Spine

SPINE Volume 40, Number 10, pp 734-739

Can We Convert Between Outcome Measures of Disability for Chronic Low Back Pain?

Tom Morris, PhD,* Siew Wan Hee, PhD,† Nigel Stallard, PhD,† Martin Underwood, MD,† and Shilpa Patel, DHealthPsy†

Study Design. Retrospective database analysis.

Objective. A range of patient-reported outcomes were used to measure disability due to low back pain. There is not a single back pain disability measurement commonly used in all randomized controlled trials. We report here our assessment as to whether different disability measures are sufficiently comparable to allow data pooling across trials.

Summary of Background Data. We used individual patient data from a repository of data from back pain trials of therapist-delivered interventions.

Methods. We used data from 11 trials (n = 6089 patients) that had at least 2 of the following 7 measurements: Roland-Morris Disability Questionnaire, Chronic Pain Grade disability score, Physical Component Summary of the 12- or 36-Item Short Form Health Survey, Patient Specific Functional Scale, Pain Disability Index, Oswestry Disability Index, and Hannover Functional Ability Questionnaire. Within each trial, the change score between baseline and short-term follow-up was computed for each outcome and this was used to calculate the correlation between the change scores and the Cohen's κ for the 3-level outcome of change score of less than, equal to, and more than zero. It was considered feasible to pool 2 measures if they were at least moderately correlated (correlation >0.5) and have at least moderately similar responsiveness (κ >0.4).

Results. Although all pairs of measures were found to be positively correlated, most correlations were less than 0.5, with only 1 pair of outcomes in 1 trial having a correlation of more than 0.6. All κ statistics were less than 0.4 so that in no cases were the criteria for acceptability of pooling measures satisfied.

Conclusion. The lack of agreement between different outcome measures means that pooling of data on these different disability measurements in a meta-analysis is not recommended.

Key words: agreement between measurements, correlation, crosswalking, individual participant data, low back pain, meta-analysis, patient-reported outcome, responsiveness.

Level of Evidence: 2

Spine 2015;40:734–739

Patient-reported outcome measures (PROMs) are commonly used in low back pain (LBP) research comparing therapist-delivered interventions. These outcomes are used to measure participants' perspectives on their symptoms, capabilities, performance functioning, treatment preferences, and general well-being.

Investigators tend to choose instruments with which they are familiar or those recommended in consensus statements. Although all these instruments aim to measure similar constructs, there is little information on their compatibility and comparability. To compare results based on different measures, it is important to know if summary measures such as treatment effect sizes from one instrument have the same interpretation as that from another instrument. The commonest outcome measures used in randomized controlled trials for LBP, and the ones that researchers are most familiar with interpreting are the Roland-Morris Disability Questionnaire (RMDQ) score and the Oswestry Disability Index (ODI).[1] Being able to standardize outcomes to measure in one of these would improve quality of the interpretation of outcomes. The importance of being able to crosswalk scores between different measures was identified by the National Institutes of Health Task Force on Research Standards in Chronic Low Back Pain as an important research priority.[2]

If the measures are comparable, then it is possible to compare data from studies using different measures and to pool these data in a meta-analysis. If the measures are not

From the *Leicester Clinical Trials Unit, University of Leicester, Leicester, United Kingdom; and †Warwick Medical School, University of Warwick, Coventry, United Kingdom.

Acknowledgment date: November 25, 2014. First revision date: February 5, 2015. Acceptance date: February 23, 2015.

The manuscript submitted does not contain information about medical device(s)/drug(s).

National Institute for Health Research under its Programme Grants for Applied Research Programme (grant reference number RP-PG-0608-10076) funds were received in support of this work. This project benefitted from facilities funded through Birmingham Science City Translational Medicine Clinical Research and Infrastructure Trials Platform, with support from Advantage West Midlands and the Wolfson Foundation.

Relevant financial activities outside the submitted work: board membership, grants, payment for lectures, employment, travel/accommodations/meeting expenses.

