

Please cite the Published Version

Parkes, M, Callaghan, Michael, Tive, L, Lunt, M and Felson, D (2018) Responsiveness of Single versus Composite Measures of Pain in Knee Osteoarthritis. *Journal of Rheumatology*, 45 (9). pp. 1308-1315. ISSN 0315-162X

DOI: <https://doi.org/10.3899/jrheum.170928>

Publisher: Journal of Rheumatology

Version: Accepted Version

Downloaded from: <https://e-space.mmu.ac.uk/619761/>

Usage rights: © In Copyright

Additional Information: This is an Author Accepted Manuscript provided by Journal of Rheumatology of a paper accepted for publication in Journal of Rheumatology.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

Responsiveness of Single versus Composite Measures of Pain in Knee

Osteoarthritis

Matthew J Parkes^{1,2}, Michael J Callaghan^{1,2,3,4}, Leslie Tive⁵, Mark Lunt^{1,2}, David T Felson^{1,2,6}

¹Arthritis Research UK Centre for Epidemiology, Centre for Musculoskeletal Research, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK.

²NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK.

³Faculty of Health, Psychology, and Social Care, Department of Health Professions, Manchester Metropolitan University, Manchester, UK

⁴Central Manchester University Hospitals NHS Foundation Trust, Manchester, UK

⁵Pfizer Inc., New York, NY, USA

⁶Clinical Epidemiology Unit, Boston University School of Medicine, Boston, MA, USA.

Address correspondence:

Matthew Parkes, Research Statistician, Research in Osteoarthritis Manchester (ROAM), Arthritis Research UK Centre for Epidemiology, Centre for Musculoskeletal Research, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UNITED KINGDOM, M13 9PT (matthew.parkes@manchester.ac.uk)

Word counts:

Abstract: 243

Main Text: 3,495

ORCIDs of authors, where available:

Matthew Parkes, ORCID: 0000-0002-1574-9933

Michael Callaghan, ORCID: 0000-0003-3540-2838

David Felson, ORCID: 0000-0002-2668-2447

Mark Lunt, ORCID: 0000-0002-2391-5575

Data Sharing and Integrity

The corresponding author (MJP) had full access to the data presented in this study, and takes responsibility for the integrity of the data, and the accuracy of the data analysis.

Abstract

Objective

In rheumatoid arthritis, composite outcomes constructed from a combination of outcome measures are widely used to enhance responsiveness (sensitivity to change) and comprehensively summarize response. WOMAC pain is the primary outcome measure in many osteoarthritis (OA) trials. Information from other outcomes, such as rescue medication use, and other WOMAC subscales, could be added to create composite outcomes, but the sensitivity of such a composite has not been tested.

Method

We used data from a completed trial of Tanezumab for knee OA (NCT00733902). The WOMAC questionnaire and rescue medication use were measured at multiple time points, up to 16 weeks. Pain and rescue medication outcomes were standardised and combined into 3 composite outcomes via principal components analysis to produce one score (composite outcome) and their responsiveness was compared to WOMAC pain, the standard. We pooled all treatment doses of Tanezumab into one 'treatment' group, for simplicity, and compared this to the control group (placebo).

Results

The composite outcomes showed modestly but not statistically significantly greater responsiveness when compared to WOMAC pain alone. Adding information on rescue medication to the composite improved responsiveness. While improvements in sensitivity were modest, the required sample sizes for trials using composites was 20-40% less than trials using WOMAC pain alone

Conclusion

Combining information from related, but distinct, outcomes considered relevant to a particular treatment improved responsiveness, could reduce sample size requirements in OA trials and might offer a way to better detect treatment efficacy in OA trials.

Keywords: Osteoarthritis; Outcomes; Pain; Sensitivity to Change; Responsiveness.

Running headline: Sensitivity to Change of Pain Outcomes

1 Significance & Innovations:

- 2 • This study attempts to evaluate meaningful ways of combining single outcomes in a way that
3 improves responsiveness, gaining more power to detect treatment effects without collecting more
4 data.
- 5 • This can improve efficiency in future clinical trials, as it helps improve detection of smaller
6 treatment effects with fewer participants.
- 7 • Combining outcomes appears to produce composites with greater sensitivity to change than
8 constituent parts.

