Special Interest Group

Outcome Evaluations in Gout

H. RALPH SCHUMACHER, WILLIAM TAYLOR, NANCY JOSEPH-RIDGE, FERNANDO PEREZ-RUIZ, LAN X. CHEN, NAOMI SCHLESINGER, DINESH KHANNA, DANIEL E. FURST, MICHAEL A. BECKER, NICOLA DALBETH, and N. LAWRENCE EDWARDS

ABSTRACT. Methods to measure outcomes in gout still require consensus and validation. This Special Interest Group was assembled to identify domains of interest and is now evaluating a series of outcomes for features of acute gouty arthritis and chronic gout. To accomplish this, working groups have been formed and domains identified. Delphi methodology has been used to address gouty flares as an outcome of greatest interest. Studies addressing other outcome measures were reported at the OMERACT 8 meeting and validation has begun on some outcomes. There has been progress on developing a definition of a flare, and validating reproducibility of some chronic gout outcome measures in some domains, such as tophus size and patient perceptions. Use of these outcomes as well as a health-related quality of life measure are being studied in clinical trials. Pain on a Likert scale appears to be a valid outcome in acute gout. Final validation of these outcomes has not yet been achieved. In summary, the unique problems of evaluating outcomes in gout are finally being addressed. While no measures are available for use yet, an agenda has been developed. (J Rheumatol 2007;34:1381–5)

Key Indexing Terms:
GOUT OUTCOMES OMERACT URIC ACID

The Special Interest Group for gout outcomes met in Malta to review progress since 2002 and to plan continued studies. At OMERACT 7, the focus was on identification of core domains that were published in that report[1]. In studies of chronic gout our group identified 9 necessary domains: (1) serum urate, (2) gout flares, (3) tophus size, (4) joint imaging, (5) health-related quality of life (HRQOL), (6) musculoskeletal function, (7) patient global, (8) participation, and (9) safety and tolerability[1]. Discussions of whether cardiovascular outcomes of hyperuricemia should be considered have continued. Committees have made progress in several areas.

From the Division of Rheumatology, University of Pennsylvania School of Medicine, and Veterans Affairs Medical Center, Philadelphia, Pennsylvania, USA.

Supported by TAP Pharmaceuticals.

H.R. Schumacher, MD, Veterans Affairs Medical Center, Philadelphia, Pennsylvania; W. Taylor, MD, Wellington School of Medicine and Health Sciences, Rehabilitation Teaching and Research Unit, University of Otago, Wellington, New Zealand; N. Joseph-Ridge, MD, TAP Pharmaceuticals, Lake Forest, Illinois; F. Perez-Ruiz, MD, Sección de Reumatología, Hospital de Cruces, Vizcaya, Spain; L.X. Chen, MD, PhD, Department of Rheumatology, University of Pennsylvania; N. Schlesinger, MD, Department of Medicine, UMDNJ-Robert Wood Johnson Medical School, New Brunswick, New Jersey; D. Khanna, MD, University of Cincinnati, Cincinnati, Ohio; D.E. Furst, MD, UCLA Medical School, Los Angeles, California; M.A. Becker, MD, Department of Medicine, University of Chicago Medical Center, Chicago, Illinois; N. Dalbeth, MD, Department of Medicine, University of Auckland, Auckland, New Zealand; N.L. Edwards, MD, Department of Medicine, University of Florida, Miami, Florida, USA.

Address reprint requests to Dr. H.R. Schumacher, VA Medical Center 151K, University and Woodland Avenues, Philadelphia, PA 19104, USA. E-mail: schumacr@mail.med.upenn.edu

Chronic Gout

Flare definition. Before being able to validate gout flares as an outcome we had to attempt to define flares. A Delphi exercise to define flare was conducted by Taylor, *et al*[2] with 34 rheumatologists who initially expressed an interest and responded to the first questionnaire: 22 questionnaires were received for a second round, and 21 were received on a third round. The first round identified 43 potential items, of which 6 items had agreement that the items be included in a gout flare definition (median score 7 to 9 on a 9-point scale) and 9 items had agreement that the items not be included in the definition (median score 1 to 3). The remaining 28 items were rescored in a third round, with median scores 4 to 6 indicating that these were neither appropriate nor inappropriate. Since there was apparent uncertainty whether such criteria should be applicable only to those already known to have gout, it has been proposed that the Delphi survey be repeated. In addition, observational studies are required to formally test the accuracy and reliability of such items or combinations of items for determining the presence of a gout flare. Once defined, gout flare reduction could become a useful means of determining response to treatment in the context of intervention studies and in practice.

