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†, Shilpa Patel, DHealthPsy†

From *Leicester Clinical Trials Unit, University of Leicester, U.K.; and †Warwick Medical School, University of Warwick, U.K.

Collaborators:

Dr Christer Carlsson, Dr Francesca Cecchi, Dr Ninna Dufour, Dr Heinz Endres, Dr Mark Hancock, Professor Elaine Hay, Dr Michael Von Korff, Professor Sarah Lamb, Dr Luciana Macedo, Dr Hugh MacPherson, Professor Chris Maher, Professor Suzanne McDonough, Professor Rob Smeets, Professor David Torgerson, and Professor Claudia Witt who contributed anonymised individual participant data from randomised controlled trials.

Acknowledgment:

We would like to acknowledge the Low Back Pain Repository team for their earlier contributions to this work.

Funding:

This paper summarises independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Grant Reference Number RP-PG-0608-10076). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.
This project benefitted from facilities funded through Birmingham Science City Translational Medicine Clinical Research and Infrastructure Trials Platform, with support from Advantage West Midlands (AWM) and the Wolfson Foundation.

Address correspondence and reprint requests to:

Siew Wan Hee
Division of Health Sciences, Warwick Medical School, The University of Warwick,
Coventry CV4 7AL, United Kingdom.
Email: s.w.hee@warwick.ac.uk
Tel: +44-(0)2476151652
Abstract

Study Design. Retrospective database analysis.

Objective. A range of patient-reported outcomes are used to measure disability due to low back pain. There is not a single back pain disability measurement commonly used in all randomised 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 eleven trials (n = 6089 patients) that had at least two of the following seven measurements: the Roland Morris Disability Questionnaire, the Chronic Pain Grade disability score, the physical component summary of the SF-12/36, the Patient Specific Functional Scale, the Pain Disability Index, the Oswestry Disability Index, and the 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 Cohen’s kappa for the three-level outcome of change score < 0, change score = 0 and change score > 0. It was considered feasible to pool two measures if they were at least moderately correlated (correlation > 0.5) and have at least moderately similar responsiveness (kappa > 0.4).

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

Conclusions. 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.
**Introduction**

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. In order 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 from another instrument. The commonest outcome measures used in RCT for low back pain, and the ones that researchers are most familiar with interpreting, are the Roland Morris Disability Score and the Oswestry Disability Index\(^1\). Being able to standardise 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 Institute of Health task force on research standards in chronic LBP 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 comparable then such comparisons and any meta-analysis using different measures may not be robust.
We have developed a large pooled dataset of individual patient data from 19 trials \( (n = 9328) \) of therapist delivered interventions for LBP which will be a resource for researchers working in the field (report submitted to NIHR)\(^3\). All included trials in this pooled dataset used at least one of the five PROMs designed to measure back pain related disability listed above or included generic-based health-related quality of life instruments such as SF-12\(^4\) or SF-36\(^5\). However, no common instrument was used by all these trials.

In this paper 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 utilised in the research, each to varying degrees. In our dataset we had data available on six PROMs that aim to measure back pain related disability, namely, the Chronic Pain Grade (CPG) disability score which is one of the two domains in the CPG that aims to grade chronic pain status\(^6\), the Hannover Functional Ability Questionnaire (FFbHR)\(^7\), the Oswestry Disability Index (ODI)\(^8\), the Pain Disability Index (PDI)\(^9\), the mean score of three items from the Patient Specific Functional Scale (PSFS)\(^{10}\) and the Roland Morris Disability Questionnaire (RMDQ)\(^{11}\).

Eleven \( (n = 6089) \) of the 19 trials included two 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 or SF-36, recorded at baseline and short term follow-up
(2-3 months post-randomisation). 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 at 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, where the latter is explained below. Correlation was assessed by calculation of Pearson’s 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. 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.

Responsiveness is the ability to detect a change in condition. If two 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 two outcome measures do not have similar responsiveness then combining them in a meta-analysis may introduce heterogeneity which could be falsely attributed to other sources, such as the treatment effect.
Similarity of responsiveness of two outcome measures was examined by categorising the change scores as negative change (change score < 0), no change (change score = 0) or positive change (change score > 0), and calculating Cohen’s kappa from these categorisations.25 A priori we considered kappa > 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. Figures 1(A) – 1(F) show a selection of scatterplots of standardised change scores of these outcome measures. The other scatterplots are available as supplementary materials (see supplementary Figures 1(A) – 1(D)). It is clear from these plots that although instruments appear to be positively correlated, there is a large amount of disagreement between the outcomes.

Correlations and kappa 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 three trials were very similar, all of about 0.5812,14,22. Another three 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\textsuperscript{14,16,20}, and four trials had both SF-12/36 PCS and RMDQ with range 0.38 to 0.52, again similar\textsuperscript{13,16,17,20}. However, correlations between other outcomes were quite widely ranging across trials; between CPG and RMDQ (3 trials, range 0.21 to 0.47)\textsuperscript{16,20,21} and between PSFS and RMDQ (3 trials, range 0.40 and 0.70)\textsuperscript{15,17,18}.