9

10

11 Introduction

12 Clinical trialists have a tendency to measure many outcomes. Several of these outcomes (often
13 deliberately) cover overlapping 'domains', attempting to ensure that the 'signal' of a true change in an
14 outcome following an intervention is captured. Pain is a good example; researchers will often use a
15 variety of similar pain-related outcomes in interventional trials.

16 Pain is a complex, multidimensional measure(1,2), and generating just one scale or item that adequately
17 captures most, if not all, aspects of pain is challenging. Furthermore, as pain is strongly related to
18 functional limitation(3), the most appropriate pain outcome might cover aspects of both pain and
19 function. The optimal clinical trial pain outcome(s) should additionally be sensitive to change following
20 an intervention, by which we mean the outcome's ability to detect a change, often also termed an
21 outcome's responsiveness(4), discriminating well between a true signal, (treatment effect) and noise
22 (random variation).

23 Composite outcomes are a way of combining (often related) indices or scores to form one overall
24 outcome. This approach, which has been used in many disease areas, including osteoarthritis(5),
25 rheumatoid arthritis(6–8) and asthma(9), may improve the capture of a domain more completely as it
26 takes account of more information than one outcome alone. Pain measurement appears particularly
27 suited to this approach, given its complexity. Combining information from several different domains may
28 improve a composite's ability to detect a change when one truly occurs, and therefore 'responsiveness'
29 may also be improved.

30 ConstructingComposites:AvailableMethods

31 There are several methods for combining outcomes into composites. Some of these facilitate domain
32 coverage; others increase responsiveness. Ideally, the method used should improve both. The simplest

33 method of combining two or more outcomes is through summing or averaging them(5). This method
34 assumes that the constituent outcomes have equal weighting in the composite, and that units from the
35 constituent outcomes are comparable and exchangeable.

36 A second method of combining multiple outcomes is through the use of weights to assign 'importance' of
37 constituent outcomes. The composite is produced by multiplying each constituent outcome by its weight,
38 and then summing these scores. An example of this is the DAS28(6,7). Weights can be derived from a
39 variety of sources, including statistical modelling (as with the DAS), but also from group consultation, for
40 example via a Delphi exercise (10–14).

41 Another data-driven approach uses principal components analysis, a data-reduction method which
42 inherently concentrates as much variance from constituent outcomes into as few factors as possible. This
43 method may produce a composite outcome which more completely captures the variance from an
44 underlying multidimensional process, such as pain.

45 Theoretically, combining several outcomes purporting to measure aspects of pain and its consequences,
46 such as function loss and rescue medication use, should increase domain coverage (as each outcome
47 contributes some information about the pain signal), and therefore responsiveness. Since all of the
48 contributing outcomes should measure that same latent factor (pain), the analysis model used should
49 assume a priori a one factor solution, rather than generating multiple outcomes. This way, we can
50 combine all outcomes related to pain into one composite outcome, which will hopefully show maximal
51 responsiveness in pain.

52 This study sought to combine several pain outcomes using principal components analysis, taken from a
53 large completed clinical trial of a treatment that reduced pain, and compare the relative responsiveness
54 of these composites to the uncombined WOMAC pain subscale score alone, to establish whether the

55 inclusion of additional pain information improves responsiveness following administration of an
56 intervention.

57 Assessment of responsiveness is optimal in certain trial designs. The ideal trial should contain a
58 treatment arm with an intervention which is known to truly change the construct of interest (pain, for
59 example); a control arm which is known to truly not change the construct of interest, and at least two
60 (ideally more) time points in both arms, over which the change in each outcome is assessed. The trial we
61 selected had these features. If the outcome of interest is not changed during the study, then it is not
62 possible to assess responsiveness.

63

64 Methods

65 The data used in this study were from a large completed clinical trial of Tanezumab in participants with
66 knee osteoarthritis (NCT00733902). This trial was a 32 week four-arm parallel-group phase III trial,
67 comparing 3 doses of tanezumab (2.5, 5, or 10 mg) against placebo. Participants were observed at
68 baseline, 2, 4, 8, 12, 16, 24, and 32 weeks; we used data from the 2 week visit to the 16 week visit, as data
69 for rescue medication use were collected only at these visits. For simplicity, we pooled all tanezumab
70 doses (2.5mg, 5mg, and 10mg) together into one 'treatment' group and compared this to the placebo
71 group. Further details regarding the trial's design as well as data on unstandardised outcome scores in
72 have been published previously(15). This study is a reanalysis of completed clinical trial data, and is
73 exempt from ethical review under the NHS Health Research Authority Guidelines.