Address correspondence and reprint requests to Siew Wan Hee, PhD, Division of Health Sciences, Warwick Medical School, The University of Warwick, Coventry CV4 7AL, United Kingdom; E-mail: s.w.hee@warwick.ac.uk; shilpa.patel@warwick.ac.uk

DOI: 10.1097/BRS.0000000000000866

comparable, then such comparisons and any meta-analysis using different measures may not be robust.

We have developed a large pooled data set of individual patient data from 19 trials (n = 9328) of therapist-delivered interventions for LBP that will be a resource for researchers working in the field (report submitted to the National Institute for Health Research).[3] All included trials in this pooled data set used at least one of the 6 PROMs designed to measure the aforementioned back pain–related disability or included generic-based health-related quality-of-life instruments such as 12- (SF-12)[4] or 36-Item Short Form Health Survey (SF-36).[5] However, no common instrument was used by all these trials.

In this article, we assess the agreement between the instruments by determining their correlation and responsiveness to detect positive, zero, or negative change at an individual participant level with the intention of calibrating measures against each other to allow data pooling using a single common scale.

MATERIALS AND METHODS

Trials, Instruments, and Change Scores

There are a number of back pain–related disability outcome measures used in the research, each to varying degrees. In our data set, we had data available on 6 PROMs that aim to measure back pain–related disability, namely, the Chronic Pain Grade (CPG) disability score, which is one of the 2 domains in the CPG that aims to grade chronic pain status,[6] the Hannover Functional Ability Questionnaire (FFbHR),[7] the ODI,[8] the Pain Disability Index (PDI),[9] the mean score of 3 items from the Patient Specific Functional Scale (PSFS),[10] and the RMDQ.[11]

Eleven of the 19 trials (n = 6089) included 2 or more measures of back pain–related disability or included data that allowed us to calculate the Physical Component Summary (PCS) from the generic-based SF-12/36, recorded at baseline and short-term follow-up (2–3 mo postrandomization).[12–22] We used individual patient data from these trials to make comparisons between back pain–specific measures and SF12/36 PCS to facilitate indirect comparisons between back pain–specific measures.

The change score for each individual patient was defined as the difference between the score at short-term follow-up and baseline with sign allocated so that a positive change score indicates an improvement in disability in each case. We compared change scores of each instrument within each trial.

Correlation and Responsiveness

In order for conversion between outcome measures to be meaningful, the change in each measure should be correlated and have similar responsiveness,[23] where the latter is explained as follows. Correlation was assessed by calculation of the Pearson correlation coefficient and illustrated using scatterplots. *A priori* we considered correlations greater than 0.5 (a large effect size) to indicate a level of correlation that would allow pooling of data collected from different measures.[23,24] This criterion was lower than the one used (0.7) in a similar study that examined the justification of combining scores for meta-analyses in chronic obstructive pulmonary disease.[23]

Responsiveness is the ability to detect a change in condition. If 2 measures are similarly responsive when a patient's condition improves or worsens over time, then this should be reflected by a change in the patient's score on both measures. If 2 outcome measures do not have similar responsiveness, then combining them in a meta-analysis may introduce heterogeneity that could be falsely attributed to other sources, such as the treatment effect.

Similarity of responsiveness of 2 outcome measures was examined by categorizing the change scores as negative change (change score <0), no change (change score $= 0$), or positive change (change score >0), and calculating Cohen's κ from these categorizations.[25] *A priori* we considered κ more than 0.4 to indicate sufficiently similar responsiveness.[26] These broad categories were chosen to demonstrate whether or not the outcome measures had similar responsiveness in the most basic sense (improved, worsened, or no change). All analyses were run in R.[27]

RESULTS

We included data from 11 trials (n = 6089) in this analysis (Table 1), allowing 21 pairwise comparisons between outcomes within trials. Figure 1A–F shows a selection of scatterplots of standardized change scores of these outcome measures. The other scatterplots are available as supplementary materials (see Supplementary Digital Content Figure 1A–D, available at: http://links.lww.com/BRS/A974, http://links.lww.com/BRS/A975, http://links.lww.com/BRS/A976, and http://links.lww.com/BRS/A977). It is clear from these plots that although instruments seem to be positively correlated, there is a large disagreement between the outcomes.

