Application of the OMERACT Filter to Measures of Core Outcome Domains in Recent Clinical Studies of Acute Gout

William J. Taylor, David Redden, Nicola Dalbeth, H. Ralph Schumacher, N. Lawrence Edwards, Lee S. Simon, Markus R. John, Margaret N. Essex, Douglas J. Watson, Robert Evans, Keith Rome, and Jasvinder A. Singh

ABSTRACT. ***Objective.*** To determine the extent to which instruments that measure core outcome domains in acute gout fulfill the Outcome Measures in Rheumatology (OMERACT) filter requirements of truth, discrimination, and feasibility.

Methods. Patient-level data from 4 randomized controlled trials of agents designed to treat acute gout and 1 observational study of acute gout were analyzed. For each available measure, construct validity, test-retest reliability, within-group change using effect size, between-group change using the Kruskall-Wallis statistic, and repeated measures generalized estimating equations were assessed. Floor and ceiling effects were also assessed and minimal clinically important difference was estimated. These analyses were presented to participants at OMERACT 11 to help inform voting for possible endorsement.

Results. There was evidence for construct validity and discriminative ability for 3 measures of pain [0 to 4 Likert, 0 to 10 numeric rating scale (NRS), 0 to 100 mm visual analog scale (VAS)]. Likewise, there appears to be sufficient evidence for a 4-point Likert scale to possess construct validity and discriminative ability for physician assessment of joint swelling and joint tenderness. There was some evidence for construct validity and within-group discriminative ability for the Health Assessment Questionnaire as a measure of activity limitations, but not for discrimination between groups allocated to different treatment.

Conclusion. There is sufficient evidence to support measures of pain (using Likert, NRS, or VAS), joint tenderness, and swelling (using Likert scale) as fulfilling the requirements of the OMERACT filter. Further research on a measure of activity limitations in acute gout clinical trials is required. (J Rheumatol First Release Jan 15 2014; doi:10.3899/jrheum.131245)

Key Indexing Terms:
GOUT OUTCOME MEASURES PSYCHOMETRICS

From the Department of Medicine, University of Otago, Wellington, New Zealand; Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama, USA; Department of Medicine, University of Auckland, Auckland, New Zealand; University of Pennsylvania and Veterans Affairs (VA) Medical Center, Philadelphia, Pennsylvania, USA; Department of Medicine, University of Florida, Gainesville, Florida, USA; SDG LLC, Cambridge, Massachusetts, USA; and Integrated Hospital Care Franchise, Immunology, Novartis Pharma AG, Basel, Switzerland; Pfizer Inc., New York, New York, USA; Epidemiology, Merck Sharp & Dohme Corp., Whitehouse Station, New Jersey, USA; Clinical Sciences, Regeneron Pharmaceuticals, Tarrytown, New Jersey, USA; Health & Rehabilitation Research Institute and School of Podiatry, Auckland University of Technology, Auckland, New Zealand; and Birmingham VA Medical Center and University of Alabama at Birmingham, Birmingham, Alabama, USA.

Supported with resources and use of facilities at the Birmingham VA Medical Center, Alabama, USA (J.A.S. and D. Redden). N. Dalbeth has received consulting fees from Ardea Biosciences, Metabolex, Novartis, and Takeda. Her institution has received funding from Fonterra, and she is a named inventor on a patent related to milk products and gout. H.R. Schumacher has been a consultant for Regeneron, Novartis, Pfizer, Savient, Ardea, Metabolex, and BioCryst, and he has received a grant from Takeda. N.L. Edwards has received consultant fees from Novartis, Takeda Pharmaceutical, Savient Pharmaceutical, Ardea Biosciences, Regeneron Pharmaceuticals, Metabolex Pharmaceuticals, and BioCryst Pharmaceuticals. L.L. Simon has served on the board of directors for Savient Pharmaceuticals, and has consulted for Takeda. M.R. John is employed by Novartis and sometimes owns shares in the company. M.N. Essex is employed by Pfizer and owns shares in the company. D.J. Watson is an employee of and owns stock in Merck & Co. Inc.; the marketing authorization holder for etoricoxib and sponsor of the etoricoxib clinical trials that contributed data for this work. R. Evans is employed by Regeneron and owns shares of stock. J.A. Singh has received research grants from Takeda and Savient and consultant fees from Savient, Takeda, Ardea, Regeneron, Allergan, URL Pharmaceuticals and Novartis. He is a member of the executive of OMERACT, an organization that develops outcome measures in rheumatology and receives arms-length funding from 36 companies; a member of the American College of Rheumatology's Guidelines Subcommittee of the Quality of Care Committee; and a member of the Veterans Affairs Rheumatology Field Advisory Committee.