74 VariableDefinitions

75 Single Outcomes:

76 We used the following pain-related outcomes featured in NCT00733902: WOMAC pain, stiffness, and
77 function subscales; and number of rescue medication pills taken per week.

78 Composite Outcomes:

79 Including information from at least two, and up to four outcomes in each composite gives 11 possible
80 combinations available from which composites could be generated. We generated a total of three
81 composite outcomes which were felt to be the most meaningful and pragmatic of the 11 possible
82 combinations. Composite one consisted of the WOMAC pain subscale plus rescue medication. Composite
83 two consisted of all three WOMAC subscales (pain, stiffness, and function). Composite three consisted of
84 all three WOMAC subscales, plus the rescue medication outcome. Composite outcomes were derived by

85 including the selected combination of variables in a principal components analysis (PCA), which assumed
86 a one factor solution. We opted for PCA, given its propensity to maximise the amount of variance
87 captured in the first (and in this case, only) derived component. We assumed that all included outcomes
88 measured different aspects of one latent (multidimensional) pain variable, and forcing a one component
89 solution therefore ensured that this variable was derived. Angst et al. (2005) found that unrestricted
90 factor analysis of individual WOMAC items established new factors which drew from both the pain and
91 function subscales, and merged them together(16), supporting this idea. It also simplifies the analysis, as
92 it creates only one composite outcome, rather than allowing many composite factors to be generated in
93 each PCA model. We constructed three PCA models, each generating one of three composite outcomes.
94 Rotation of the factor solution (of any type, varimax, promax, or other) was not indicated in our
95 approach, as a one factor solution has only one possible orientation.

96 We pooled together data from all study visits in the analysis models (rather than using data from baseline
97 only, for example) assuming that it was best to include the maximum available number of observations in
98 the PCA models.

99 AnalysisApproach

100 All composite outcome measures were compared to the WOMAC pain, assuming this as the standard.

101 All of the single outcomes (WOMAC pain subscale score, WOMAC function subscale score, WOMAC
102 stiffness subscale score, and number of rescue medication pills taken) were standardised prior to
103 inclusion in the factor analysis models, and the composites (composites one, two, and three detailed
104 above) were also standardised. Having all variables standardised (as z-scores) allows direct comparison

105 of outcomes with different units.

106 We used a random-effects panel linear regression (via SAS's PROC MIXED) to assess change in the
107 standardised outcome score over time, with outcome type, study visit, and treatment group (either
108 tanezumab or placebo) and all possible interactions, as predictor variables. Constructing the data in 'long
109 format', and using outcome type as a categorical dummy-coded variable allows direct testing for
110 differences in responsiveness between all outcomes in one statistical model (further detail on model
111 terms in online appendix). SAS's PROC MIXED command uses a likelihood-based approach, treating
112 missing observations as missing-at-random.

113 We used linear combinations of coefficients from the regression model (using SAS PROC ESTIMATE) to
114 produce the difference in standardised change between the WOMAC pain subscale and each composite
115 outcome, at each study time point. This formally tests whether the outcomes differed from the WOMAC
116 pain subscale in terms of responsiveness at each of the five time points in the study.

117 Statistical analysis used SAS® software version 9.3; (SAS Institute Inc., Cary, NC, USA). A nominal alpha
118 level of 0.05 was used for all confidence intervals.

119

120 Results

121

122 StudySampleDemographics

123 At baseline, the placebo group (N=172) comprised 119 females (69.2%), with a mean age of 62.2 years,
124 Kellgren-Lawrence grades 2, 3 and 4 of 39.5%, 47.7%, and 12.8% respectively, mean WOMAC pain
125 subscale score (0-10) of 7.1, and mean WOMAC function subscale score (0-10) of 6.6. The pooled
126 tanezumab group (N=518) at baseline had 301 females (58.1%), with a mean age of 61.4 years, Kellgren-
127 Lawrence grades 2, 3 and 4 of 38.4%, 46.3%, and 14.5% respectively, mean WOMAC pain subscale score
128 (0-10) of 7.1, and mean WOMAC function subscale score (0-10) of 6.8.

129

130 Ten participants had missing observations for all outcomes at the time points of interest, giving a total
131 sample size for this analysis of 680, compared with the original trial sample size of 690, with 509 in the
132 pooled tanezumab group, and 171 in the placebo group. Data for the 680 included patients could have
133 been collected on 7 outcomes, at 5 time points, giving a total of 23,800 possible observations. Of these,
134 20,597 were actual observed data points, with 3,203 observations missing (13.5%).