Other studies recently performed in gout have investigated flares with nonvalidated methods, but do emphasize that flares appear to be common and numbers are reduced with effective treatment. Studies coordinated by Joseph-Ridge, *et al* have examined flares defined by each investigator in trials of urate-lowering agents. In one study[3], the incidence of gout flares during coadministration of colchicine with febuxostat or

placebo was 8%-13%, compared with 30%-42% during febuxostat or placebo administration alone. These self-defined flares by each investigator gradually declined over time with chronic urate-lowering therapy.

In a 52-week study, the effect of febuxostat or allopurinol 300 mg/day was evaluated in 760 subjects[4]. In that study, gout flares were further defined as those requiring treatment. Flares occurred in 21% to 36% of subjects receiving antiinflammatory prophylaxis, across all treatment groups. After prophylaxis was withdrawn during Weeks 8 to 16, 43% to 53% of subjects required treatment for gout flare. The incidence of gout flare diminished thereafter in all groups, and by Weeks 48 to 52 flares were seen in 6%–11%.

Flares, even crudely defined as in these studies, correlated with eventual serum urate levels, and responded to the effect of prophylactic colchicine. Thus, these trials have encouraged us to focus prospectively on flares as chronic outcomes.

Tophus size. No standardized or validated tool is currently available to measure tophus size as an outcome. Several studies have suggested techniques that can be evaluated and encourage further study. A Manual Measurement of Tophi Validation Study has been completed and published[5]. The quantitative evaluation was the area (in mm^2) of each measurable tophus independently measured by 2 raters at each center. Intra- and interrater reproducibility were examined by calculating the mean difference and average percentage difference in areas between visits and raters, respectively. Fifty-two tophi were measured in a small group (13 subjects): 22 tophi were located on the hand/wrist, 16 on the elbow, and 14 on the foot/ankle. The mean (± SD) difference in tophus area between visits was 0 ± 835 mm^2 (95% CI –162 to 162 mm^2) and the mean average percentage difference was 29% ± 33%. The mean average percentage difference between raters was 32% ± 27%. The largest variability in measurements was noted for elbow tophi, which was likely due to bursal fluid. It was felt that study of manual measurement was useful if only tophi on hands and feet are selected.

A recent clinical study (760 subjects treated for 52 weeks) utilized this technique for direct physical measurement of tophi to evaluate change with treatment[6]. Subjects were randomized to febuxostat 80 mg/day or 120 mg/day or allopurinol 300 mg/day. One hundred fifty-six subjects had tophi. When elbows were excluded from the analysis (bursal fluid), the median percentage changes from baseline in tophus area were –87.0%, –72.5%, and –28.7% with febuxostat 80 mg and 120 mg and allopurinol 300 mg, respectively. Thus this technique was able to show a significant difference ($p < 0.05$) between the febuxostat 80 mg group and allopurinol.

Ultrasound (US) and magnetic resonance imaging (MRI) were compared by Perez-Ruiz, *et al* for the evaluation of periarticular tophi[7]. US detected 80% of those found by MRI. MRI detected some tophi anatomically not accessible for US (intercondylar, infrapatellar, and synovial locations), but all periarticular tophi detected by MRI were also detected by US. Thus, US may be as useful as MRI for selected target accessible tophi for followup. Interobserver and intraobserver accuracy of US were tested with 2 examiners (a radiologist and a rheumatologist). The smallest detectable difference (SDD) for intraobserver was small (4 mm) but was greater for interobserver (9 mm), with intraclass correlation > 0.8 (very good: range 0.83–0.92). It was concluded that US shows good reproducibility, and the SDD is small enough to be useful for followup in 6–12 month studies, depending on the range of urate-lowering while on therapy, but preferably the same observer should be used.

