A Systematic Review of Outcome Measures Use, Analytical Approaches, Reporting Methods, and Publication Volume by Year in Low Back Pain Trials Published between 1980 and 2012: Respice, adspice, et prospice

Robert Froud1,-, Shilpa Patel2, Dévan Rajendran3,*, Philip Bright3, Tom Bjørkli1, Rachelle Buchbinder4, Sandra Eldridge5, Martin Underwood2

1 Department of Health Sciences, Kristiania University College, Oslo, Norway, 2 Warwick Medical School, University of Warwick, Coventry, United Kingdom, 3 European School of Osteopathy, Maidstone, Kent, United Kingdom, 4 Monash Department of Clinical Epidemiology, Cabrini Institute and Department of Epidemiology and Preventive Medicine, School of Public Health and Preventative Medicine, Monash University, Malvern, Victoria, Australia, 5 Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom

These authors contributed equally to this work.

* r.froud@warwick.ac.uk

Abstract

Background

Increasing patient-reported outcome measures in the 1980s and 1990s led to the development of recommendations at the turn of the millennium for standardising outcome measures in non-specific low back pain (LBP) trials. Whether these recommendations impacted use is unclear. Previous work has examined citation counts, but actual use and change over time, has not been explored. Since 2011, there has been some consensus on the optimal methods for reporting back pain trial outcomes. We explored reporting practice, outcome measure use, and publications over time.

Methods

We performed a systematic review of LBP trials, searching the European Guidelines for the management of LBP, extending the search to 2012. We abstracted data on publications by year, outcome measure use, analytical approach, and approaches taken to reporting trials outcomes. Data were analysed using descriptive statistics and regression analyses.

Results

We included 401 trials. The number of published trials per year has increased by a factor of 4.5 from 5.4 (1980–1999) to 24.4 (2000–2012). The most commonly used outcome measures have been the Visual Analogue Scale for pain intensity, which has slowly increased in use since 1980/81 from 20% to 60% of trials by 2012, and the Roland-Morris Disability
design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist, apart from RF and MU who have read the journal’s policy and the authors of this manuscript have the following competing interests: They are shareholders and directors of Clinvivo Ltd, a company that provides electronic outcome measurement services. MU declares that he is lead author on two of the 401 trials included in this systematic review. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

Questionnaire, which rose to 55% in 2002/2003, and then fell back to 28% by 2012. Most trialists (85%) report between-group mean differences. Few (8%) report individual improvements, and some (4%) report only within-group analyses. Student’s t-test, ANOVA, and ANCOVA regression, or mixed models, were the most common approaches to analysis.

Conclusions
Recommendations for standardising outcomes may have had a limited or inconsistent effect on practice. Since the research community is again considering outcome measures and modifying recommendations, groups offering recommendations should be cognisant that better ways of generating trialist buy-in may be required in order for their recommendations to have impact.

Introduction
Patient-reported outcome measures (PROMs) are outcomes that are reported by patients, rather than being objectively assessed or involving third-party (e.g. clinician) judgement. Throughout the 1980s and 1990s, multiple back-specific PROMs were developed and began to dominate as outcome measures used in non-specific low back pain (nsLBP) trials. Between 1998 and 2000 recommendations were made to standardise outcome measure use to facilitate cross-trial comparisons, pooling of data, and encourage scale familiarity. [1–3] More recently, researchers and clinicians have again begun to question whether the right things are being measured and there have been calls to review the outcome measures used in trials. [4–8] It is not clear whether the millennial recommendations for standardisation had an effect on practice. Previous studies have explored the number of times back-specific measures have been cited, but not actual use; also, trends of use over time have not hitherto been explored. [9, 10]

Results from trials using PROMs can be reported differently, and this is known to affect clinicians’ interpretations of effectiveness and subsequent decision-making. [11, 12] With this in mind, recommendations for reporting outcomes in back pain trials were made in 2011 and 2014. [13, 14]

We aimed to explore actual use of outcome measures in nsLBP trials, between 1980 and 2012, spanning the publication point of the millennial core-set recommendations. Our objectives were to identify the most commonly used outcomes, and the domain coverage of back-specific PROMs, and to consider whether there was any change in the trajectories of outcome measure use over the period of interest. Additionally, we reviewed the number of publications over time, reporting methods and analytical approaches for the most commonly used outcome measures to provide a baseline assessment of current practice so that any future change may be monitored.