Cohen’s kappa statistics calculated for the three by three table with numbers of patients with positive change, no change or negative change on each outcome were less than 0.4 for all 21 comparisons. Some were similar between trials, namely for PCS and FFbHR (range, 0.27 to 0.30)\textsuperscript{12,14,22} and for PCS and CPG (range 0.27 to 0.31)\textsuperscript{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 randomised 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 eleven randomised controlled clinical trials from a large pooled dataset 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 within-trial comparisons between pairs of seven different outcome measures, with some pairwise comparisons repeated based on data from a number of different trials. We are not aware of any similar comparison conducted on this scale. A weakness has been the small sample sizes for some trials. As 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, Pearson’s 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, 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 kappa statistics between a pair of measures to be obtained from different datasets 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 three trials. This may not be surprising as these three trials were conducted by the same group, tested the same intervention (acupuncture) and recruited from similar German populations.
On the other hand 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 and BeST 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 one month, whereas in the Von Korff trial the time period was the last three months.\textsuperscript{16,20,21} This may explain the weaker association between CPG and RMDQ in the Von Korff trial as the RMDQ was designed to measure if their back pain had been interfering with their daily activities on the day they were assessed.

Our comparison has been based on the change from baseline to short-term follow-up (2-3 months post-randomisation). This time point was chosen as data were available in all trials. Nearly all of the improvement from baseline seen in intervention and control arms of randomised controlled trials of LBP is seen by around three months.\textsuperscript{29} Thus, there would be little additional advantage in additionally considering long term outcomes. Many of the trials also had mid-term (6 months) and long term (1 year) follow-up. We performed the same analyses on these data, and the results were similar.

**Conclusions**

We used data from eleven randomised clinical trials (\(n = 6089\) patients) in LBP to compare the following seven measurements: the Roland Morris Disability Questionnaire, the Chronic Pain Grade disability score, the physical component summary of the SF-12/36, the Patient Specific Functional Scale, the Pain Disability Index, the Oswestry Disability Index, and the Hannover Functional Ability Questionnaire.
Pairs of measures were found to be positively correlated, but correlations were mostly less than the 0.5 we specified a priori, with only one pair of outcomes in one 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 cut-off than other investigators. However, all kappa 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 cross walking 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 NIH taskforce 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.

**Figure legends**

Figure 1 – Scatterplots of standardised 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. Abbreviations: PCS, Physical Component Summary of SF-12 or SF-36; CPG, Chronic Pain Grade disability scale; ODI, Oswestry Disability Index; PDI, Pain Disability Index; FFbHR, Hannover Functional Ability Questionnaire;
RMDQ, Roland Morris Disability Questionnaire; and PSFS, Patient Specific Functional Scale.

Supplemental Figure 1 – Scatterplots of standardised change scores of outcome measures. (A) PCS against FFbHR; (B) PCS against RMDQ; (C) PCS against PSFS; and (D) FFbHR against PDI. Abbreviations: PCS, Physical Component Summary of SF-12 or SF-36; FFbHR, Hannover Functional Ability Questionnaire; RMDQ, Roland Morris Disability Questionnaire; PSFS, Patient Specific Functional Scale; and PDI, Pain Disability Index.
References


## Tables

Table 1. Instruments used and number of patients by trial.

<table>
<thead>
<tr>
<th>Trial</th>
<th>n</th>
<th>Outcome measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK BEAM&lt;sup&gt;20&lt;/sup&gt;</td>
<td>885</td>
<td>RMDQ, CPG, PCS</td>
</tr>
<tr>
<td>BeST&lt;sup&gt;16&lt;/sup&gt;</td>
<td>426</td>
<td>RMDQ, CPG, PCS</td>
</tr>
<tr>
<td>Brinkhaus&lt;sup&gt;12&lt;/sup&gt;</td>
<td>281</td>
<td>PCS, FFbHR, PDI</td>
</tr>
<tr>
<td>Haake&lt;sup&gt;14&lt;/sup&gt;</td>
<td>1110</td>
<td>CPG, FFbHR, PCS</td>
</tr>
<tr>
<td>Hancock&lt;sup&gt;15&lt;/sup&gt;</td>
<td>235</td>
<td>RMDQ, PSFS</td>
</tr>
<tr>
<td>HULLEXPORB&lt;sup&gt;13&lt;/sup&gt;</td>
<td>203</td>
<td>RMDQ, PCS</td>
</tr>
<tr>
<td>Macedo&lt;sup&gt;17&lt;/sup&gt;</td>
<td>158</td>
<td>RMDQ, PCS, PSFS</td>
</tr>
<tr>
<td>Pengel&lt;sup&gt;18&lt;/sup&gt;</td>
<td>232</td>
<td>RMDQ, PSFS</td>
</tr>
<tr>
<td>Von Korff&lt;sup&gt;20&lt;/sup&gt;</td>
<td>227</td>
<td>RMDQ, CPG</td>
</tr>
<tr>
<td>Witt&lt;sup&gt;22&lt;/sup&gt;</td>
<td>2229</td>
<td>PCS, FFbHR</td>
</tr>
<tr>
<td>YACBAC&lt;sup&gt;19&lt;/sup&gt;</td>
<td>206</td>
<td>PCS, ODI</td>
</tr>
</tbody>
</table>