W.J. Taylor, PhD, FRACP, Associate Professor, Department of Medicine, University of Otago; D. Redden, PhD, Associate Professor, Department of Biostatistics, School of Public Health, University of Alabama at Birmingham; N. Dalbeth, MD, FRACP, Associate Professor, Department of Medicine, University of Auckland; H.R. Schumacher, MD, Professor, University of Pennsylvania and VA Medical Center; N.L. Edwards, MD, Professor, Department of Medicine, University of Florida; L.S. Simon, MD, SDG LLC; M.R. John, MD, Global Program Medical Director, Integrated Hospital Care Franchise, Immunology, Novartis Pharma AG; M.N. Essex, PharmD, Senior Medical Director, Pfizer Inc.; D.J. Watson, PhD, FISPE, Epidemiology, Merck Sharp & Dohme Corp.; R. Evans,

PharmD, Senior Director, Clinical Sciences, Regeneron Pharmaceuticals; K. Rome, PhD, Professor, Health and Rehabilitation Research Institute and School of Podiatry, Auckland University of Technology; J.A. Singh, MBBS, MPH, Associate Professor, Birmingham VA Medical Center and University of Alabama at Birmingham.

Address correspondence to Prof. Taylor, Department of Medicine, University of Otago, PO Box 7343, Wellington, New Zealand. E-mail: will.taylor@otago.ac.nz

At the 11th Outcome Measures in Rheumatology (OMERACT) meeting, held in May 2012, the focus of the Gout Module was to obtain endorsement of specific instruments that measure each of the 5 core domains identified at OMERACT 9 as key outcomes in acute gout trials[1]. To assist participants in determining whether specific instruments met the OMERACT filter of truth, discrimination, and feasibility necessary for adequate technical performance of outcome instruments, we aimed to calculate the key psychometric properties from recent trials or observational studies of acute gout.

MATERIALS AND METHODS

Patient-level data were generously provided by Merck Sharp & Dohme Corp. (MSD), Novartis, Pfizer, and Regeneron concerning 4 trials of treatment with etoricoxib, canakinumab, celecoxib, and rilonacept, respectively. Treatment allocation was not made available for the canakinumab study (Novartis) because trial results were in publication at the time of this analysis[2]; nor for the etoricoxib (MSD) dataset. In addition, data from a small observational cohort study of acute gout was provided by Professor Keith Rome (Auckland University of Technology)[3]. The key characteristics of each study are shown in Table 1 and 2. Note that all studies were active-controlled, although the celecoxib study included an arm with a lower than recommended dose of celecoxib. These studies were pragmatically selected on the basis of availability of patient-level data with which to perform secondary analysis, studies with drugs of different biological mechanisms, and studies of both randomized controlled trials (RCT) and longitudinal observational studies. A systematic review of published trials of acute gout was performed separately and is reported in a companion article[4].

Each of the included studies had previously received ethical approval from appropriate ethical review board, and provision of patient-level data to the authors was within the permission given by patients at informed consent.

Construct validity, or the extent to which the instrument was closely associated with similar concepts and not closely associated with dissimilar concepts, was assessed using Spearman correlation coefficients between each instrument measured at the baseline timepoint. Floor and ceiling effects were calculated as the percentage of participants scoring the minimum and maximum possible at baseline and final visit. Within-group discrimination was assessed within each study by pooling the change scores of each instrument and calculating the effect size (ES). Between-group discrimination was assessed by calculating the Kruskal-Wallis statistic for the difference between the final reported value of each measure across treatment arms. Within- and between-group change was also assessed using repeated measures generalized estimating equations with ordinal regression to maximize information available from multiple timepoints (for example, pain was measured at several timepoints).