Correlations and κ statistics are shown in Table 2. The correlations ranged from 0.21 to 0.70, confirming that these instruments are positively correlated and with the linear associations between them ranging from weak to moderately strong. Where several trials include the same pair of measures, it is interesting to compare the correlations obtained. Three trials had both SF-12/36 PCS and FFbHR data, and the correlations in the 3 trials were very similar, all of about 0.58.[12,14,22] Another 3 trials had both SF-12/36 PCS and CPG, and the correlations between these measures in the different trials were reasonably similar, ranging from 0.41 to 0.56,[14,16,20] and 4 trials had both SF-12/36 PCS and RMDQ, with range from 0.38 to 0.52, again similar.[13,16,17,20] However, correlations between other outcomes were quite widely ranging across trials: between CPG and RMDQ (3 trials; range, 0.21–0.47)[16,20,21] and between PSFS and RMDQ (3 trials; range, 0.40–0.70).[15,17,18]

Cohen's κ statistics calculated for the 3 by 3 table with the number of patients with positive change, no change, or negative change on each outcome was less than 0.4 for all 21 comparisons. Some were similar between trials, namely, for PCS and FFbHR (range, 0.27–0.30)[12,14,22] and for PCS and CPG

TABLE 1. Instruments Used and Number of Patients by Trial

Trial	n	Outcome Measures		
UK BEAM[20]	885	RMDQ	CPG	PCS
BeST[16]	426	RMDQ	CPG	PCS
Brinkhaus *et al*[12]	281	PCS	FFbHR	PDI
Haake *et al*[14]	1110	CPG	FFbHR	PCS
Hancock *et al*[15]	235	RMDQ	PSFS	
HULLEXPROB[13]	203	RMDQ	PCS	
Macedo *et al*[17]	158	RMDQ	PCS	PSFS
Pengel *et al*[18]	232	RMDQ	PSFS	
Von Korff *et al*[21]	227	RMDQ	CPG	
Witt *et al*[22]	2229	PCS	FFbHR	
YACBAC[19]	206	PCS	ODI	

RMDQ indicates Roland-Morris Disability Questionnaire; CPG, Chronic Pain Grade disability score; PCS, Physical Component Summary of 12- or 36-Item Short Form Health Survey; FFbHR, Hannover Functional Ability Questionnaire; PDI, Pain Disability Index; PSFS, Patient Specific Functional Scale; ODI, Oswestry Disability Index.

(range, 0.27–0.31).[14,16,20] However, the level of agreement was never more than fair.

DISCUSSION

A number of patient-reported outcomes are commonly used to measure disability in randomized controlled trials of interventions for LBP, with little consensus as to a preferred measure. High correlation and similar responsiveness are necessary conditions for outcome measures to be comparable enough that one could be used to predict another so that they could be pooled, for example, in a meta-analysis. Our work reported here has used data from 11 randomized controlled clinical trials from a large pooled data set of individual participant data to assess the extent to which these criteria are satisfied for pairs of measures.

We found that for each pair of outcome measures, correlation and similarity in responsiveness were low. In all cases, these were below the threshold set to consider it feasible to convert between the outcome measures or combine them in an individual participant data meta-analysis.