135

136 PrincipalComponentsAnalysisResults

137 The PCA process generated composites with component loadings shown in table 1. WOMAC pain and
138 stiffness subscales consistently had the greatest, and indeed equal, loading, closely followed by the

139 WOMAC function subscale. When all 3 WOMAC subscale variables were included in the PCA model (in
140 composite 3), the rescue medication's loading dropped considerably.

141

142 CompositeOutcomePerformance

143 All composites showed responsiveness greater than at least some of their constituent outcomes on their
144 own, and this difference was consistent across multiple time points (figure 1). Composite one showed

145 consistently greater responsiveness than the WOMAC pain subscale alone. The remaining two composites
146 displayed responsiveness greater than all other constituent outcomes, except the WOMAC stiffness
147 subscale. None of the single or composite outcomes showed consistently statistically significantly better
148 responsiveness than that observed in the WOMAC pain subscale at the chosen alpha level (table 3).

149

150 We next examined the impact of the observed differences in responsiveness on sample size requirements
151 for a hypothetical new trial featuring the same design (table 2). For example, the WOMAC pain subscale
152 between-groups standardised change at four weeks was a difference of -0.37. A hypothetical new trial of
153 identical design observing this between-group difference for the WOMAC pain outcome would require
154 236 participants (118 per group) to achieve 80% power with a two sided 5% type-I error rate. In
155 contrast, using composite 1 (i.e. including information on rescue medication as well as the WOMAC pain
156 subscale score) as the primary outcome which had an observed difference at four weeks of -0.41, the
157 same trial would need 190 participants (95 per group) to achieve 80% power with this difference - a
158 saving of 46 participants. When the observed differences between treatments is smaller, the reduction in
159 sample size was more extreme: the WOMAC pain difference at 16 weeks (-0.26) would require 476
160 participants for 80% power in a hypothetical new trial, compared to only 364 participants when using

161

162

co
mp
osi
te 1
(us
ing
the
obs
erv
ed
diff
ere
nce
of
0.2
9),
a
sav
ing
of
11
2
par
tici
pa
nts.

163 Discussion

164 We found that composite outcomes generally had moderately greater responsiveness in a large OA trial
165 than WOMAC pain – the usual standard outcome of these trials. That suggests if one of these composite
166 outcomes were used as the primary outcome in an OA trial, fewer subjects would be needed to
167 demonstrate treatment efficacy.

168

169 The improvements in responsiveness did not meet the criteria for a statistically significant difference, but
170 perhaps a more salient measure of their import was to determine what effect using these outcomes had
171 on the sample size needed to be likely to show statistically significant effects of treatment vs. placebo. We
172 found that the reduction in sample size was substantial, ranging from roughly 20 to 40%. Thus,
173 composites could substantially diminish the sample sizes needed in an osteoarthritis trial whose main
174 outcome is pain.

175

176 Eigenvalues from the three composite models all were much greater than the 1.0 cut-off typically used to
177 select retained factors in a PCA model(17), and a large proportion of the variance in the outcomes was
178 captured by the first component in the PCA model, as anticipated (table 1). The second factor listed in the
179 model output (which was not extracted in this analysis) in all cases had an eigenvalue much less than 1,

180 lending support to the idea that the selected correlated outcomes are well captured in in one 'pain'
181 component.

182

183 Rescue medication use, whilst contributing to the 'pain' component the least (table 1), appeared to still
184 improve responsiveness: composites including this outcome - composites one (WOMAC pain plus rescue
185 medication use) and three (WOMAC pain, stiffness, and function, plus rescue medication) - showed slight
186 improvements in responsiveness compared with composite 2, which excluded rescue medication.

187

188 Aside from the methods used to combine outcomes, the method chosen to assess responsiveness is also
189 important(18,19). Several methods are commonly cited to quantify responsiveness: the standardised
190 response mean (SRM)(20), the effect size (ES)(18), either Glass' Δ (21) or Cohen's d (22) (depending on
191 the standard deviation used), or Guyatt's responsiveness index (GRI) (23). All of these methods have two
192 important limitations. First, all methods calculate responsiveness over two time points, and cannot easily
193 be generalised to a study which has three or more time points. This prevents assessment of how
194 responsiveness may fluctuate over time, and limits the definition of responsiveness only to the
195 magnitude of change relative to its variance, rather than the speed of response. Second, these methods do
196 not directly assess statistical inference; a; differences in responsiveness coefficients are assessed
197 descriptively only. Methods have been proposed (modified jackknife procedure (5,24,25), bootstrapping
198 (26)) to address this issue, but other methods which directly perform statistical inference as part of the
199 method generating the coefficient are desirable.