Test-retest reliability was calculated using patient global assessment (PGA) of response to identify a subset of participants who perceived no change. To identify a stable group in the etoricoxib clinical trial we selected cases with the same patient perception of response at days 2 and 5 and at days 5 and 8, in 2 separate estimations of reliability. In the celecoxib clinical trial we selected the low-dose celecoxib cases for the analysis over the first 12 h and cases with poor or fair response at Day 9 for the analysis over 9 days. The intraclass correlation (ICC) used a mixed-effects model for single measure absolute agreement in stable cases. The standard error of measurement (SEM) was calculated as the square root of the error variance from the analysis of variance table from whence the ICC was calculated. Smallest detectable difference (SDD) was calculated as $SEM \times \sqrt{2} \times 1.96$[5]. The minimal clinically important difference (MCID) was calculated as the median value of change in each measure for the "fair" category of patient global response to treatment, where this was available[6].

RESULTS

Feasibility (time to completion, cost, respondent burden) were not formally assessed in any study, but all instruments appear to be easy to complete with no or minimal need for training and no or little cost.

Pain Measures

Three pain measures were used in different studies: 0–4 point Likert-like scale, 0–100 mm visual analog scale (VAS), and 0–10 numeric rating scale (NRS). Data for the NRS were derived from a single unpublished study, and therefore most discussion focused on the Likert scale and

Table 1. Data sources for validation data related to measures of 5 acute gout domains.

Source	n	Inclusion	Treatment Groups	Publication (reference)
Merck Sharp & Dohme	150	Onset within 48 h, 1977 ARA criteria, at least moderate pain	Etoricoxib, indomethacin	(7)
Pfizer	402	Onset within 48 h, 1977 ARA criteria, at least moderate pain	Celecoxib 50, 400/200, 800/400, indomethacin	(8)
Regeneron	225	Onset within 48 h, 1977 ARA criteria, at least moderate pain	Indomethacin, rilonacept and indomethacin, rilonacept	Not published [NCT00855920]
Novartis (2 replicate studies)	424	Onset of acute flare within 5 days, 1977 ARA criteria, at least 3 flares within prior 12 mo, pain at least 50 mm on 100 mm VAS	Canakinumab 150 mg SQ, triamcinolone 40 mg IM	(2)*
Auckland University of Technology	20	Observational study, acute gout flare at baseline, 1977 ARA criteria	Not applicable	(3)

* Not published prior to data analyses and presentation but now published; dataset provided to investigators was a 90% random subsample of the main study dataset (n = 456). ARA: American Rheumatology Association; VAS: visual analog scale.

Table 2. Instruments available for each data source.

Source	Pain	Disability	Joint Swelling/ tenderness	Patient Global
Merck Sharp & Dohme	Likert 0–4	NA	Likert 0–3*	Response to treatment (Likert 0–4 point)
Pfizer	Likert 0–4	NA	Likert 0–3*	Response to treatment (Likert 0–4 point)
Regeneron	Likert 0–4, NRS 0–10	Activity limitations, NRS 0–10 (from WPAI:SHP v2.0)	NA†	No measure† available
Novartis	Likert 0–4, VAS 0–100	HAQ-DI	Likert 0–3*	Response to treatment (Likert 0–4 point)
Auckland University of Technology	VAS 0–100	HAQ-II	Swollen and tender joint count	VAS 0–100

* Index joint assessed by a physician; † Likert 0–3 grade for joint tenderness and swelling was used in the actual trial but those data were not available for the current analysis. NA: no measure available; HAQ-II: Health Assessment Questionnaire version II; HAQ-DI: Health Assessment Questionnaire Disability Index; VAS: visual analog scale; WPAI:SHP: Worker Productivity and Activity Impairment Index (Specific Health Problem)[9]; NRS: numeric rating scale.