A strength of our work has been the use of individual participant data from a large number of trials using different combinations of outcome measures. This has enabled us to conduct 21 within-trial comparisons between pairs of 7 different outcome measures, with some pairwise comparisons repeated on the basis of data from a number of different trials. We are not aware of any similar comparison conducted on this scale. A weakness of this study has been the small sample size for some trials. Because comparisons were conducted within trial, this means that some estimates may not be precise. A further weakness is that although all but one of the outcome measures are ordinal, we have treated them as continuous in our analysis. Specifically, the Pearson correlation coefficient requires that the variables in question are continuous. Although it is common practice for ordinal variables with a large number of points on their scales to be treated as though they are continuous, some authors consider this to be a mistreatment of such variables,[28] but we felt that applying a more complicated method would have been an attempt to account for a richer structure than was actually present.

The lack of agreement between different outcome measures taken on the same patient is probably due to the fact that the questionnaires measure disability in different ways. Indeed, it would be hard to justify the time-consuming process of creating a new questionnaire if the end result were to be very similar to another already-existing questionnaire.

Data from several trials including the same pairs of measures enabled the correlation coefficients and κ statistics between a pair of measures to be obtained from different data sets and compared. Of particular note is the correlation between PCS of SF-12/36 and FFbHR, which were about 0.58 and were very similar across the 3 trials. This may not be surprising because these 3 trials were conducted by the same group, tested the same intervention (acupuncture), and recruited from similar German populations.[12,14,22]

On the contrary, the correlations between CPG and RMDQ ranged from 0.21 to 0.47. There were slight variations in the version of CPG instrument that was used in these trials. The UK BEAM[20] and BeST[16] trials used the modified version of CPG, which asked patients how much their back trouble had been interfering with their daily activities in the last 1 month, whereas in the Von Korff trial[21] the time period was the last 3 months. This may explain the weaker association between