Abbreviations: RMDQ, Roland Morris Disability Questionnaire; CPG, Chronic Pain Grade disability scale; PCS, Physical Component Summary of SF-12 or SF-36; FFbHR, Hannover Functional Ability Questionnaire; PDI, Pain Disability Index; PSFS, Patient Specific Functional Scale; and ODI, Oswestry Disability Index.
Table 2. Pearson correlation and Cohen’s kappa for each pair of instruments.

<table>
<thead>
<tr>
<th>Outcome measure 1</th>
<th>Outcome measure 2</th>
<th>Trial</th>
<th>Pearson correlation</th>
<th>Cohen’s kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPG</td>
<td>RMDQ</td>
<td>UK BEAM&lt;sup&gt;20&lt;/sup&gt;</td>
<td>0.47</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BeST&lt;sup&gt;16&lt;/sup&gt;</td>
<td>0.44</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Von Korff&lt;sup&gt;21&lt;/sup&gt;</td>
<td>0.21</td>
<td>0.12</td>
</tr>
<tr>
<td>CPG</td>
<td>FFbHR</td>
<td>Haake&lt;sup&gt;14&lt;/sup&gt;</td>
<td>0.48</td>
<td>0.25</td>
</tr>
<tr>
<td>PCS</td>
<td>RMDQ</td>
<td>UK BEAM&lt;sup&gt;20&lt;/sup&gt;</td>
<td>0.51</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BeST&lt;sup&gt;16&lt;/sup&gt;</td>
<td>0.38</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HULLEXPROB&lt;sup&gt;13&lt;/sup&gt;</td>
<td>0.45</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Macedo&lt;sup&gt;17&lt;/sup&gt;</td>
<td>0.52</td>
<td>0.27</td>
</tr>
<tr>
<td>PCS</td>
<td>CPG</td>
<td>UK BEAM&lt;sup&gt;20&lt;/sup&gt;</td>
<td>0.56</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BeST&lt;sup&gt;16&lt;/sup&gt;</td>
<td>0.41</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Haake&lt;sup&gt;14&lt;/sup&gt;</td>
<td>0.49</td>
<td>0.27</td>
</tr>
<tr>
<td>PCS</td>
<td>FFbHR</td>
<td>Brinkhaus&lt;sup&gt;12&lt;/sup&gt;</td>
<td>0.59</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Haake&lt;sup&gt;14&lt;/sup&gt;</td>
<td>0.58</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Witt&lt;sup&gt;22&lt;/sup&gt;</td>
<td>0.59</td>
<td>0.27</td>
</tr>
<tr>
<td>PCS</td>
<td>PSFS</td>
<td>Macedo&lt;sup&gt;17&lt;/sup&gt;</td>
<td>0.36</td>
<td>0.17</td>
</tr>
<tr>
<td>PCS</td>
<td>ODI</td>
<td>YACBAC&lt;sup&gt;19&lt;/sup&gt;</td>
<td>0.60</td>
<td>0.28</td>
</tr>
<tr>
<td>RMDQ</td>
<td>PSFS</td>
<td>Hancock&lt;sup&gt;15&lt;/sup&gt;</td>
<td>0.70</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Macedo&lt;sup&gt;17&lt;/sup&gt;</td>
<td>0.40</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pengel&lt;sup&gt;18&lt;/sup&gt;</td>
<td>0.53</td>
<td>0.18</td>
</tr>
<tr>
<td>PDI</td>
<td>FFbHR</td>
<td>Brinkhaus&lt;sup&gt;12&lt;/sup&gt;</td>
<td>0.55</td>
<td>0.32</td>
</tr>
<tr>
<td>PDI</td>
<td>PCS</td>
<td>Brinkhaus&lt;sup&gt;12&lt;/sup&gt;</td>
<td>0.54</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Abbreviations: CPG, Chronic Pain Grade disability scale; RMDQ, Roland Morris Disability Questionnaire; FFbHR, Hannover Functional Ability Questionnaire; PCS, Physical Component Summary of SF-12 or SF-36; PSFS, Patient Specific Functional Scale; ODI, Oswestry Disability Index; and PDI, Pain Disability Index.