VAS scales, for which there were data from more than 1 RCT and more than 1 class of drugs (Table 2).

Likert-like scale. A 0–4 point Likert scale was used in most studies with categories of "none" (0), "mild," "moderate," "severe," and "extreme" (4) pain. The Likert scale had good construct validity (Table 3): strong correlation with patient global (Spearman's correlation coefficient, 0.72) and NRS pain score (0.55 and 0.73), moderate-strong correlation with disability (0.58 and 0.31) and moderate correlation with joint tenderness (0.34, 0.36, 0.13), but weaker correlation with joint swelling (0.18, 0.18, 0.19).

ES ranged from 1.20 to 2.84, demonstrating a large ES

Table 3. Construct validity showing Spearman correlation coefficients for each measure.

Source Measure	Pain (VAS or NRS)	Joint Tenderness	Joint Swelling	Activity Limitations*	Patient Global‡
MSD					
Pain (Likert)	NA	0.34	0.18	NA	NA
Joint tenderness			0.25	NA	NA
Joint swelling				NA	NA
Activity limitations					NA
Pfizer					
Pain (Likert)	NA	0.36	0.18	NA	NA
Joint tenderness			0.37	NA	NA
Joint swelling				NA	NA
Activity limitations					NA
Regeneron					
Pain (Likert)	0.75	NA	NA	0.31	NA
Pain (NRS)		NA	NA	0.39	NA
Joint tenderness			NA	NA	NA
Joint swelling				NA	NA
Activity limitations					NA
Novartis					
Pain (VAS)	0.55	0.13	0.19	0.58	0.72
Pain (Likert)		0.15	0.17	0.58	0.70
Joint tenderness			0.46	0.18	0.56
Joint swelling				0.25	0.47
Activity limitations					0.50
AUT					
Pain VAS	NA	NA	NA	0.66	0.73
Joint tenderness			NA	NA	NA
Joint swelling				NA	NA
Activity limitations†					0.73

* Activity limitations measured by single 0–10 NRS in Regeneron data, HAQ-II in AUT data, and HAQ-DI in Novartis data. † In addition the HAQ-II correlated highly with measures of specific foot function in this dataset. ‡ Changes in each measure were correlated with patient global because the patient global represented perception of change (except for the AUT dataset). NA: measure not available in the dataset; AUT: Auckland University of Technology; VAS: visual analog scale; NRS: numeric rating scale; MSD: Merck Sharp & Dohme Corp.

over time (Table 4). The Likert scale discriminated well between treatment groups, with minimal clinically important difference (MCID) ranging from a change of 1 to 2. Floor effects were appreciable at final visit and ceiling effects were appreciable at baseline (Table 5).

Pain visual analog scale (VAS). A VAS pain scale 0 to 100 mm was used in 2 studies. The VAS pain scale had good construct validity: strong correlation with patient global (0.72 and 0.73 in 2 studies), and with disability (0.58 and 0.66), but weak correlation with joint swelling (0.19) or joint tenderness (0.13).

ES ranged from 1.58 to 4.46, demonstrating a large ES over time. VAS pain scale discriminated well between treatment groups as recently reported[7], with MCID of 19 on 0–100 mm scale. Minimal floor effects were appreciable at final visit (14%) and minimal ceiling effects were appreciable at baseline (13%).

Numeric rating scale. One study of rilonacept used both Likert scale and NRS. Based on this single study, NRS pain seemed to have face, content, and construct validity, and was sensitive to change (within and between group).

Joint Swelling

A 0–3 point Likert scale used in most studies was examined in this analysis, typical categories being "no swelling" (0), "palpable," "visible," and "bulging beyond the joint margins" (3) in the index joint, as assessed by a physician. The Likert scale had evidence for construct validity with moderate correlation with patient global (0.47) and activity limitation as measured by Health Assessment Questionnaire (HAQ; 0.25) and with joint tenderness (0.25, 0.37) and weak correlation with pain (0.14, 0.18). In treatment trials of canakinumab, Likert scale showed between-group, as reported[2], and within-group differences (Table 6). ES ranged from 2.3 to 2.9. In this analysis, the MCID for joint swelling corresponded to a change of 1 on the Likert scale. Significant floor effects were appreciable at final visit (47 to 64%) and ceiling effects (27 to 56%) were appreciable at baseline.