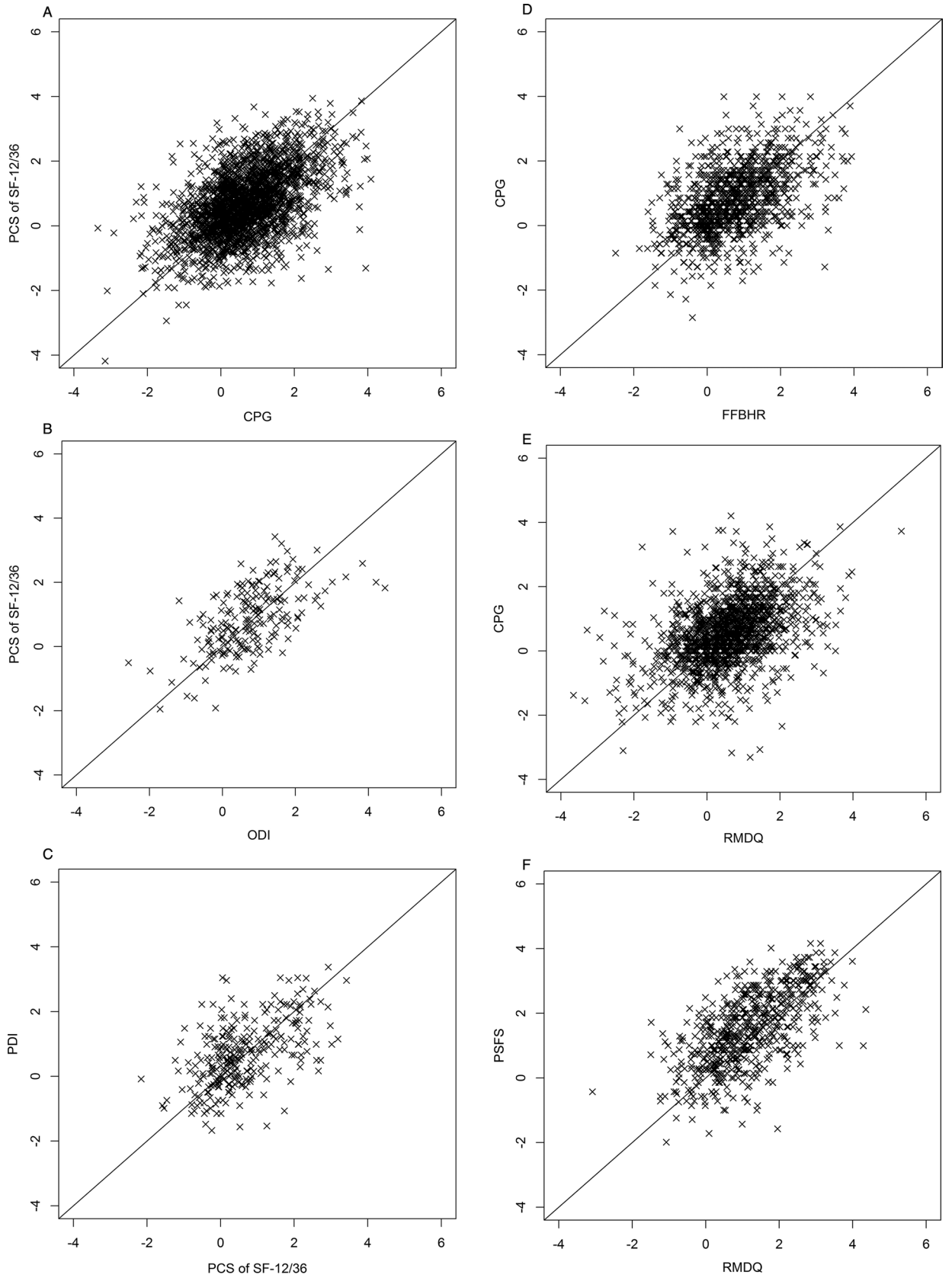

Figure 1. Scatterplots of standardized change scores of outcome measures: (**A**) PCS against CPG; (**B**) PCS against ODI; (**C**) PDI against PCS; (**D**) CPG against FFbHR; (**E**) CPG against RMDQ; and (**F**) PSFS against RMDQ. PCS indicates Physical Component Summary of the 12- or 36-Item Short Form Health Survey; CPG, Chronic Pain Grade disability score; ODI, Oswestry Disability Index; PDI, Pain Disability Index; FFbHR, Hannover Functional Ability Questionnaire; RMDQ, Roland-Morris Disability Questionnaire; and PSFS, Patient Specific Functional Scale.

TABLE 2. Pearson Correlation and Cohen's κ for Each Pair of Instruments

Outcome Measure 1	Outcome Measure 2	Trial	Pearson Correlation	Cohen's κ
CPG	RMDQ	UK BEAM[20]	0.47	0.27
		BeST[16]	0.44	0.22
		Von Korff *et al*[21]	0.21	0.12
CPG	FFbHR	Haake *et al*[14]	0.48	0.25
PCS	RMDQ	UK BEAM[20]	0.51	0.33
		BeST[16]	0.38	0.17
		HULLEXPROB[13]	0.45	0.29
		Macedo *et al*[17]	0.52	0.27
PCS	CPG	UK BEAM[20]	0.56	0.31
		BeST[16]	0.41	0.27
		Haake *et al*[14]	0.49	0.27
PCS	FFbHR	Brinkhaus *et al*[12]	0.59	0.30
		Haake *et al*[14]	0.58	0.29
		Witt *et al*[22]	0.59	0.27
PCS	PSFS	Macedo *et al*[17]	0.36	0.17
PCS	ODI	YACBAC[19]	0.60	0.28
RMDQ	PSFS	Hancock *et al*[15]	0.70	0.38
		Macedo *et al*[17]	0.40	0.26
		Pengel *et al*[18]	0.53	0.18
PDI	FFbHR	Brinkhaus *et al*[12]	0.55	0.32
PDI	PCS	Brinkhaus *et al*[12]	0.54	0.31

CPG indicates Chronic Pain Grade disability score; RMDQ, Roland-Morris Disability Questionnaire; FFbHR, Hannover Functional Ability Questionnaire; PCS, Physical Component Summary of 12- or 36-Item Short Form Health Survey; PSFS, Patient Specific Functional Scale; ODI, Oswestry Disability Index; PDI, Pain Disability Index.