Sensitivity to change was evaluated during a 12-month open-label followup study while on urate-lowering therapy[8]. Average serum urate (sUA) ranged from 3.9 to 7.2 mg/dl. The effect size was 2.8 (Guyatt's method); 76% of patients with sUA < 6.0 mg/dl showed reductions in tophus size that were greater than the SDD, while 78% of patients with average sUA levels > 6 mg/dl did not show changes in tophi measurements. US has been shown to meet the OMERACT filter requirements, and further evaluation of changes in tophus size using US in clinical trials is warranted.

Functional status. Functional status has not often been assessed in studies of gout. The original development of the Steinbrocker scale occurred in a population of patients with rheumatoid arthritis (RA), osteoarthritis, and gout[9], but few studies since that time have reported functional measures in patients with gout. The Health Assessment Questionnaire Disability Index (HAQ-DI) is a key patient reported outcome in the majority of rheumatology clinical trials, and is frequently used as part of the core set of the American College of Rheumatology (ACR) clinical response criteria in RA clinical trials[10]. In patients with RA, the HAQ-DI is a strong predictor of healthcare utilization[11], work disability[12], morbidity, and mortality[13].

Taylor, *et al* reported a small observational cohort study of 2 groups of clinic patients with gout (n = 53) and RA (n = 51)[2]. Clinical and functional measures were correlated with HAQ-DI scores, and Rasch analysis was used to determine differences in performance characteristics between patients with gout and RA. Clinical indices correlated highly with HAQ-DI scores in gout patients, particularly with other measures of physical function (SF-36 physical function, $r_s = –0.81$; ACR functional class, $r_s = 0.89$; Sollerman hand function test, $r_s = –0.81$; Disability of Shoulder, Arm and Hand score, $r_s = 0.87$). A stronger relationship between days of sick leave and HAQ-DI was observed in gout patients ($R^2 = 0.44$, $p < 0.001$) compared to patients with RA ($R^2 = 0.20$, $p = 0.02$). HAQ-DI scores showed a bimodal distribution in gout and evidence of floor effects in gout and RA. The HAQ-DI items fitted a Rasch measurement model with an item separation index of 2.19, and Cronbach's alpha was high (0.94). However, there was evidence of differential item functioning and a slightly different relationship between original HAQ-DI scores and Rasch modeled scores observed in gout compared to RA. Two

of 4 items that showed differences in item performance may have been due to a higher proportion of Maori/Pacific people in this New Zealand gout sample. It was concluded that, while HAQ-DI has good construct validity in gout, scores have a different meaning in gout compared to RA, so that direct comparative evaluations across disorders are difficult with this instrument.

Quality of life. No disease-specific validated instrument is currently available to measure gout's effect on daily life. Khanna reported an ongoing study led by Hirsch that is also addressing the risk for gout patients of experiencing decreased health-related quality of life. Researchers in San Diego, Cincinnati, and Minneapolis, USA, assessed the psychometric properties of a gout-specific patient-reported outcomes (PRO) instrument designed to investigate the effect of gout during and between attacks in a community population.

The instrument content was based on a previously developed version used in clinical trials, interviews with gout patients, and an expert panel consisting of rheumatologists and psychometricians[14]. Adults diagnosed with gout (ACR preliminary criteria) were recruited from clinics (rheumatology, family practice, internal medicine) and surrounding communities in 3 US cities. As part of the validation, subjects completed the gout-specific PRO instrument, symptom questions (e.g., attack frequency and duration), and the SF-36 via mail survey or in the clinic. There were 4 scales: "Gout impact" (GI), "Gout pain and severity between flares" (GPSB), "Well-being impact" (WBI), and "Productivity" (P). Confirmation of gout diagnosis, presence of tophi, and physician-rated severity were obtained from subjects' physicians.

