BMJ

RESEARCH

# Reporting and interpretation of SF-36 outcomes in randomised trials: systematic review

Despina G Contopoulos-Ioannidis, assistant professor,[1,2] Anastasia Karvouni, research fellow,[3] Ioanna Kouri, research fellow,[3] John P A Ioannidis, professor[3,4]

[1]Department of Paediatrics, University of Ioannina School of Medicine, Ioannina, Greece
[2]Department of Paediatrics, George Washington University, School of Medicine and Health Sciences, Washington, DC, USA
[3]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Greece
[4]Institute for Clinical Research and Health Policy Studies, Tufts University School of Medicine, Boston, MA, USA

**Correspondence to: J P A Ioannidis, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece**
**jioannid@cc.uoi.gr**

Cite this as: *BMJ* 2009;339:a3006
doi:10.1136/bmj.a3006

## ABSTRACT

**Objective** To determine how often health surveys and quality of life evaluations reach different conclusions from those of primary efficacy outcomes and whether discordant results make a difference in the interpretation of trial findings.

**Design** Systematic review.

**Data sources** PubMed, contact with authors for missing information, and author survey for unpublished SF-36 data.

**Study selection** Randomised trials with SF-36 outcomes (the most extensively validated and used health survey instrument for appraising quality of life) that were published in 2005 in 22 journals with a high impact factor.

**Data extraction** Analyses on the two composite and eight subdomain SF-36 scores that corresponded to the time and mode of analysis of the primary efficacy outcome.

**Results** Of 1057 screened trials, 52 were identified as randomised trials with SF-36 results (66 separate comparisons). Only eight trials reported all 10 SF-36 scores in the published articles. For 21 of the 66 comparisons, SF-36 results were discordant for statistical significance compared with the results for primary efficacy outcomes. Of 17 statistically significant SF-36 scores where primary outcomes were not also statistically significant in the same direction, the magnitude of effect was small in six, moderate in six, large in three, and not reported in two. Authors modified the interpretation of study findings based on SF-36 results in only two of the 21 discordant cases. Among 100 additional randomly selected trials not reporting any SF-36 information, at least five had collected SF-36 data but only one had analysed it.

**Conclusions** SF-36 measurements sometimes produce different results from those of the primary efficacy outcomes but rarely modify the overall interpretation of randomised trials. Quality of life and health related survey information should be utilised more systematically in randomised trials.

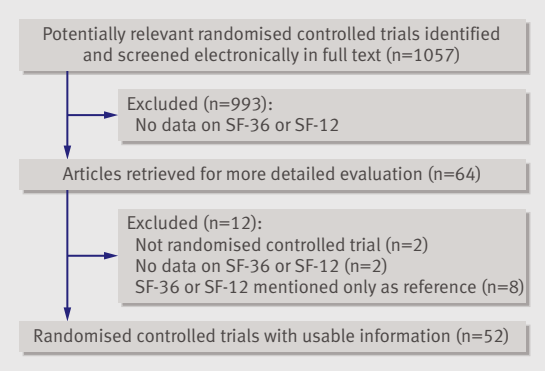
## INTRODUCTION

Quality of life outcomes and surveys of overall health status are considered useful to incorporate in randomised trials.[1-7] Such data would be important to collect and report systematically, regardless of whether the results agree with the primary outcomes or not. It is unknown whether it is common for quality of life and health survey results to reach different conclusions from those of the primary efficacy outcomes, whether there is selective reporting of outcomes, or whether discordant results in these outcomes modify the conclusions of these trials. We therefore evaluated recently published trials (2005) in 22 leading journals. Many generic and disease specific quality of life and health survey measures exist.[2 8] Some are difficult to compare or are prone to methodological shortcomings and suboptimal validation.[1 2 9-11] To maximise comparability across trials covering diverse diseases and interventions we focused on trials using short form-36 (SF-36). Originally developed as a multipurpose health survey instrument, SF-36 has been translated in more than 50 countries as part of the international quality of life assessment project and has become the most extensively validated and used generic instrument for measuring quality of life. It is an instrument that has extensive applications for population health surveys, comparisons of relative burden of diseases, and differentiation of health benefits across groups produced by diverse interventions.[11-14]