CPG and RMDQ in the Von Korff trial because the RMDQ was designed to measure if their back pain had been interfering with their daily activities on the day they were evaluated.

Our comparison has been based on the change from baseline to short-term follow-up (2–3 mo postrandomization). This time point was chosen because data were available in all trials. Nearly all of the improvement from baseline seen in intervention and control arms of randomized controlled trials of LBP is seen by around 3 months.[29] Thus, there would be little advantage in additionally considering long-term outcomes. Many of the trials also had mid-term (6 mo) and long-term (1 yr) follow-up. We performed the same analyses on these data, and the results were similar.

CONCLUSION

We used data from 11 randomized clinical trials (n = 6089 patients) in LBP to compare the following 7 measurements: RMDQ, CPG disability score, PCS of the SF-12/36, PSFS, PDI, ODI, and FFbHR.

Pairs of measures were found to be positively correlated, but correlations were mostly less than the 0.5 we specified *a priori*, with only 1 pair of outcomes in 1 trial having a correlation of more than 0.6. Correlations between the SF-12/36 PCS and other PROMs, namely, CPG, FFbHR, ODI, and PDI, were moderately positive (between 0.40 and 0.60). We note, however, that we set a less rigorous cutoff than other investigators. However, all κ statistics, including those comparing these pairs of outcomes, were less than 0.4. In no cases were the criteria we had set for acceptability of pooling measures satisfied.

These data do not support the notion that crosswalking between scores on different LBP outcomes measures is justifiable. Future researchers need to settle on a single outcome measure for trials of back pain treatments. Adoption of the core set suggested by the National Institutes of Health Task Force is an important step that will allow a better understanding of the differences and similarities from results from different studies.[2]

We conclude that the lack of agreement between different outcome measures means that pooling of data on these different disability measurements in a meta-analysis is not recommended.

➢ Key Points

- ❑ Changes from baseline to short-term follow-up of 6 back pain–related disability outcomes and 1 generic-based health-related quality-of-life outcome were compared on the basis of data from 11 trials of therapist-delivered interventions for chronic LBP.
- ❑ Correlations between the measures and Cohen's κ statistics comparing the number of patients for whom the change scores were greater than, equal to, or less than zero were calculated. Correlations and κ statistics were found to be low.
- ❑ It is not recommended that data on the different patient-reported outcomes studied should be pooled in a meta-analysis.

Supplemental digital content is available for this article. Direct URL citation appears in the printed text and is provided in the HTML and PDF versions of this article on the journal's Web site (www.spinejournal.com).