Materials and Methods
Two independent reviewers (RF and SP) identified randomised controlled trials (RCTs) of any intervention for nsLBP published in or after 1980, from COST-B13’s European Guidelines for the Management of Low Back Pain (EGLBP), which included a comprehensive systematic search of all interventions for nsLBP, and the systematic reviews reported in the EGLBP. [15] As the COST-B13 search ended in November 2002, we extended the search to January 1, 2007 using the Cochrane Library, EMBASE, Lilacs, PsycINFO, and PubMed databases, and we
hand-searched the Health Technology Assessment (HTA) journal. We later updated the search to January 1, 2012, using the Cochrane Library, PubMed, and EMBASE database. We omitted PsycINFO, Lilacs, and the HTA journal in this extension due to good cross-coverage from the other databases (see Discussion). An example search strategy is included as a supplementary file (S1 Text). We combined material from the EGLBP and the extended searches, removed duplicates, and short-listed by title and abstract. Full-texts were obtained if the titles and abstract alone contained insufficient information for assessment against the criteria listed in Table 1.

Data abstraction and validation

Two reviewers (either RF, SP, TB, PB, or DR) independently abstracted data on outcome measure use, details of primary outcome analysis, and reporting methods. An outcome was identified as ‘primary’ if (1) the outcome was nominated as such; if no outcome was nominated, or multiple outcomes were nominated, we used (2) the outcome measure on which the sample size calculation was based; if this was not reported, we identified (3) the first outcome measure referred to in the abstract; and if this was not identified, we used (4) the first outcome mentioned in the paper. We identified the primary time point of interest, or used the first follow-up time point in cases when this was not clear. This approach has been taken in other methodological reviews. [16–20] For comparison, a sensitivity analysis of primary outcome measure use for the most commonly used outcome measures was performed using only criterion 1 and 2. Disagreements were resolved through discussion and, if necessary, with arbitration and a third reviewer (RF, SE, or MU).

Using Microsoft Visual Basic 6.3 (Microsoft, Washington) and Microsoft Office Excel 2003 (Microsoft, Washington), we developed a front-end program to assist the data abstraction process and manage abstracted data, which validated entries and provided alerts in the case of missed fields.

For outcome measure identification, we used expert validation of 20% of papers, as has been done in other methodological reviews. [21] Half (i.e. 10%) were selected at random and half (i.e. 10%) were purposively sampled (papers that we anticipated might lead to disagreement) from the papers marked for full-text extraction and then given to an independent reviewer (either SE, RF, SP, TB, PB, or DR), for independent abstraction. Since early in the process we

| Table 1. Inclusion and exclusion criteria. |
|------------------|------------------|
| Order | Inclusion criterion |
| 1 | RCTs of nsLBP |
| 2 | Non-English language reports |
| 3 | Studies that were not RCTs or presented insufficient information for us to determine whether randomisation was used to allocate participants |
| 4 | Reports that self-identified as pilot/feasibility studies |
| 5 | Cross-over designs (because of limited utility in the LBP field) |
| 6 | RCTs with mixed samples (e.g. neck or thoracic pain in addition to LBP), samples of participants with radiating leg pain, or referred pain extending past the knee, or samples including LBP specific pathology (e.g. cancer, ankylosing spondylitis, or disc herniation) or pregnancy |
| 7 | Non-inferiority trials (because of limited utility in the LBP field) |
| 8 | Follow-up studies with no new outcome measures, and multiple publications. In the case of multiple publications, we included the first published article and excluded subsequent publications |

RCT = Randomised controlled trial; nsLBP = non-specific low back pain; LBP = low back pain.

doi:10.1371/journal.pone.0164573.t001
observed good (> 80%) agreement on outcome measurement identification, but inadequate agreement on analytical and reporting methods, full independent abstraction was subsequently used to identify analytical approach and reporting methods used. Disagreements were settled by arbitration involving one of the statisticians (either RF or SE). Quality of included trials was not evaluated, since we were interested in all non-specific back pain trials, regardless of the trials’ methodological quality.

Analysis

To report the proportion of outcome measure use by year, the total number of trials is needed as the denominator. For this reason, we first explored the number of published nsLBP trials by year. Prevalence of outcome measure use and in the case of PROMs, domain of measurement, was then calculated for primary and secondary outcomes, by year. We were particularly interested in PROM use, and we did not differentiate between different types of objective outcome measure use, or clinical judgements.