## METHODS

We considered randomised trials with data on SF-36 published in 2005 in five major general medicine journals (*New England Journal of Medicine*, *JAMA*, *Lancet*, *BMJ*, *PLoS Medicine*) and 17 specialty journals with the highest impact factor among those that publish research from clinical trials for 2005 using the 2005 Journal Citation Reports (*Circulation*, *Journal of the American College of Cardiology*, *Gastroenterology*, *Hepatology*, *Journal of the National Cancer Institute*, *Journal of Clinical Oncology*, *Blood*, *Annals of Internal Medicine*, *Diabetes*, *Diabetes Care*, *Brain*, *Annals of Neurology*, *American Journal of Respiratory and Critical Care Medicine*, *Journal of American Society of Nephrology*, *Arthritis and Rheumatism*, *American Journal of Psychiatry*, *Archives of General Psychiatry*).

We considered randomised trials to be eligible that reported on any of the two composite (physical, mental) and eight subdomain SF-36 scores (physical functioning, role physical, bodily pain, general health,




Flow chart of papers through trial

vitality, social functioning, role emotional, mental health). When referral was made to additional separate publications reporting primary efficacy or SF-36 outcomes, these were also retrieved. We also considered trials using SF-12, a shorter version of SF-36 (for composite scores). No restriction was set on disease and compared interventions. Whenever information was not reported on all 10 scores, we asked authors for missing information.

We searched the 22 target journals through PubMed using limits for randomised clinical trial (type of study) and 2005 (year of publication). Identified articles were downloaded in PDF format and screened electronically using Acrobat Reader "Find" tool for keywords: quality of life, SF36, SF 36, SF-36, short form 36, short form-36, SF-12, SF12, mental composite score, physical composite score, medical outcome study, MOS 36, MOS-36, and Ware. Articles passing electronic screening were further evaluated by two independent investigators (AK and IK). Disagreements were resolved by consensus. Remaining disagreements were resolved by DGC-I.

To probe whether SF-36 data may have remained unpublished we communicated (three emails, each sent three weeks apart) with the corresponding authors of 100 trials randomly selected among those not reporting SF-36 data. Selection was based on a list of 100 numbers generated randomly and applied to the 1057 retrieved articles, ordered serially per journal, after excluding the 52 eligible articles.

### Data extraction

Data were extracted by three independent investigators (IK, AK, and DGC-I). Discrepancies were resolved by consensus. Remaining disagreements were resolved by JPAI.

From each eligible article we extracted information on authors, journal, design (superiority or non-inferiority), condition, interventions compared, sample size (randomised, analysed for SF-36), definition of primary efficacy outcome (as reported; if not clarified, we selected the outcome used for sample size calculations), time points and statistical analysis for the primary outcome and SF-36 assessments, whether SF-36 was a co-primary outcome, and whether any other quality of life and health related survey scales were used. We also recorded which SF-36 scores were reported and for which we could obtain missing information from authors.

### Discordant results

For the primary efficacy outcome and for each of the presented SF-36 assessments we recorded whether the difference between compared arms was statistically significant ($P<0.05$) favouring the experimental arm, non-statistically significant, or statistically significant favouring the control arm. For trials with more than two arms we considered the comparison of each experimental intervention against control separately. We considered all comparisons and also present results separately for superiority and non-inferiority trials.[15]

Data on SF-36 outcomes were extracted for the reported analyses that corresponded as closely as possible to the same time points as for primary outcome data. Specifically, when measurements for primary or SF-36 outcomes were carried out at several time points, for primary efficacy outcomes we preferred analyses accounting for multiple measurements (for example, repeated measurement analysis) than analyses of single time points. If the primary outcome was a time to event analysis or incorporated serial longitudinal measurements, we preferred the analysis of serial longitudinal SF-36 measurements; if this was unavailable, we recorded whether there was formal statistically significant difference at any time points when SF-36 had been appraised. When the primary outcome was appraised at a single time point, we recorded the SF-36 outcomes at the single same (or closest) time point. In two comparisons where co-primary outcomes existed and could not be prioritised, we based the evaluation of statistical significance on overall authors' interpretation.

We considered SF-36 results as statistically significant when at least one of the composite or subdomain scores showed a statistically significant result in favour or against the experimental intervention. There were no situations where some of the specific SF-36 scores were significant for the experimental intervention and others were significant against.