References

1. Müller U, Duetz MS, Roeder C, et al. Condition-specific outcome measures for low back pain, part I: validation. *Eur. Spine J* 2004;13:301–13.
2. Deyo RA, Dworkin SF, Amtmann D, et al. Report of the NIH Task Force on Research Standards for Chronic Low Back Pain. *J Pain* 2014;15:569–85.
3. Patel S, Hee SW, Mistry D, et al. Improving outcomes from the treatment of low back pain. *Programme Grants Appl Res*. In press.
4. Ware JE, Kosinski M, Turner-Bowker DM, et al. *How to Score Version 2 of the SF-12 Health Survey (With a Supplement Documenting Version 1)*. Lincoln, RI: QualityMetric Inc; 2002.
5. Ware J, Kosinski M, Dewey J. *How to Score Version 2 of the SF-36 Health Surveyed*. Lincoln, RI: QualityMetric Inc; 2000.
6. Von Korff M, Ormel J, Keefe FJ, et al. Grading the severity of chronic pain. *Pain* 1992;50:133–49.
7. Kohlmann T, Raspe H. [Hannover Functional Questionnaire in ambulatory diagnosis of functional disability caused by backache]. *Rehabilitation (Sttutg)* 1996;35:I–VIII.
8. Fairbank JC, Couper J, Davies JB, et al. The Oswestry Low Back Pain Disability Questionnaire. *Physiotherapy* 1980;66:271–3.
9. Tait RC, Chibnall JT, Krause S. The Pain Disability Index: psychometric properties. *Pain* 1990;40:171–82.
10. Stratford P, Gill C, Westaway M, et al. Assessing disability and change on individual patients: a report of a patient specific measure. *Physiother Can* 1995;47:258–63.
11. Roland M, Morris R. A study of the natural history of back pain, part I: development of a reliable and sensitive measure of disability in low-back pain. *Spine* 1983;8:141–4.
12. Brinkhaus B, Witt CM, Jena S, et al. Acupuncture in patients with chronic low back pain: a randomized controlled trial. *Arch Intern Med*. 2006;166:450–7.
13. Carr J, Klaber Moffett J, Howarth E, et al. A randomized trial comparing a group exercise programme for back pain patients with individual physiotherapy in a severely deprived area. *Disabil Rehabil* 2005;27:929–37.
14. Haake M, MüLler H, Schade-Brittinger C, et al. German Acupuncture Trials (GERAC) for chronic low back pain: randomized, multicenter, blinded, parallel-group trial with 3 groups. *Arch Intern Med* 2007;167:1892–8.
15. Hancock MJ, Maher CG, Latimer J, et al. Assessment of diclofenac or spinal manipulative therapy, or both, in addition to recommended first-line treatment for acute low back pain: a randomised controlled trial. *Lancet* 370:1638–43.
16. Lamb SE, Hansen Z, Lall R, et al. Group cognitive behavioural treatment for low-back pain in primary care: a randomised controlled trial and cost-effectiveness analysis. *Lancet* 2010;375:916–23.
17. Macedo LG, Latimer J, Maher CG, et al. Effect of motor control exercises *versus* graded activity in patients with chronic nonspecific low back pain: a randomized controlled trial. *Phys Ther* 2012;92:363–77.
18. Pengel LHM, Refshauge KM, Maher CG, et al. Physiotherapist-directed exercise, advice, or both for subacute low back pain: a randomized trial. *Ann Intern Med* 2007;146:787–96.
19. Thomas KJ, MacPherson H, Ratcliffe J, et al. Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain. *Health Technol Assess* 2005;9:iii–iv, ix–x, 1–109.
20. UK BEAM Trial Team. United Kingdom back pain exercise and manipulation (UK BEAM) randomised trial: effectiveness of physical treatments for back pain in primary care. *BMJ* 2004;329:1377.
21. Von Korff M, Balderson BHK, Saunders K, et al. A trial of an activating intervention for chronic back pain in primary care and physical therapy settings. *Pain* 2005;113:323–30.
22. Witt CM, Jena S, Selim D, et al. Pragmatic randomized trial evaluating the clinical and economic effectiveness of acupuncture for chronic low back pain. *Am J Epidemiol* 2006;164:487–96.
23. Puhan M, Soesilo I, Guyatt G, et al. Combining scores from different patient reported outcome measures in meta-analyses: when is it justified? *Health Qual Life Outcomes* 2006;4:94.
24. Cohen J. The significance of a product moment r_s. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988:79–81.
25. Agresti A. *Categorical Data Analysis*. 2nd ed. Hoboken, NJ: Wiley; 2002.
26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
27. R Development Core Team. *R: A Language and Environment for Statistical Computing* [computer program]. 2014. Available at: http://www.R-project.org. Accessed August 18, 2014.
28. Parsons NR. Proportional-odds models for repeated composite and long ordinal outcome scales. *Stat Med* 2013;32:3181–91.
29. Artus M, van der Windt DA, Jordan KP, et al. Low back pain symptoms show a similar pattern of improvement following a wide range of primary care treatments: a systematic review of randomized clinical trials. *Rheumatology (Oxford)* 2010;49:2346–56.