Preliminary analysis has been performed on a subset of participants (n = 151) who were primarily male (85%), White (73%), and had a mean age of 61.1 years. Physicians rated severity as mild (51.1%), moderate (35.6%), and severe (13.3%), while the average subject's rating of disease severity was moderate (54.6 on a 100 mm visual analog scale). Twenty-four percent of subjects had tophi. The OMERACT filters of truth (face, content, construct validity), discrimination, which includes reliability and sensitivity to change (test-retest, internal consistency), and feasibility were assessed. Face and content validity was demonstrated by input from 2 focus groups and expert opinion. Convergent construct validity was demonstrated by moderate correlations between GPSB and the SF-36 Physical Summary Scale (product moment correlation, r = –0.40) and "Well-being impact" with the SF-36 Mental Summary Scale (r = –0.35). Divergent construct validity was supported by significant differences in the "Gout impact" and "Gout pain and severity between flares" scales among subjects with increasing gout severity [physician-rated severity and report of tophi and patient-reported number of attacks last year ($p < 0.05$ for all)]. Internal consistency (Cronbach's alpha) was > 0.70, and test-retest correlations were between 0.68 and 0.81 for the 4 scales.

This gout-specific PRO instrument demonstrated acceptable validity and reliability for measuring the impact of gout across patients with differing gout severity characteristics. Further study will examine factor structure, the ability to detect change over time, and define minimal clinically important differences.

Imaging. Joint imaging as a chronic outcome domain has had a preliminary validation as a scoring method for radiological damage by Dalbeth, *et al* (data not published). The evaluation of damage in individual joints affected by chronic gout is a preliminary step toward developing a radiographic damage index. Following a structured review of plain radiographs of an initial small group of 12 patients with chronic gout, 3 rheumatologists independently scored 95 hand proximal interphalangeal joints on a scale of 0 to 10 (normal to severely affected). After a consensus exercise, the final ratings were averaged to form the consensus global score, which was used as a "gold standard" index of joint damage. The same joints were independently assessed by a radiologist for Sharp-van der Heijde (S-vdH) erosion score, s-vdH joint space narrowing score, Ratingen destruction score, and Steinbrocker global score. Each score and combinations of these scores were compared with the consensus global score.

Analysis showed that the combination of the S-vdH erosion and narrowing scores was most strongly correlated with the global consensus scores ($r = 0.88$, $p < 0.001$), and that these 2 scores independently predicted the global consensus scores. Further, the limits of agreement for the mean difference between the scoring method and the consensus global scores were narrowest for the combined S-vdH erosion and narrowing scores. Thus, the combined S-vdH erosion and narrowing score adequately represents radiological damage in individual joints affected by chronic gout. Further work will determine the number and sites of joints that should be incorporated into a chronic gout radiology scoring method. A discussion of the use of ultrasound and MRI to evaluate tophi is included above.

Patient-oriented domains. Edwards is carrying out a patient survey to better appreciate which patient-oriented domains might ultimately be included in the core set of domains for outcome measures. An open-ended questionnaire was sent to 50 patients from the VA Medical Center in Gainesville, Florida, who had been followed by the Rheumatology Service and had a definite diagnosis of gout. Twenty-eight questionnaires were returned in the 6-week time limit required. The questionnaire contained 2 questions, and subjects were asked to submit as many responses as they felt appropriate for each question. The questions were designed to elicit feelings about the disease itself and about the therapy for gout. Question 1 stated, "Of all the ways that gout affects your life and the way you feel, which are the greatest problems for you?". The second question stated, "Of all the ways your gout treatment/therapy affects your life, which are the greatest problem for you?". The 28 responders submitted a total of 320 replies for Question 1 and 186 replies for Question 2, for an average of

11.4 for Question 1 and 6.6 for Question 2. The 320 responses to Question 1 were divided into 22 separate bins, and the 186 responses to Question 2 were divided into 14 separate response bins.

