Published in final edited form as:
J Rheumatol. 2014 March ; 41(3): 574–580. doi:10.3899/jrheum.131245.

Application of the OMERACT filter to measures of core outcome domains in recent clinical studies of acute gout

William J Taylor, PhD, FRACP[1] **[Associate Professor]**, **David Redden, PhD**[2] **[Associate Professor]**, **Nicola Dalbeth, MD, FRACP**[3] **[Associate Professor]**, **H Ralph Schumacher, MD**[4] **[Professor]**, **N Lawrence Edwards, MD**[5] **[Professor]**, **Lee S Simon, MD**[6] **[Professor]**, **Markus R John, MD**[7] **[Global Program Medical Director]**, **Margaret N Essex, PharmD**[8] **[Senior Medical Director]**, **Douglas J Watson, PhD, FISPE**[9], **Robert Evans, Pharm D**[10] **[Senior Director]**, **Keith Rome, PhD**[11] **[Professor]**, and **Jasvinder A Singh, MBBS, MPH**[12] **[Associate Professor]**

[1]Department of Medicine, University of Otago, Wellington, New Zealand [2]Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama, USA [3]Department of Medicine, University of Auckland, New Zealand [4]University of Pennsylvania and VA Medical Center, Philadelphia PA, USA [5]Department of Medicine, University of Florida, Gainesville FL, USA [6]SDG LLC, Cambridge, MA, USA [7]Integrated Hospital Care Franchise - Immunology, Novartis Pharma AG, Basel, Switzerland [8]Pfizer Inc., New York, NY, USA [9]Epidemiology, Merck Sharp & Dohme Corp. Whitehouse Station NJ, USA [10]Senior Director,

Corresponding author and reprint requests: Associate Professor William Taylor, Department of Medicine, University of Otago, PO Box 7343, Wellington, New Zealand Phone +64 4 385 5541 x4801, Fax +64 4 389 5427, will.taylor@otago.ac.nz.

Disclosures:

1. WJT declares no conflicts of interest.
2. DR declares no conflicts of interest.
3. ND 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.
4. HRS has been a consultant for Regeneron, Novartis, Pfizer, Savient, Ardea, Metabolex & Biocryst. He has receieved a grant from Takeda.
5. NLE has received consultant fees from Novartis, Takeda Pharmaceutical, Savient Pharmaceutical, Ardea Biosciences, Regeneron Pharmaceuticals, Metabolex Pharmaceuticals and Biocryst Pharmaceuticals.
6. LSS has served on the Board of Directors for Savient Pharmaceuticals, and as a consultant for Takeda
7. MRJ is employed by Novartis and sometimes owns shares in the company.
8. MNE is employed by Pfizer and owns shares in the company.
9. DJW 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.
10. RE is employed by Regeneron and owns shares of stock.
11. KR declares no conflicts of interest.
12. JAS has received research grants from Takeda and Savient and consultant fees from Savient, Takeda, Ardea, Regeneron, Allergan, URL pharmaceuticals and Novartis. JAS 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.

Clinical Sciences, Regeneron Pharmaceuticals, Tarrytown NJ, USA [11]Health & Rehabilitation Research Insituitute and School of Podiatry, AUT University, Auckland, New Zealand [12]Birmingham Veterans Affairs Medical Center and University of Alabama at Birmingham, Birmingham, Alabama, USA

Abstract

Objective—To determine the extent to which instruments that measure core outcome domains in acute gout fulfil the OMERACT filter requirements of truth, discrimination and feasibility.

Methods—Patient-level data from four randomised controlled trials of agents designed to treat acute gout and one observational study of acute gout were analysed. 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 generalised estimating equations were assessed. Floor and ceiling effects were also assessed and MCID 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, 0 to 100 mm visual analogue scale). 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.

Conclusions—There is sufficient evidence to support measures of pain (using Likert, numeric rating scale or visual analogue scales), 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.

Key Indexing Terms

gout; outcome measures; psychometrics

Introduction

At OMERACT 11 (May 2012), the focus of the Gout Module was to obtain endorsement of specific instruments that measure each of the five 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.

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) since the results of the trial 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 randomised controlled trials and longitudinal observational studies. A systematic review of published trials of acute gout was performed separately and is reported in a companion article.

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 time-point. 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. 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 Generalised Estimating Equations with ordinal regression to maximise information available from multiple time-points (for example, pain was measured at several time-points).