Table 4. Indices of discrimination.

Measure	Source	Within Group (pooled)		Between Group	
		Effect Size	† GEE, Wald Chi-square	KW Statistic	† GEE, Wald Chi-square
Pain (Likert)	MSD*,	2.32	NA	NA	NA
	Pfizer,	2.72	816, p < 0.001	17.6, p = 0.001	16.8, p = 0.001
	Regeneron,	1.20	NA	26.7, p < 0.001	NA
	Novartis*	2.84	446.8, p < 0.001	NA	NA
Pain (VAS)	AUT*,	1.58	NA	NA	NA
	Novartis*	4.46	602.2, p < 0.001	NA	NA
Pain (NRS)	Regeneron	1.62	NA	26.6, p < 0.001	NA
Joint tenderness	MSD*,	3.2	NA	NA	NA
	Pfizer,	2.5	542, p < 0.001	1.7, p = 0.67	12, p = 0.01
	Regeneron,	NA	NA	NA	NA
	Novartis*,	2.25	598, p < 0.001	NA	NA
	AUT*	NA	NA	NA	NA
Joint swelling	MSD*,	2.9	NA	NA	NA
	Pfizer,	2.3	561, p < 0.001	2.2, p = 0.54	4.0, p = 0.26
	Regeneron,	NA	NA	NA	NA
	Novartis*,	2.5	523, p < 0.001	NA	NA
	AUT*	NA	NA	NA	NA
Activity limitations	MSD*,	NA	NA	NA	NA
	Pfizer,	NA	NA	NA	NA
	Regeneron,	0.81	NA	5.4, p = 0.067	NA§
	Novartis*,	1.04	159, p < 0.001	NA	NA
	AUT*	1.72	NA	NA	NA
Patient global	MSD*,	NA‡	NA	NA	NA
	Pfizer†,	NA‡	NA	5.5, p = 0.14	NA§
	Regeneron,	NA	NA	NA	NA
	Novartis*,	NA‡	NA	NA	NA
	AUT*¶	1.46	NA	NA	NA

* MSD: Merck Sharp & Dohme Corp. Treatment allocation not available or not relevant therefore between-group discrimination was not assessable. † Repeated measures GEE with ordinal regression performed in Pfizer and Novartis datasets; ‡ No baseline measure since it assessed response to treatment; § Not measured at multiple timepoints; ¶ PGA measured with 100 mm visual analog scale for current status (all other studies used global response to treatment). NA: not available or not applicable; GEE: generalized estimating equations; NRS: numeric rating scale; AUT: Auckland University of Technology; PGA: patient global assessment.

Table 5. Floor (percentage of participants at minimum possible value) and ceiling (percentage of participants at maximum possible value) effects.

Measure	Source	Floor (%)		Ceiling (%)	
		Baseline	Final*	Baseline	Final*
Pain (Likert)	MSD,	0	42	21.7	1.8
	Pfizer,	0	35.4	17.7	0.8
	Regeneron,	0	11.6	11.1	4.0
	Novartis	0.24	28.25	12.59	1.25
Pain (VAS)	AUT†,	0	33.3	5	0
	Novartis	0	14.4	2.1	0
Pain (NRS)	Regeneron	0	11.6	9.3	4.9
Joint tenderness	MSD,	0	50.0	57.5	3.0
	Pfizer,	0.5	44.1	39.1	4.7
	Regeneron,	NA	NA	NA	NA
	Novartis,	0.71	55.1	43.4	2.64
	AUT†	NA	NA	NA	NA
Joint swelling	MSD,	0	52.1	56.0	5.4
	Pfizer,	2.2	47.2	27.1	4.5
	Regeneron,	NA	NA	NA	NA
	Novartis,	1.4	63.7	35.1	2.2
	AUT†	NA	NA	NA	NA
Disability‡	MSD,	NA	NA	NA	NA
	Pfizer,	NA	NA	NA	NA
	Regeneron,	11.7	33.2	8.1	3.5
	Novartis,	5.63	46.19	0.43	0
	AUT†	0	0	0	16.7
Patient Global Assessment¶	MSD,	NA	4.5	NA	26.4
	Pfizer,	NA	2.8	NA	40.1
	Regeneron,	NA	NA	NA	NA
	Novartis,	NA	2.1	NA	39.1
	AUT†	0	0	5	0