The rank order for the responses to the function/activity question (Question 1) was determined by the number of responses in each bin. The top 5 patient-perspective responses for Question 1 were: (1) pain, (2) immobility during flares, (3) unpredictability of attacks, (4) dependency on spouse/limitation of activities of daily living, and (5) abandonment of hobbies/leisure activities.

The rank order of the patient-perspective responses to the treatment and therapy question (Question 2) was as follows: (1) the need for lifelong therapy, (2) adverse drug effects/gastrointestinal symptoms, (3) ineffectiveness of therapies, (4) flares during uric acid-lowering treatment, and (5) the need for multiple drugs to treat gout.

The top 10 responses for both Question 1 and Question 2 were presented at the meeting. Phase II of this Delphi study will involve rank-order queries to 130 patients with active gout in the Gainesville VA Medical Center, including 42 patients with only acute, intermittent symptoms and 78 patients with evidence of advanced gout. SIG members in Cincinnati, Ohio, and Barcelona, Spain, have offered to expand this study on their very different patient populations. Other sites will also be sought.

Acute Gout

Validation of core set domains. We have identified 3 new clinical trials on urate-lowering in which outcomes can be validated. For evaluation of outcomes in acute gouty arthritis we have identified 5 core domains: (1) pain, (2) inflammation, (3) function, (4) patient global assessment, and (5) safety. A Delphi survey of importance of the different domains has just been completed.

Schlesinger and colleagues examined the validity of the 0–4 Likert pain scale in acute gouty arthritis (data not published). They used the OMERACT filter paradigm to estimate the validity of acute gouty arthritis patients' reported pain assessment within each of the filter's 3 components: truth, discrimination, and feasibility. The data source was 339 subjects from 2 identical parallel-group, 7-day randomized controlled trials for treatment of acute gout[15,16]. They assessed different properties of validation — construct validity, discriminant validity, and responsiveness.

The results showed that the patient pain assessment question had good construct validity measured by Spearman correlation coefficients of 0.48–0.57 (all $p < 0.001$). The question had good face validity in that it is clear and unambiguous. The question was able to discriminate between patient groups categorized according to responses given on the Patient's Global Assessment of Response to Treatment (PGART), Investigator's Global Assessment of Response to Treatment (IGART), and discontinuation due to lack of efficacy. The question was responsive to changes over time, as shown by large effect-size statistics. Since this was a 1-item assessment, measures of reliability were not applicable. It was concluded that the categorical daily patient-reported pain assessment (Likert scale) is a valid and sensitive measure to assess treatment in acute gout.

Classification

Disease definition. Although classification of disease is not a focus of OMERACT it has become obvious, for example, that definition of a flare is essential before we can validate its use as an outcome or response criterion. In addition, definition of gout can influence who is included in studies and the validation of outcomes. An ongoing study at Philadelphia and Pittsburgh Veterans Affairs Medical Centers[17] examined whether two ICD-9 diagnoses of gout in the electronic record can be validated as useful to confirm actual gout through examination of patient records. One hundred sixty-five charts with two ICD-9 coded encounters for gout have been reviewed. Outpatient visits from all clinics and inpatient records were examined. By various criteria no more than 58 (35%) met any clinical criteria of gout. Interestingly, 157 patients (95%) had used gout medicines (allopurinol, colchicine). Only 66 (40%) of those taking gout medications actually had any evidence that active gout or gout management was addressed at visits. Seventy-four percent of these selected records with some mention of gout in the clinic notes met the published Rome criteria and 44% met the ACR preliminary criteria for gout. Documentation in the medical record may be incomplete or the presence of gout may be recorded incorrectly. Variations among patients seen by rheumatologists versus others still must be studied. Further prospective studies of criteria are under way.

Agenda

We will be able to validate a definition of a gouty arthritis flare as we have received an ACR-EULAR response criteria grant to do so. This will involve a new Delphi exercise, a nominal group meeting, clinical validation, and finally, examination of ability to detect change in planned longitudinal clinical therapy trials that will provide needed large numbers of patients.