For statistically significant SF-36 effects when the respective primary efficacy outcome was discordant, we extracted information on the effect size of SF-36. Roughly, standardised mean differences of less than 0.30 standard deviations are small effects, 0.30-0.80 are moderate, and more than 0.80 are large.[16-20] The corresponding cut-offs for raw scores are less than 4, 4-10, and more than 10 points.

For comparisons with discordant statistical significance on SF-36 and primary outcome results, we recorded whether the authors had discussed the SF-36 results at all, whether they commented on the discrepancy and if so with what arguments, and if SF-36 findings changed the interpretation of the trial results.

## RESULTS

Overall 1057 trials were screened and 52 eligible trials identified[w1-w52] with 66 eligible comparisons (figure and web extra table). Additional data were presented in other published articles on primary efficacy for one trial[w43] and SF-36 for eight trials.[w4 w21 w24 w29 w35 w36 w46 w51] Additional SF-36 data were provided directly by the authors in 11 trials with 13 comparisons (see web extra fig 1). Forty two trials (56 comparisons) addressed superiority, and 10 (10 comparisons) non-inferiority. In seven trials (10 comparisons)[w2 w8 w35 w39 w40 w44 w45] SF-36 was described as a co-primary outcome. Additional quality of life or health survey instruments appeared in 16 trials (16 comparisons).

Eventually, data for physical composite score and mental composite score were available for 34 trials (39 comparisons) and 35 trials (40 comparisons, see web extra fig 1). Data on at least one of the eight subdomain scores were available for 36 trials (48 comparisons). Data on all possible SF-36 scores were available for 18 trials (eight published, 10 obtained from authors). Six trials[w6 w23 w29 w31 w35 w44] had collected information a priori only for specific subdomains.

### Concordance of results

Of the 66 comparisons, 21 (32%) had discordant statistical significance for primary efficacy and SF-36 results (table 1). Moreover, of the 56 comparisons of superiority trials 19 had discordant primary efficacy and SF-36 results (see web extra fig 2).

In one[w44] of the 21 discrepancies, SF-36 was a co-primary outcome. In seven discrepancies, additional quality of life or health survey instruments were also used. In two trials[w14 w51] the additional instruments agreed with SF-36, and in five[w12 w15 w31 w44 w47] they agreed with the primary efficacy outcome.

In the 13 discordant comparisons with only SF-36 significant results (nine comparisons in favour and four against the experimental intervention; in seven trials[w7 w14 w21 w31 w44 w46 w47] and three trials,[w15 w43 w51] respectively) there were 17 statistically significant specific scores (five normalised, 10 raw, two reporting only statistical significance without effect size); effect sizes were small in six, moderate in six, and large in three.

**Table 1** | Concordance of statistical significance in SF-36 and primary outcome results

| | SF-36 results | | | |
|---|---|---|---|---|
| Primary outcome | Significant* | Non-significant | Significant (against)† | Total |
| Significant | 21 | 8 | 3 | 32 |
| Non-significant | 9‡ | 23 | 1 | 33 |
| Significant (against) | 0 | 0 | 1 | 1 |
| Total | 30 | 31 | 5 | 66 |

κ coefficient 0.33 (95% confidence interval 0.06 to 0.59) for concordance of primary outcome against SF-36. No situations occurred where specific SF-36 scores were significant for experimental intervention and others were significant against.
*At least one of composite or subdomain scores shows statistically significant result in favour of experimental intervention.
†At least one of composite or subdomain scores shows statistically significant result against experimental intervention.
‡This category contains the only two studies (w31 and w44) where interpretation of study findings was modified based on SF-36 results.

### Interpretation of trial findings in discordant settings

**Improved primary outcome only**—SF-36 results did not modify the trial's interpretation of these 11 comparisons (eight trials, table 2).[w4 w12 w16 w18 w41 w42 w43 w51] In five comparisons (four trials), SF-36 outcomes were only tabulated or alluded to in the results, without further discussion.[w12 w16 w18 w42] In the other four trials the authors focused on other non-primary outcomes,[w4] claimed that SF-36 was not sensitive enough to detect improvements,[w41] adopted a non-intention to treat analysis for SF-36 with significant results,[w43] or dismissed the importance of the negative effects on SF-36 in the face of benefits in disease-free survival.[w51]