200 Our approach made use of z-scores (standard scores) (27). Converting each outcome's absolute score to
201 a z-score allows direct comparison of change in an outcome at different time points, thereby allowing
202 direct assessment of change over time, and direct comparison between different outcomes. This

203 methodology has been used previously to compare non-composite outcomes(28).

204

205 The PCA approach assumes that an intervention will alter several related aspects of a common construct,

206 therefore combining all the multidimensional aspects of pain together to form one outcome should

207 increase responsiveness. However, if one aspect of pain is changed alone, then the inclusion of other

208 aspects of pain which do not change may decrease the sensitivity of the composite. Our finding that the

209 WOMAC stiffness subscale was the most sensitive outcome may fit this explanation: It may be, at least in
210 this trial, that the WOMAC stiffness subscale was the closest correlate to the actual latent pain factor
211 altered by the treatment, hence the greatest responsiveness, and inclusion of other subscales or rescue
212 medication eroded it. Our finding may be limited to tanezumab alone – as the agent’s anti-nerve growth
213 factor effect may have a greater impact on the stiffness sensation than other pain subscales(15,29).

214

215 Freemantle et al. (2003) provide a comprehensive discussion on the use of composite outcomes in
216 clinical trials(30), highlighting how composite outcomes can obfuscate changes in constituent outcomes.
217 This is particularly problematic when outcomes are unrelated (for example, a composite which combines
218 cardiovascular events and mortality), although they note the statistical advantages (increased power and
219 sensitivity) that arise through the construction of composites(30,31). This discussion highlights how both
220 the outcomes used in the composite, and the method by which they are combined, are important. The
221 present study combined the three WOMAC subscales, pain, stiffness, and function, into one composite
222 outcome. We assumed that these three subscales were all aspects of the same construct (pain). The PCA
223 (table 1) produced extremely high factor loading in all three subscales, supporting this notion - at least in
224 this trial. In contrast, if pain and function were discrete constructs, then the PCA should fail, with either
225 pain or function alone loading on the latent factor. Both Ryser et al. (1999), and Angst et al. (2005) found

226 close association between pain and function WOMAC subscales, partly supporting this finding(16,32). In
227 addition, an item overlap analysis on the WOMAC pain and function subscales by Stratford et al. (33)
228 found significant item redundancy between the pain and function subscales, and a further factor analysis
229 on the WOMAC items found clustering of items not by subscale, but by activity(34), suggesting that the
230 WOMAC subscales are not distinct.

231

232 We surmised that responsiveness in the outcomes may differ over time, as well as in magnitude. In this
233 study, all outcomes appeared to have responded at the same time point, and retained their relative
234 positions consistently over time (none of the outcome's trajectories crossed over each other over time,
235 figure 1).

236

237 There are limitations to this analysis. We observed only very few statistically significant differences
238 between outcomes. The trial was designed to observe a difference in the primary outcome between
239 treatment groups (a relatively large difference), and was not designed to compare treatment differences
240 between outcomes (much smaller differences). Therefore even the large sample size in the trial provides
241 relatively low power to detect differences between outcomes. Ideally in future, this analysis would be
242 designed in to the trial prior to commencement, with appropriate sample size and power. We also
243 allowed many interaction effects, which increased model-to-data fit at the expense of statistical power.
244 We have assumed in this analysis that the covariate structure of the pain outcomes, and the relationship
245 between the outcomes and the latent (unobserved) pain outcome are consistent between studies, and
246 therefore generalisable across other studies. This is a relatively strong assumption, requiring validation
247 in other datasets to allow wider generalisation to other trials with confidence.

248

249 While the aim of this approach was to include additional information on pain from rescue medication
250 data, this outcome may not be optimal. Rescue medication is a challenging variable to collect data on
251 accurately, and therefore the likelihood is that measurement error of this variable is high. This may
252 provide an explanation for why the improvement in sensitivity of composites including rescue
253 medication are small.