We will expand our efforts started on patient-oriented domains based on the OMERACT 8 plenary session on the International Classification of Functioning, Disability and Health (ICF).

Each of the projects presented as progress reports will undergo full validation with the OMERACT filter for presentation and development of consensus at OMERACT 9.

A thorough review article will be prepared to examine the range and properties of outcome measures in gout studies.

REFERENCES

1. Schumacher HR, Edwards NL, Perez-Ruiz F, et al. Outcome measures for acute and chronic gout. J Rheumatol 2005;32:2452-5.

2. Grainger R, Harrison AA, Taylor WJ. Preliminary identification of potential items for a definition of "gout flare" using Delphi methodology [abstract]. Arthritis Rheum 2005; 52 Suppl:S105.
3. Becker MA, Schumacher HR, Wortmann RL, et al. Febuxostat, a novel nonpurine selective inhibitor of xanthine oxidase. Arthritis Rheum 2005;52:916-23.
4. Becker MA, Schumacher HR, Wortmann RL, et al. Reduction in gout flares in subjects with chronic gout treated with febuxostat or allopurinol for 52-weeks: FACT trial [abstract]. Arthritis Rheum 2005;52 Suppl:S108.
5. Schumacher HR, Becker MA, Palo WA, Streit J, MacDonald PA, Joseph-Ridge N. Tophaceous gout: quantitative evaluation by direct physical measurement. J Rheumatol 2005;32:2368-72.
6. Wortmann RL, Schumacher HR, Becker MA, et al. Reduction in tophus size in subjects with chronic gout treated with febuxostat or allopurinol for 52 weeks — FACT trial [abstract]. Arthritis Rheum 2005;52 Suppl:S108.
7. Perez-Ruiz F, Martin JI, Canteli B, Sole JMN. Ultrasonography (US) is as useful as magnetic resonance imaging (MRI) for the evaluation of deep periarticular tophi in chronic gout [abstract]. Arthritis Rheum 2005;52 Suppl:S106.
8. Perez-Ruiz F, Martin JI. Serum urate levels correlate with changes in tophi measured with US during urate-lowering therapy. Rheumatol Clin 2006;2 Suppl:11.
9. Steinbrocker O. Prognosis for employability in the major arthritides rheumatoid arthritis, osteoarthritis and gout. Pa Med 1969;72:82-5.
10. Felson DT, Anderson JJ, Boers M, et al. 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;36:729-40.
11. Michaud K, Messer J, Choi HK, Wolfe F. Direct medical costs and their predictors in patients with rheumatoid arthritis — A three-year study of 7,527 patients. Arthritis Rheum 2003;48:2750-62.
12. Wolfe F, Hawley DF. The longterm outcomes of rheumatoid arthritis: Work disability: a prospective 18 year study of 823 patients. J Rheumatol 1998;25:2108-17.
13. Wolfe F, Michaud K, Gefeller O, Choi HK. Predicting mortality in patients with rheumatoid arthritis. Arthritis Rheum 2003; 48:1530-42.
14. Osterhaus JT, Patel P, Palo WA, Bakst A, Joseph-Ridge N. Patient-reported outcomes associated with gout: baseline results from two clinical trials [abstract]. Arthritis Rheum 2005;52 Suppl:S401.
15. Schumacher HR, Boice JA, Daikh DI, et al. Randomized double-blind trial of etoricoxib and indomethacin in treatment of acute gouty arthritis. BMJ 2002;324:488-92.
16. Rubin BR, Burton R, Navarra S, et al. Efficacy and safety profile of treatment with etoricoxib 120 mg once daily compared with indomethacin 50 mg three times daily in acute gout: A randomized controlled trial. Arthritis Rheum 2004;50:598-600.
17. Schumacher HR, Chen LX, Kwan-Morley J, Malik A, Dinnella JE, Kwoh CK. Validation of medical record gout diagnoses for research using cases in the VA database (abstract). Arthritis Rheum 2006;54 Suppl:S644.