**Improved SF-36 only**—SF-36 modified the interpretation of only two trials.[w31 w44] The authors favoured the peer modelling videotape for breast cancer based on the significant and large improvement on SF-36 vitality despite no improvement on the IES-R (revised impact of events scale) (both were co-primary outcomes, table 2).[w44] In the chronic renal failure anaemia trial the benefit in vitality score from erythropoietin was acknowledged as clinically important.[w31] In the other five comparisons (three trials), benefits on SF-36 did not change the interpretation.[w7 w14 w21] One trial dismissed the SF-36 difference as transient and weak,[w14] one trial considered the non-statistically significant benefits in efficacy as clinically important, whereas the significant improvements in SF-36 vitality scores were considered clinically unimportant and the authors then even questioned the use of SF-36 in trials on diabetes,[w21] and in another trial the authors considered that the clinical significance of statistically significant differences in SF-36 domains in patients with fibromyalgia could not be evaluated.[w7]

**Improved SF-36, non-inferiority on primary outcome**—SF-36 did not modify the interpretation of these two trials.[w46 w47] Both trials already concluded favourably for the experimental intervention that achieved the desired non-inferiority, and in one of them[w47] the observed benefit in SF-36 was considered possibly due to chance.

**Only SF-36 worsened**—In one trial[w15] where SF-36 worsened with the experimental intervention, the investigators interpreted the results as showing no consistent differences in quality of life, because an additional instrument (EQ5D) showed no significant differences.

### Probing unpublished data

Authors of 69 of 100 additional randomly selected trials responded. SF-36 data had actually been collected from five trials. The data had been analysed for only one trial and did not show any statistically significant differences for SF-36 or the primary efficacy outcome.

## DISCUSSION

In one third of the trial comparisons in our empirical evaluation, differential effects on primary efficacy outcomes compared with SF-36 were identified. However, when SF-36 compared with efficacy

**Table 2** | **Trial comparisons with discordant SF-36 and primary efficacy outcome results (21 discrepant comparisons in 16 trials)**