* Refers to Day 5 unless mentioned specifically; † Final value at 6 to 8 wks; ‡ Measured by Health Assessment Questionnaire version II in AUT, Health Assessment Questionnaire Disability Index in Novartis, Worker Productivity and Activity Impairment Index 0–10 in Regeneron; ¶ Final value at Day 9 for Pfizer and Novartis. NA: measure not available; VAS: visual analog scale; NRS: numeric rating scale; AUT: Auckland University of Technology; HAQ-II: Health Assessment Questionnaire version II; HAQ-DI: Health Assessment Questionnaire Disability Index; WPAI:SHP: Worker Productivity and Activity Impairment Index (Specific Health Problem)[9]; MSD: Merck Sharp & Dohme Corp.

Joint tenderness

Joint tenderness was also measured using a 0–3 point Likert scale in most studies. An example of a 0–3 point Likert scale used in the Novartis studies: no pain (0), patient states that "there is pain" (1), patient states "there is pain and winces" (2), and patient states "there is pain, winces and withdraws" on palpation or passive movement of the affected study joint, as assessed by a physician (3). Joint tenderness Likert scale had strong correlation with patient global (0.56), moderate correlation with joint swelling (0.25, 0.37, 0.46) and with pain (0.19, 0.34, 0.36; Table 3). The ES for the Likert scale ranged 2.3 to 3.2, and the measure discriminated between treatment groups in 1 study that we analyzed, as well as a recently published analysis of duplicate RCT for canakinumab[2]. The MCID for joint tenderness ranged from 1 to 2. We observed significant floor effects at final visit (44 to 55%) and ceiling effects (39 to 58%) at baseline.

Patient Global Assessment

The patient global measure used in most studies was a 0–4 point Likert scale of global assessment of response to therapy. For example, in the etoricoxib clinical trial, the global response to treatment was assessed with the question: "How would you rate the study medication you received for gout?" with these response options: Excellent = 0, Very good = 1, Good = 2, Fair = 3, Poor = 4. The only study that used a global assessment of current status was the Auckland University of Technology observational study that used a 100 mm VAS, asking participants to rate how well they were doing overall.

PGA is usually the external benchmark for all other outcome measures, including several described above. Therefore, it has face, content, and construct validity almost by definition. Typically PGA relate to assessment of current disease status; however, all but 1 study provided data for

Table 6. Indices of test-retest reliability, smallest detectable difference (SDD) and minimal important difference (MID).

