 2003;30:1648-54.
10. Anderson JJ, Baron G, van der Heijde D, Felson DT, Dougados M. Ankylosing spondylitis assessment group preliminary definition of short-term improvement in ankylosing spondylitis. Arthritis Rheum 2001;44:1876-86.
11. van Tubergen A, van der Heijde D, Anderson J, Landewe R, Dougados M, Braun J, et al. Comparison of statistically derived ASAS improvement criteria for ankylosing spondylitis with clinically relevant improvement according to an expert panel. Ann Rheum Dis 2003;62:215-21.
12. International Society on Multiple Criteria Decision Making. [Internet. Accessed March 18, 2011.] Available from: http://www.mcdmsociety.org/intro.html
13. Neogi T, Aletaha D, Silman AJ, Naden RL, Felson DT, Aggarwal R, et al. The 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for rheumatoid arthritis: phase 2 methodological report. Arthritis Rheum 2010;62:2582–91.
14. Transport Analysis guidance — WebTAG. [Internet. Accessed March 16, 2011.] London: Department for Transport; 2010. Available from: http://www.dft.gov.uk/webtag/
15. Hadorn DC, Holmes AC. The New Zealand priority criteria project. Part 1: Overview. BMJ 1997;314:131-4.
16. Hansen P, Ombler F. A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives. J Multi-Crit Decis Anal 2009;15:87-107.