| Author (reference) | Condition | Comparison | Primary efficacy outcome | Interpretation |
|---|---|---|---|---|
| Improved primary outcome only: | | | | |
| Campbell[w4]* | Refractory ascites | Transjugular intrahepatic portal-systemic shunt+large volume paracentesis (as needed) *v* large volume paracentesis (as needed) | Ascites recurrence | Improvement in primary outcome considered not worth it because of no survival benefit and possible worsening of encephalopathy; these competing effects considered to nullify any changes in quality of life |
| Devière[w12] | Gastroesophageal reflux | Endoscopic implantation of biocompatible non-resorbable copolymer (Enteryx; Boston Scientific) *v* sham procedure | Reduction in use of proton pump inhibitor | SF-36 outcomes only alluded to in results, without discussion |
| Fairbank[w16] | Chronic low back pain | Surgical treatment *v* intensive rehabilitation | Oswestry disability index score | SF-36 outcomes tabulated only in results, without discussion |
| Gilron[w18] | Neuropathic pain | Garbapentin+morphine *v* morphine | Mean daily pain score | SF-36 outcomes alluded to only in results, without discussion |
| Shaheen†w41 | Ulceration after band ligation | Pantoprazole *v* placebo | Size of oesophageal ulcer | SF-36 considered not sensitive enough to detect improvements |
| Sherman[w42] | Chronic low back pain | Yoga *v* self care book; yoga *v* exercise | Roland-Morris disability questionnaire score | SF-36 outcomes alluded to only in results, without discussion |
| Singh†‡[w43] | Atrial fibrillation | Amiodarone *v* sotalol; amiodarone *v* placebo; sotalol *v* placebo | Recurrence of atrial fibrillation | Amiodarone *v* sotalol, and amiodarone *v* placebo: briefly mentioned in results that amiodarone associated with significantly worse mental health scores (P=0.005, no effect size provided) and focused on non-intention to treat analysis (patients on sinus rhythm *v* with recurrent arrhythmia); sotalol *v* placebo: focused on non-intention to treat analysis (patients on sinus rhythm *v* with recurrent arrhythmia) |
| Whelan*[w51] | Breast cancer | Letrozol *v* placebo | Disease-free survival | Negative effects on SF-36 dismissed in face of significant benefits in disease-free survival; three subdomains affected, but SF-36 effects considered transient and small (‹0.2 SD) |
| Improved SF-36 only: | | | | |
| Stanton§¶w44 | Breast cancer (re-entry phase after surgery) | Peer modelling videotape *v* standard print material alone | Revised impact of events scale score | Peer modelling videotape was favoured based on significant and large (0.92 SD) improvement on SF-36 vitality despite no improvement on IES-R; SF-36 and IES-R were co-primary outcomes |
| Parfrey¶w31 | Anaemia in renal failure | Erythropoietin for high *v* low target haemoglobin | Left ventricular volume index | Benefit in vitality score (0.35 SD) acknowledged as clinically important and as "only consistent benefit conferred by normalizing hemoglobin in patients with chronic kidney disease" |
| Crofford[w7] | Fibromyalgia | Pregabalin (150 mg) *v* placebo; pregabalin (300 mg) *v* placebo | Pain score | Pregabalin 150 mg and 300 mg offered moderate benefits in general health (4.6 and 5.9 points, respectively), even though their clinical significance was deemed uncertain |
| EVAR 1[w14] | Abdominal aortic aneurysm | Endovascular *v* open repair of aneurysm | All cause mortality | Benefit in physical component score dismissed because difference was transient and not strong (0.17 SD) |
| Hill-Briggs*[w21] | Type 2 diabetes | Nurse case manager *v* usual care; community health worker *v* usual care | Haemoglobin $A1_C$ | Nurse case manager *v* usual care: non-statistically significant decrease in haemoglobin $A1_C$ levels was considered "clinically important" and significant improvements of moderate magnitude (8.53 points) in SF-36 vitality scores considered clinically unimportant; community health worker *v* usual care: as above with improvement in vitality score of 6.34 points. SF-36 use in diabetes trials was questioned |
| Improved SF-36, non-inferiority on primary outcome: | | | | |
| Sweeney*[w46] | Fast ventricular tachycardia | Antitachycardia pacing *v* shock first (with implantable cardioverter defibrillator) | Duration of fast tachycardia | Experimental intervention was favoured because it "is highly effective, equally safe and improves quality of life." Effect sizes for significant SF-36 scores were moderate or large (5, 12, 20 points) |
| UKATT[w47] | Alcoholism | Social behaviour and network therapy *v* motivational enhancement | Days of abstinence | Experimental intervention worth adopting based on primary efficacy results. Observed benefit in physical composite score was small (0.13 SD) and considered as possibly "due to chance" |
| Only SF-36 worsened: | | | | |
| EVAR 2[w15] | Abdominal aortic aneurysm (unfit for repair) | Endovascular repair of aneurysm *v* no repair | All cause mortality | Authors already concluded against experimental intervention because of no survival benefit and need for continuous surveillance and reintervention. Authors saw no clear and consistent differences in quality of life. EQ5D quality of life showed no significant differences and SF-36 detrimental effect was small (0.19 SD) |

EQ5D=EuroQoL-5 dimension; IES-R=revised impact of events scale score. Only discordant comparisons per trial are shown; additional comparisons may exist in same trial.
*Retrieved trial was separate publication for quality of life results. Corresponding publications with primary efficacy trial results: for Campbell[w4] was Sanyal et al. *Gastroenterology* 2003;124:634-41; for Hill-Briggs[w21] was Garry et al. *Prev Med* 2003;37:23-32; for Sweeney[w46] was Wathen et al. *Circulation* 2004;110:2591-6; and for Whelan[w51] was Goss et al. *N Eng J Med* 2003;349:1793-802.
‡Retrieved trial was separate publication for primary efficacy outcomes only. Corresponding publication with SF-36 results for Singh[w43] was Singh et al. *J Am Coll Cardiol* 2006;48:721-30.
§SF-36 questionnaire was a co-primary outcome for Stanton[w44] among two primary outcomes.
¶Overall interpretation of study was modified by SF-36 results.
†Authors had provided additional SF-36 data after contact.

## WHAT IS ALREADY KNOWN ON THIS TOPIC

Quality of life and related health survey outcomes could be essential in deciding whether an intervention is worth adopting

It is unknown whether such outcomes reach different conclusions from those of primary efficacy outcomes or whether they affect the interpretation of current clinical trials

## WHAT THIS STUDY ADDS

Several randomised trials published in influential journals have had discordant results on primary efficacy outcomes compared with SF-36

When SF-36 and efficacy outcomes reached discordant conclusions, SF-36 rarely modified the interpretation of these trials

Quality of life and health related survey information deserves more standardised and systematic use in randomised trials

outcomes reached discordant conclusions, SF-36 rarely affected the interpretation of these trials. What we observed was generally a tendency to belittle rather than to pronounce discordant results. Several trials did not discuss the SF-36 findings at all, and most did not report all the tested SF-36 scores. Considering post hoc an instrument as insensitive or not worth reporting contradicts the initial choice to use this instrument as a trial outcome.