			ICC	SEM	SDD	MID
Pain (Likert)	MSD	Between day 2 and 5	0.56	0.51	1.41	1
		Between day 5 and 8	0.80	0.42	1.17	2
	Pfizer*	Between 2 and 4 h	0.81	0.39	1.08	
		Between 2 and 8 h	0.72	0.50	1.39	
		Between 2 and 12 h	0.60	0.62	1.72	
	Pfizer†	Between day 1 and 9	0.07	0.76	2.1	2
		Between day 2 and 9	0.15	0.76	2.1	2
		Between day 5 and 9	0.59	0.54	1.50	2
	Novartis	Baseline to 7 days postdose	0.35	0.85	2.36	1.0
		24 h postdose to 7 days	0.55	0.64	1.77	
		48 h postdose to 7 days	0.71	0.49	1.36	
Pain (VAS)	Novartis	Baseline to 7 days postdose	0.35	3.66	10.15	19
		24 h postdose to 7 days	0.57	2.93	8.12	
		48 h postdose to 7 days	0.76	2.23	6.18	
Joint tenderness	MSD	Between day 2 and day 5	0.50	0.46	1.28	2
		Between day 5 and day 8	0.79	0.34	0.94	1
	Pfizer	Between Day 1 and 9	0.06	0.66	1.8	2
		Between Day 5 and 9	0.11	0.59	1.6	2
	Novartis	Baseline to 7 days postdose	0.0**	1.06	2.93	1.0
		24 h postdose to 7 days	0.50	0.54	1.51	
		48 h postdose to 7 days	0.49	0.54	1.50	
		72 h postdose to 7 days	0.49	0.54	1.50	
Joint swelling	MSD	Between day 2 and day 5	0.48	0.53	1.47	1
		Between day 5 and day 8	0.77	0.43	1.18	1
	Pfizer	Between Day 1 and 9	0.13	0.64	1.8	1
		Between Day 5 and 9	0.37	0.73	2.0	1
	Novartis	Baseline to 7 days	0.0	1.07	2.97	1
		24 h postdose to 7 days	0.44	0.65	1.80	
		48 h postdose to 7 days	0.44	0.65	1.79	
		72 h postdose to 7 days	0.44	0.65	1.79	
Activity limitations	Novartis		0.55	0.45	1.25	0.5

* Pain assessed as "current" level of pain; † pain assessed as "over the last 24 h;" ** Statistical software indicated that estimation of a negative variance parameter was attempted. MSD: Merck Sharp & Dohme Corp.; VAS: visual analog scale; ICC: intraclass correlation; SEM: standard error of measurement.

PGA of response to treatment. Application of the OMERACT filter to a transition scale such as this is problematic. Reliability could not be determined, because we used the responses on this measure to define a stable subgroup. Within-group change was not meaningful for a measure that had no meaning at baseline. For the single study that used a conventional PGA, an ES of 1.46 suggested adequate within-group change sensitivity for that format.

In the only RCT that provided both treatment allocation and measured a global response to treatment (celecoxib study), we did not observe a between-group difference (Table 5).

Activity Limitation

Activity limitation data were available from 3 studies. Two studies used the HAQ-disability index or HAQ-II, and one study used a 0–10 NRS item from the Worker Productivity and Activity Index: Specific Health Problem (WPAI:SHP) scale as a measure of activity limitations.

Health Assessment Questionnaire

HAQ scores showed strong correlation with patient global (0.50, 0.73), moderate correlation with joint swelling (0.31), moderate to strong correlation with pain (0.26, 0.33, 0.37, 0.66), and moderate correlation with joint tenderness (0.46). The ES was moderate to large, ranging from 1.04 to 1.72, suggesting adequate within-group discrimination. Unfortunately, in the only RCT that used the HAQ, treatment allocation data were not made available to us, so between-group discrimination could not be ascertained, and the data on change in HAQ were not reported in the recent publication from that study[2]. MCID for HAQ-DI was estimated at 0.5 in the 2 replicate clinical trials of canakinumab. There was floor effect at followup visits (33 to 46%), but ceiling effect was minimal (0 to 17%).

0–10 NRS from WPAI:SHP

This single item used only in the Regeneron study was expressed at the baseline visit as "During the past 7 days prior to your gout attack, how much did your gout attack

affect your ability to do your regular daily activities, other than work at a job?" and the response is given on a 0 ("Gout attack had no effect on my daily activities") to 10 ("Gout attack prevented me from doing my daily activities"). This was administered as one of several items from the WPAI:SHP. At the followup visit at Day 7, the question was reworded slightly as "During the past 7 days, how much did your gout attack affect your ability to do your regular daily activities other than work at a job?" This item showed moderate correlation with pain measures (0.31, 0.39) and floor effects at the Day 7 visit (33.2%). We observed a trend toward between-group discrimination for this single item measured at Day 7 (Table 5).