254

255 Even though the between-outcome differences were not statistically significant, even a small
256 improvement in responsiveness can impact upon sample size calculations (table 3). This produces gains
257 in efficiency without collecting any novel data simply by reanalysing the data using a method which
258 produces a more sensitive, and therefore efficient, outcome. We could have included further assessment
259 of other composites made from different combinations of the 11 possible from the four single outcomes
260 used, for example one using WOMAC pain plus WOMAC stiffness. We opted to create the three
261 composites which would have the most pragmatic impact on outcome inclusion/exclusion when
262 designing a trial. The alternative, generating all 11 possible combinations and comparing them head to
263 head, would further reduce the statistical power to discern differences between composite outcomes.

264

265 The PCA approach to generating a composite outcome by its nature produces a unitless score. While the
266 generated score may have increased responsiveness compared to one of the constituent outcomes, it is
267 more difficult to ascertain the clinical importance of the observed effect, in comparison to another
268 outcome with meaningful units and an agreed minimally clinical importance difference (MCID). A
269 downside PCA composites is the absence of known values of MCID, but this could be established if a
270 specific composite were widely used.

271

272 The choice of primary and secondary outcomes in this trial limited the choice of outcomes available to
273 combine into a composite. Ideally, we would have preferred to use a trial featuring a wider range of pain
274 outcomes, particularly the more recent KOOS(35) and ICOAP(36) questionnaires; however a dataset
275 using these outcomes amongst others, and featuring the other requirements was not available.

276

277

278 The present findings are similar to our previous paper, which used data from two other completed
279 clinical trials of non-drug interventions(28). In both of these trials, the WOMAC stiffness subscale also
280 showed an increased, but non-statistically-significant, degree of responsiveness compared to the other
281 two WOMAC subscales. Angst et al. (2001; 2008), in contrast, found the WOMAC pain subscale to be the
282 most sensitive outcome to change(5,24), however these studies did not examine rescue medication, and
283 used a two-time point approach only. Further, the two studies previously analysed were both
284 prospective cohort studies lacking a control group. Thus, optimizing the detection of treatment effect
285 over placebo was not possible in the two Angst et al. analyses.

286

287 In summary, we investigated whether collapsing several measures of a multidimensional construct into
288 one composite outcome through the use of PCA could help improve responsiveness following an
289 intervention. Adding rescue medication alongside other elements of the WOMAC showed improved
290 responsiveness, greater than the constituent outcomes.

291

292

293 Acknowledgements

294 We would like to acknowledge the contributions of Pfizer in allowing our team access to the completed trial datasets, and their
295 support in using their trial analysis platform, specifically Pamela Singletary, Daireen Garcia, Glenn Pixton, and Michael Smith

296 for their help and support. We would also like to acknowledge the contributions of the ROAM team to this project. The ROAM
297 group is supported by the Manchester Academic Health Sciences Centre (MAHSC). Arthritis Research UK also continues to
298 support the Centre for Epidemiology (grant number 20380). This report includes independent research supported by (or
299 funded by) the National Institute for Health Research Biomedical Research Unit Funding Scheme. The views expressed in this
300 publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the
301 Department of Health. The funding agencies had no role in any of the following: design and conduct of the study; collection,
302 management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript; and decision to
303 submit the manuscript for publication. Prof. Felson is supported by NIH AR4778. The authors would also like to acknowledge

304 the assistance given by Séamus Byers, Contracts, IT Services and the use of the Computational Shared Facility at The
305 University of Manchester.

306

307

308 Contributions of authors

309 DTF initially proposed the study.

310 Wrote the manuscript: MJP, DTF

311 Analysis and interpretation of data: MJP, ML

312 Reviewed drafts of the paper: MJP, MJC, ML, LT, DTF

313

314

315 Competing Interest Statement (Financial Support)

316 MJP, MJC, ML, and DTF receive salary support from the National Institute for Health Research, as part of the Manchester