In most trials for chronic conditions, quality of life and surveys of health status are useful to consider. SF-36 was reported in fewer than 5% of the trials we screened, and our author survey suggested that some additional trials (at least five of 100) had collected information on SF-36 but without analysing or publishing it two or three years after the publication of the main trial results. Quality of life seems to remain undervalued in clinical research: few trials collect quality of life related data, fewer report on them, data are only partially presented, and quality of life rarely affects the trial interpretation.

We should acknowledge some caveats. Firstly, by selecting high impact journals we identified trials with high visibility and probably also high quality.[21] It is unlikely that this strategy would have selected for discordant results between outcomes. Secondly, selective analysis and reporting bias may affect primary outcomes and not just SF-36,[22-25] but this should not have increased the perceived rate of discrepancies between outcomes. Thirdly, discordance at the level of statistical significance does not necessarily mean that results for different outcomes differ beyond chance. Among statistically significant results, chance findings and non-clinically important differences are possible, and primary outcomes should be given more weight in the discussion than secondary outcomes. Given that trials are typically powered to address the primary outcome, a significant result in the primary outcome with a non-significant result in quality of life or health survey assessments may sometimes simply reflect lack of power for the quality of life or health survey outcome. Therefore we also examined the SF-36 effect sizes and the circumstances and discussion of discordant results. Fourthly, we did not carry out the same in-depth evaluation for trials where efficacy and SF-36 outcomes were concordant. It is unlikely that authors would then have modified their inferences, but SF-36 may have strengthened the conclusions. Finally, we did not examine trials using only other quality of life or health survey instruments beyond SF-36. However, SF-36 is the most robustly standardised and widely used one, and we wanted to maximise comparability. Although other scales may also be used, one study found that only 4.2% of trials reported any quality of life outcome.[2]

Although quality of life and health survey scales have been used in clinical trials for over 25 years, several issues remain debated.[26] Besides problems of fragmented, selectively reported information, it is sometimes impossible to say whether and which analyses are based on a priori analytical plans.[11 27] Proper attention to the importance of these outcomes should be given in clinical trials. Otherwise, with a growth in the clinical trials' administrative paperwork,[28] outcomes such as SF-36 may become routine compulsory assessments without a genuine interest to learn from them.

Overall, quality of life and health survey assessments provide a different window into patient outcomes and deserve to be included in more trials with complete reporting of results, and standardised interpretation. Unbiased data on these outcomes may enhance our ability to improve clinical decision making.

We thank the following for clarifying their results or providing additional data: N Assefi, D Buchwald, and C Jacobsen (for Assefi et al[w1]); A Avenell and JA Cook (for Avenell et al[w2]); J Carratala, LJ Crofford, J Pepper, and B Lees (for De Arenaza et al[w10]); H Escobar-Morreale, RM Greenhalgh, and LC Brown (for EVAR 1[w14] and EVAR 2[w15]); I Gilron, S Hewlett, F Hill-Briggs, CA Hukins, P Jellema, and DA van der Windt (for Jellema et al[w23]); JA Klaber-Moffett, K Linde, PS Parfrey, and RN Foley (for Parfrey et al[w31]); DJ Torgerson (for Porthouse et al[w32]); D Revicki and JM Miranda (for Revicki et al[w35]); M Rienstra and IC van Gelder (for Rienstra et al[w36]); BL Rollman, MF Scheier, and S Colvin (for Scheier et al[w40]); N Shaheen, BN Singh, AL Stanton, and G Bleijenberg (for Stulemeijer et al[w45]); D van der Heijde and JH Stone (for Wegener's Granulomatosis Etanercept Trial[w50]); and WA Whitelaw.

**Contributors:** JPAI conceived the study and is guarantor. All authors designed the protocol, analysed and interpreted the data, and approved the final manuscript. DGC-I, IK, and AK collected the data. DGC-I and JPAI drafted the manuscript. IK and AK critically revised the manuscript for important intellectual content.