Test-retest reliability was calculated using patient global assessment 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 day 2 and 5 and at day 5 and 8, in two separate estimations of reliability. In the celecoxib clinical trial we selected the low-dose celecoxib cases for the analysis over the first 12 hours and cases with poor or fair response at day 9 for the analysis over 9 days. The intra-class 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 ANOVA table from whence the ICC was calculated. Smallest detectable difference (SDD) was calculated as SEM x $\sqrt{2} \times 1.96$ (4). 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 (5).

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

(i) 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).

Effect size ranged from 1.20 to 2.84, demonstrating a large effect size over time (Table 5). 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.

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

Effect size ranged from 1.58 to 4.46, demonstrating a large effect size over time. VAS pain scale discriminated well between treatment groups as recently reported (6), 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%).

(iii) Numeric rating scale (NRS)—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, 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 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, The Likert scale showed between group as reported in (2) and within group differences (Table 6). Effect size 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.

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, 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 effect size for the Likert scale ranged 2.3 to 3.2, and the measure discriminated between treatment groups in one study that we analysed as well as a recently published analysis of duplicate RCTs 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 (39–58%) at baseline.

Patient global assessment

The patient global measure used inmost 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 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 AUT observational study that employed a 100 mm VAS asking participants to rate how well they were doing overall.

Patient global assessment (PGA) is usually often the external benchmark for all other outcome measures, including several described above. Therefore, it has face, content and construct validity almost by definition. Typically patient global assessments relate to assessment of current disease status; however, all but one study provided data for patient global assessment of response to treatment. Application of the OMERACT filter to a transition scale such as this is problematic. Reliability could not be determined, since 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 effect size 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 Health Assessment Questionnaire (HAQ-disability index or HAQ-II), and one study used a 0–10

NRS item from the Worker Productivity Activity Index: Specific Health Problem (WPAI:SHP) scale as a measure of activity limitations.

Health assessment questionnaire (HAQ)—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 HAQ, treatment allocation data was not made available to us, so between-group discrimination could not be ascertained and the data on change in HAQ was not reported in the recent publication from that study (2). MCID for HAQ-DI was estimated at 0.5 in the two replicate clinical trials of canakinumab. There was floor effect at follow-up visits (33 to 46%), but ceiling effect was minimal (0 to 17%).

0–10 Numeric rating scale (NRS) from WPAI:SHP—This single item used only in the Regeneron study was expressed at the baseline visit as "During the past seven 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 Worker Productivity and Activity Impairment Index (Specific Health Problem; WPAI:SHP). At the follow-up visit at day 7, the question was re-worded slightly as "During the past seven 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 towards 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 four RCT and one cohort study. Overall, there appears to be sufficient evidence for construct validity and discriminative ability for three measures of pain (Likert, NRS, VAS). Floor and ceiling effects for pain measures suggested that either the scale for measuring pain need 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 are 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 since the two 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 patient global assessment 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 data-set 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 odds ratio of 2.19 (95% CI 1.6 to 3.1) at 72 hours 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 since 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.

Acknowledgments

Source of support:

Grants: This work was supported with the resources and the use of facilities at the Birmingham VA Medical Center, Alabama, USA (JAS, DR) and University of Otago, New Zealand (WJT).

We gratefully acknowledge Pfizer Inc. (through an Investigator Initiated Grant), Merck Sharp & Dohme Corp., Novartis, Regeneron and AUT for making available the data-sets for this study. We also acknowledge Mike Frecklington from AUT University 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. [PubMed: 19820223]
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 Nov; 71(11):1839–48. [PubMed: 22586173]
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 & research. 2012; 64(3):384–8. [PubMed: 22006512]

4. 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(1):34–42. [PubMed: 17161752]

5. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989; 10(4):407–15. [PubMed: 2691207]

6. Schumacher HR Jr, Boice JA, Daikh DI, Mukhopadhyay S, Malmstrom K, Ng J, et al. Randomised double blind trial of etoricoxib and indometacin in treatment of acute gouty arthritis. BMJ. 2002; 324(7352):1488–92. [PubMed: 12077033]

7. Schumacher HR, Berger M, Li, Yu J, Perez, Ruiz F, et al. Efficacy and Tolerability of Celecoxib in the Treatment of Moderate to Extreme Pain Associated with Acute Gouty Arthritis: A Randomized Controlled Trial. Arthritis Rheum. 2010; 62(Suppl 10):S151.

8. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. Pharmacoeconomics. 1993; 4(5):353–65. [PubMed: 10146874]

Table 1

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

Source	N	Inclusion	Treatment groups	Publication
Merck, Sharp & Dohme	150	Onset within 48 hours, 1977 ARA criteria, at least moderate pain	Etoricoxib, indomethacin	(6)
Pfizer	402	Onset within 48 hours, 1977 ARA criteria, at least moderate pain	Celecoxib 50, 400/200, 800/400, indomethacin	(7)
Regeneron	225	Onset within 48 hours, 1977 ARA criteria, at least moderate pain	Indomethacin, rilonacept and indomethacin, rilonacept	Not published (NCT00855920)
Novartis (two replicate studies)	424	Onset of acute flare within 5 days, 1977 ARA criteria, at least 3 flares within prior 12 months, 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 is now published; the dataset provided to investigators was a 90% random subsample of the main study dataset (n=456)

Table 2

Instruments available for each data source

Source	Pain	Disability	Joint swelling/tenderness	Patient global
Merck, Sharp & Dohme	Likert 0–4	No measure available	Likert 0–3 [*]	Response to treatment (Likert 0–4 point)
Pfizer	Likert 0–4	No measure available	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)	No measure [†] available	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

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) (8).

[*] Index joint assessed by a physician;

[†] a 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

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
Auckland University of Technology	Pain VAS	NA	NA	NA	0.66	0.73
	Joint tenderness			NA	NA	NA
	Joint swelling				NA	NA
	Activity limitations†					0.73

NA: measure not available in the dataset

*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.

[‡] Change in each measures were correlated with patient global, since the patient global represented perception of change (except for the AUT dataset)

Table 4

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

Measure	Source	Floor (%)		Ceiling (%)	
		Baseline	Final[*]	Baseline	Final[*]
	Novartis	NA	2.1	NA	39.1
	AUT [†]	0	0	5	0

NA measure not available

[*] Refers to Day 5 unless mentioned specifically

[†] Final value at 6 to 8 weeks

[‡] Measured by HAQ-II in AUT, HAQ-DI in Novartis, WPAI 0–10 in Regeneron

[¶] Final value at Day 9 for Pfizer and Novartis

Table 5

Indices of discrimination

Measure	Source	Within-group (pooled)		Between group	
		Effect Size	[†] GEE (Wald χ2)	KW-statistic	[†] GEE (Wald χ2)
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=0.001
	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

Measure	Source	Within-group (pooled)		Between group	
		Effect Size	[†] GEE (Wald χ2)	KW-statistic	[†] GEE (Wald χ2)
	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

NA: not available or not applicable

[*] MSD – Merck, Sharp & Dohme Corp. Treatment allocation not available or not relevant therefore between-group discrimination was not assessable

[†] Repeated measures Generalised Estimating Equations with ordinal regression performed in Pfizer and Novartis datasets

[‡] No baseline measure since it assessed response to treatment

[§] Not measured at multiple time-points.

[¶] PGA measured with 100 mm VAS for current status (all other studies used global response to treatment)

Table 6

Indices of test-retest reliability, smallest detectable difference and minimal important difference

			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 hours	0.81	0.39	1.08	--
		Between 2 and 8 hours	0.72	0.50	1.39	--
		Between 2 and 12 hours	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 post dose	0.35	0.85	2.36	1.0
		24 hours post dose to 7 days	0.55	0.64	1.77	--
		48 hours post dose to 7 days	0.71	0.49	1.36	--
Pain (VAS)	Novartis	Baseline to 7 days post dose	0.35	3.66	10.15	19
		24 hours post dose to 7 days	0.57	2.93	8.12	--
		48 hours post dose 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 Post	0.0**	1.06	2.93	1.0
		24 hours post dose to 7 days	0.50	0.54	1.51	--
		48 hours post dose to 7 days	0.49	0.54	1.50	--
		72 hours post dose 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

			ICC	SEM	SDD	MID
		24 hours post dose to 7 days	0.44	0.65	1.80	--
		48 hours post dose to 7 days	0.44	0.65	1.79	--
		72 hours post dose 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 hours’

[**] Statistical software indicated that estimation of a negative variance parameter was attempted