317 Musculoskeletal NIHR Biomedical Research Unit Grant.

318 LT owns stock or stock options in Pfizer.

319

320 References

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

1. Williams AC de C, Craig KD. Updating the definition of pain. *Pain* [Internet] 2016;157:2420-3. Available from: <http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00006396-201611000-00006>
2. Mease PJ, Hanna S, Frakes EP, Altman RD. Pain mechanisms in osteoarthritis: Understanding the role of central pain and current approaches to its treatment. *J Rheumatol* 2011;38:1546-51.
3. Neogi T. The epidemiology and impact of pain in osteoarthritis. *Osteoarthr Cartil* 2013;21:1145-53.
4. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain* 2005. p. 9-19.
5. Angst F, Aeschlimann A, Steiner W, Stucki G. Responsiveness of the WOMAC osteoarthritis index as compared with the SF-36 in patients with osteoarthritis of the legs undergoing a comprehensive rehabilitation intervention. *Ann Rheum Dis* [Internet] 2001;60:834-40. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11502609> \n <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1753825>
6. van der Heijde DM, van 't Hof MA, van Riel PL, Theunisse LA, Lubberts EW, van Leeuwen MA, et al. Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. *Ann Rheum Dis* 1990;49:916-20.
7. van der Heijde DM, van't Hof MA, van Riel PL, van Leeuwen MA, van Rijswijk MH, van de Putte LB. Validity of single variables and composite indices for measuring disease activity in rheumatoid arthritis. *Ann Rheum Dis* [Internet] 1992;51:177-81. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/1550400> \n <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1005654>
8. Ibrahim F, Tom BDM, Scott DL, Prevost AT. A systematic review of randomised controlled trials in rheumatoid arthritis: the reporting and handling of missing data in composite outcomes. *Trials* [Internet] *Trials*; 2016;17:272. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27255212> \n <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4890523>
9. Cloutier MM, Schatz M, Castro M, Clark N, Kelly HW, Mangione-Smith R, et al. Asthma outcomes: Composite scores of asthma control. *J Allergy Clin Immunol* [Internet] Elsevier Ltd; 2012;129:S24-33. Available from: <http://dx.doi.org/10.1016/j.jaci.2011.12.980>
10. Gossec L, Paternotte S, Aanerud GJ, Balanescu a, Boumpas DT, Carmona L, et al. Finalisation and validation of the rheumatoid arthritis impact of disease score, a patient-derived composite measure of impact of rheumatoid arthritis: a EULAR initiative. *Ann Rheum Dis* 2011;70:935-42.

- 353 11. Dechartres A, Albaladejo P, Mantz J, Samama CM, Collet JP, Steg PG, et al. Delphi-consensus weights for ischemic
354 and bleeding events to be included in a composite outcome for RCTs in thrombosis prevention. PLoS One
355 2011;6:10-2.
- 356 12. Rogozinska E, D'Amico MI, Khan KS, Cecatti JG, Teede H, Yeo S, et al. Development of composite outcomes for
357 individual patient data (IPD) meta-analysis on the effects of diet and lifestyle in pregnancy: A Delphi survey. BJOG
358 An Int J Obstet Gynaecol 2016;123:190-8.
- 359 13. Monchaud C, Marin B, Estenne M, Preux P-M, Marquet P. Consensus conference on a composite endpoint for
360 clinical trials on immunosuppressive drugs in lung transplantation. Transplantation 2014;98:1331-8.