**Funding:** None.

**Competing interests:** None declared.

**Ethical approval:** Not required.

**Provenance and peer review:** Not commissioned; externally peer reviewed.

1 Guyatt GH, Naylor CD, Juniper E, Heyland DK, Jaeschke R, Cook DJ. Users' guides to the medical literature. XII. How to use articles about health-related quality of life. Evidence-Based Medicine Working Group. *JAMA* 1997;277:1232-7.

2 Sanders C, Egger M, Donovan J, Tallon D, Frankel S. Reporting on quality of life in randomised controlled trials: bibliographic study. *BMJ* 1998;317:1191-4.

3 Fayers PM, Hopwood P, Harvey A, Girling DJ, Machin D, Stephens R. Quality of life assessment in clinical trials—guidelines and a checklist for protocol writers: the UK Medical Research Council experience. MRC Cancer Trials Office. *Eur J Cancer* 1997;33:20-8.

4 Food and Drug Administration. Draft guidance for industry on patient-reported outcome measures: use in medicinal product development to support labeling claims. *Fed Register* 2006;71:5862-3.

5 Committee for Medicinal Products for Human Use. Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products. www.emea.europa.eu/pdfs/human/ewp/1393104en.pdf. 2005.

6 Revicki DA. Regulatory Issues and Patient-Reported Outcomes Task Force for the International Society for Quality of Life Research. FDA draft guidance and health-outcomes research. *Lancet* 2007;369:540-2.
7 Revicki DA, Gnanasakthy A, Weinfurt K. Documenting the rationale and psychometric characteristics of patient reported outcomes for labeling and promotional claims: the PRO evidence dossier. *Qual Life Res* 2007;16:717-23.
8 Michael M, Tannock IF. Measuring health-related quality of life in clinical trials that evaluate the role of chemotherapy in cancer treatment. *CMAJ* 1998;158:1727-34.
9 Efficace F, Bottomley A, Vanvoorden V, Blazeby JM. Methodological issues in assessing health-related quality of life of colorectal cancer patients in randomised controlled trials. *Eur J Cancer* 2004;40:187-97.
10 Bottomley A, Vanvoorden V, Flechtner H, Therasse P; EORTC Quality of Life Group EORTC Data Center. The challenges and achievements involved in implementing quality of life research in cancer clinical trials. *Eur J Cancer* 2003;39:275-85.
11 Lee CW, Chi KN. The standard of reporting of health-related quality of life in clinical cancer trials. *J Clin Epidemiol* 2000;53:451-8.
12 Ware JE Jr. SF-36 health survey update. *Spine* 2000;25:3130-9.
13 Ware JE, Sherbourne CD. The MOS 36-item short form health survey (SF-36): conceptual framework and items selection. *Med Care* 1992;30:473-83.
14 Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. *BMJ* 2002;324:1417-21.
15 Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ; CONSO RT Group. Reporting of non-inferiority and equivalence randomized trials: an extension of the CONSORT statement. *JAMA* 2006;295:1152-60.
16 Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;61:102-9.
17 Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 2003;41:582-92.
18 Ringash J, O'Sullivan B, Bezjak A, Redelmeier DA. Interpreting clinically significant changes in patient-reported outcomes. *Cancer* 2007;110:196-202.
19 Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol* 1999;52:861-73.
20 Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77:371-83.
21 Nieminen P, Carpenter J, Rucker G, Schumacher M. The relationship between quality of research and citation frequency. *BMC Med Res Methodol* 2006;6:42-9.
22 Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clin Trials* 2007;4:245-53.
23 Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ* 2005;330:753.
24 Chan AW, Krleza-Jerić K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* 2004;171:735-40.
25 Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291:2457-65.
26 Osoba D. Translating the science of patient-reported outcomes assessment into clinical practice. *J Natl Cancer Inst Monogr* 2007;37:5-11.
27 Staquet M, Berzon R, Osoba D, Machin D. Guidelines for reporting results of quality of life assessments in clinical trials. *Qual Life Res* 1996;5:496-502.
28 Grimes DA, Hubacher D, Nanda K, Schulz KF, Moher D, Altman DG. The good clinical practice guideline: a bronze standard for clinical research. *Lancet* 2005;366:172-4.


**Accepted:** 19 September 2008