DISCUSSION

The measurement properties for instruments in the core domains for acute gout studies were examined in 4 RCT and 1 cohort study. Overall, there appears to be sufficient evidence for construct validity and discriminative ability for 3 measures of pain (Likert, NRS, VAS). Floor and ceiling effects for pain measures suggested that either the scale for measuring pain needs to be somewhat broader or that the patients with severe pain of acute gout respond very well to treatment and that entry criteria for a particular level of pain limited the range of possible values at baseline. There is some variation in the floor and ceiling effects for the different pain measures across all studies, which is not unexpected given the differences in instrument and study setting.

The correlation of pain with disability was high when disability was measured by HAQ but modest when measured by a single item in the Regeneron study. It is possible that the single-item instrument used to measure disability was inadequate. The correlation between pain and joint swelling was consistently weak. This is not especially surprising because the 2 concepts are quite different and the measurement of joint swelling by a 4-point scale may have insufficient variability to give strong correlation coefficients.

There appears to be sufficient evidence for a 0–3 point Likert scale to possess construct validity and discriminative ability for measuring joint swelling and joint tenderness. There was some evidence for construct validity and within-group discriminative ability for HAQ as a measure of activity limitations, but it has yet to be shown that any measure of activity limitations can discriminate between groups allocated to different treatment.

Demonstration of the psychometric properties of the PGA of response to treatment is difficult. Construct validity tends to be assumed and was not measured by any other global patient-reported outcome in the data examined to enable a sensible comparison. Test-retest reliability could not be assessed. We did not demonstrate between-group discriminative ability in the only dataset available to us in which this could be examined, but the canakinumab study has been reported recently as showing a between-group difference in global response to treatment with a proportional odds regression OR of 2.19 (95% CI 1.6 to 3.1) at 72 h and 1.97 (95% CI 1.4 to 2.8) at 7 days[2]. We did not have treatment allocation data for that dataset, so were unable to reproduce this analysis.

The assessment of reliability and the associated estimates of SDD should be considered cautiously because acute gout is a highly dynamic condition with rapid changes in clinical status. It is possible that even in patients who self-identified as showing no response to treatment, their condition had improved. Therefore, the calculated ICC values especially during the first few days of acute gout are likely to be underestimates.

At OMERACT 11, these analyses were presented to participants and were useful as a basis for discussion and final conclusions regarding measurement properties of instruments for acute gout studies. This is outlined in a companion paper.

ACKNOWLEDGMENT

We gratefully acknowledge Pfizer Inc. (through an Investigator Initiated Grant), Merck Sharp & Dohme Corp., Novartis, Regeneron and Auckland University of Technology for making available the datasets for this study. We also acknowledge Mike Frecklington from Auckland University of Technology for the data collection in that study.

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. Schlesinger N, Alten RE, Bardin T, Schumacher HR, Bloch M, Gimona A, et al. Canakinumab for acute gouty arthritis in patients with limited treatment options: results from two randomised, multicentre, active-controlled, double-blind trials and their initial extensions. Ann Rheum Dis 2012;71:1839-48.
3. Rome K, Frecklington M, McNair P, Gow P, Dalbeth N. Foot pain, impairment, and disability in patients with acute gout flares: a prospective observational study. Arthritis Care Res 2012;64:384-8.
4. Dalbeth N, Zhong CS, Grainger R, Khanna D, Khanna PP, Singh JA, et al. Outcome measures in acute gout: a systematic literature review. J Rheumatol 2014;41:xxxx.
5. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.
6. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407-15.
7. Schumacher HR Jr, Boice JA, Daikh DI, Mukhopadhyay S, Malmstrom K, Ng J, et al. Randomised double blind trial of etoricoxib and indomethacin in treatment of acute gouty arthritis. BMJ 2002;324:1488-92.
8. Schumacher HR, Berger M, Li-Yu J, Perez-Ruiz F, Vargas RB, Li C. Efficacy and tolerability of celecoxib in the treatment of moderate to extreme pain associated with acute gouty arthritis: a randomized controlled trial [abstract]. Arthritis Rheum 2010;62 Suppl 10:S151.
9. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. Pharmacoeconomics 1993;4:353-65.