- 361 14. Tong BC, Huber JC, Ascheim DD, Puskas JD, Jr BF, Blackstone EH, et al. Evidence for the Heart Team.
362 2013;94:1908-13.
- 363 15. Brown MT, Murphy FT, Radin DM, Davignon I, Smith MD, West CR. Tanezumab reduces osteoarthritic knee pain:
364 results of a randomized, double-blind, placebo-controlled phase III trial. J Pain [Internet] Elsevier Ltd;
365 2012;13:790-8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22784777>
- 366 16. Angst F, Ewert T, Lehmann S, Aeschlimann A, Stucki G. The factor subdimensions of the Western Ontario and
367 McMaster Universities Osteoarthritis Index (WOMAC) help to specify hip and knee osteoarthritis. A prospective
368 evaluation and validation study. J Rheumatol 2005;32:1324-30.
- 369 17. Kaiser HF. The application of electronic computers to factor analysis. Educ Psychol Meas [Internet] 1960;20:141-
370 51. Available from: <http://www.garfield.library.upenn.edu/classics1986/A1986E107600001.pdf>
- 371 18. Stratford PW, Riddle DL. Assessing sensitivity to change: choosing the appropriate change coefficient. Health Qual
372 Life Outcomes 2005;3:23.
- 373 19. Norman GR, Wyrwich KW, Patrick DL. The mathematical relationship among different forms of responsiveness
374 coefficients. Qual Life Res 2007;16:815-22.
- 375 20. Liang MH, Fossel a H, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med
376 Care 1990;28:632-42.
- 377 21. Hedges L V., Olkin I. Statistical methods for meta-analysis. Phytochemistry [Internet] 1985;72:369. Available from:
378 http://oldjll.sustainabilityforhealth.org/trial_records/20th_Century/1980s/hedges/hedges-kp.pdf
- 379 22. Cohen J. Statistical power analysis for the behavioral sciences [Internet]. Statistical Power Analysis for the
380 Behavioral Sciences 1988. p. 567. Available from: <http://books.google.com/books?id=T10N2IRA09oC&pgis=1>
- 381 23. Guyatt G, Walter S, Norman G. Measuring Change Over Time- Aseessing the Usefulness of Evaluative
382 Instruments. J Chronic Dis [Internet] 1987;40:171-8. Available from: <Go to ISI>://WOS:A1987G268000010
- 383 24. Angst F, Verra ML, Lehmann S, Aeschlimann A. Responsiveness of five condition-specific and generic outcome
384 assessment instruments for chronic pain. BMC Med Res Methodol [Internet] 2008;8:26. Available from:
385 <http://www.ncbi.nlm.nih.gov/pubmed/18439285>\n<http://www.biomedcentral.com/content/pdf/1471-2288-8-26.pdf>
386
- 387 25. Angst F, Goldhahn J, Drerup S, Aeschlimann A, Schwyzer H-K, Simmen BR. Responsiveness of six outcome
388 assessment instruments in total shoulder arthroplasty. Arthritis Rheum [Internet] 2008;59:391-8. Available from:
389 <http://www.ncbi.nlm.nih.gov/pubmed/18311752>

- 2 . Spadoni GF, Stratford PW, Solomon PE, Wishart LR. The Evaluation of Change in Pain Intensity: A Comparison of
6 the P4 and Single-Item Numeric Pain Rating Scales. *J Orthop Sport Phys Ther* 2004;34:187-93.
- 392 27. Kirkwood BB, Sterne J. *Essential medical statistics* [Internet]. Malden, MA: Blackwell Science 2003. 1-512 p.
393 Available from: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Essential+Medical+Statistics#0>
- 394 28. Parkes MJ, Callaghan MJ, O'Neill TW, Forsythe LM, Lunt M, Felson DT. Sensitivity to Change of Patient-Preference
395 Measures for Pain in Patients With Knee Osteoarthritis: Data From Two Trials. *Arthritis Care Res* 2016;68:1224-
396 31.
- 397 29. Lane NE, Schnitzer TJ, Birbara C a, Mokhtarani M, Shelton DL, Smith MD, et al. Tanezumab for the treatment of
398 pain from osteoarthritis of the knee. *N Engl J Med* 2010;363:1521-31.

- 399 30. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater
400 precision but with greater uncertainty? JAMA 2003;289:2554-9.
- 401 31. Freemantle N, Calvert MJ. Interpreting composite outcomes in trials. Br Med J 2010;341:c3529.
- 402 32. Ryser L, Wright BD, Aeschlimann A, Mariacher-Gehler S, Stucki G. A new look at the Western Ontario and
403 McMaster Universities Osteoarthritis Index using Rasch analysis. Arthritis Care Res [Internet] 1999;12:331-5.
404 Available from: <Go to ISI>://WOS:000083009200004\nhttp://onlinelibrary.wiley.com/store/10.1002/1529-
405 0131(199910)12:5<331::AID-ART4>3.0.CO;2-
406 W/asset/4_ftp.pdf?v=1&t=i8geidl&s=8752fa134a34dd764789f1c25066301b7320e6e6
- 407 33. Roos EM, Lohmander LS. The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to
408 osteoarthritis. Health Qual Life Outcomes [Internet] 2003;1:64. Available from:
409 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=280702&tool=pmcentrez&rendertype=abstract>
- 410 34. Hawker GA, Davis AM, French MR, Cibere J, Jordan JM, March L, et al. Development and preliminary
411 psychometric testing of a new OA pain measure - an OARSI/OMERACT initiative. Osteoarthr Cartil 2008;16:409-
412 14.

413

414

415 Figure Legend

416 Figure 1: Sensitivity to Change of Single Pain-Related Outcomes from the Tanezumab Trial. Values plotted
417 are the control-treatment differences in standard score, at different study time points. More negative
418 values indicate increased sensitivity to change