Methods for reporting the two most commonly used outcome measures and the types of analysis used were summarised using descriptive statistics and graphical methods. For reporting methods, we explored the statistics used to summarise central tendency and variance, graphical forms of representation, and use of tables. For analytical approach, we explored the statistical test used to test the null hypothesis that a between group difference was zero, or the model that was fitted to the data. We used regression analysis to explore the relationship between publications and time, and outcome measure use and time, fitting polynomial terms if relationships were non-linear. Residuals from regression analysis were examined for fit. If data were too heteroschedastic (i.e. variance of an outcome variable was dependent on the value of a predictor variable) for regression modelling, we fitted locally weighted scatter plot smoothing (Lowess) lines, which has the effect of smoothing across erratic data points, so that trends can be more easily visualised. All analyses were performed in Stata, version 12 (Statacorp, Texas). We did not publish a review protocol ahead of undertaking this work.

Results

We identified 7,066 potential articles from EBLBP and electronic databases and following removal of duplicates, titles and abstracts sifting, and full-text inspection against inclusion criteria, included 401 trials (Fig 1). Characteristics of included and excluded trials are detailed with their references as supplementary material (S1 and S2 Tables).

Publications over time

A linear regression model with a quadratic term in year was significant ($\beta_{year} = -166.187$, $P = 0.002, \beta_{year^2} = 0.041, P = 0.002$), and explained 79% of the variance (Fig 2). Some caution must be noted with respect to interpretation, since as can be seen from Fig 2, these data are slightly heteroschedastic. The number of publications increases after the millennium; the average number of publications per year from 1980 to 1999 is 5.4 and from 2000 to 2012 is 24.4.

Outcome measure use

Authors explicitly declared a primary outcome measure in 50% ($n = 201$) of trials. In trials that did not declare one, a primary outcome measure could be identified from a sample size calculation in 20% ($n = 40$) of trials. We identified the primary as the first outcome measure mentioned in the abstract, or in the paper, in the remaining 40% ($n = 161$).
The most commonly used PROMs were the Visual Analogue Scale for measuring Pain intensity (VAS-P), [22], and the Roland Morris Disability Questionnaire (RMDQ), [23] Oswestry Disability Index (ODI), [24] Numerical Rating Scale for measuring pain intensity (NRS), [25, 26] and patient-rated global assessment of improvement (i.e. a health transition question (TQ) [27]) (Table 2 and Fig 3).

We note that we class the VAS-P and NRS as back-specific, since the wording of these instruments, when used in back pain trials, typically refers specifically to a person’s back pain. In total, there were 258 different PROMs used across the trials studied within the time period.

Sensitivity analyses of primary outcome measure use within only trials that either explicitly declared the outcome measure as primary, or used it for a sample size calculation, revealed similar estimates of primary outcome measure use in across outcome measures, with the possible exception of the RMDQ (Table 2).

Apart from PROMs, objectively assessed outcome measures were also common (n = 130), particularly as secondary outcome measures, as were medication and medical services consumption (n = 66), and subjective clinical examinations (n = 19).
Pain and disability were the most commonly measured domains (Fig 4). Some outcomes, for example, adverse events, or adherence, have only ever been measured as secondary outcomes. We note that it is possible for the usage to exceed the number of included trials, which is due to some trials using more than one outcome measure to measure within these domains. The VAS-P has had a slowly increasing usage as an outcome measure in nsLBP trials, increasing from from 20% in slowly increased in use since 1980/81 to 60% of trials by 2012. There is some suggestion of a rise and fall in the use of the RMDQ, which peaks at 55% in 2002/2003. Use of the ODI has steadily increased. These data were too heteroschedastic for regression analysis and we present the data graphically with a Lowess smoother (Fig 5). We note the use of a large smoothing bandwidth to easily visualise trends; while the RMDQ

![Figure 2: Number of published non-specific low back pain trials by publication year between 1980 and 2012. The figure shows the increase in the number of published non-specific low back pain trials by year of publication and change in publication rate over time. A Lowess smoother is fitted to these data.](https://doi.org/10.1371/journal.pone.0164573.g002)

### Table 2. The most common back-specific PROMs: Frequency of use.

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Primary outcome</th>
<th>Secondary outcome</th>
<th>Total use</th>
<th>Primary use (%)</th>
<th>Sensitivity analysis (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual Analogue Scale of back pain intensity [22]</td>
<td>119</td>
<td>86</td>
<td>205</td>
<td>29.7</td>
<td>28.2</td>
</tr>
<tr>
<td>Roland-Morris Disability Questionnaire [23]</td>
<td>58</td>
<td>94</td>
<td>152</td>
<td>14.5</td>
<td>20.7</td>
</tr>
<tr>
<td>Oswestry Disability Index [24]</td>
<td>36</td>
<td>84</td>
<td>120</td>
<td>9.0</td>
<td>10.4</td>
</tr>
<tr>
<td>Pain Intensity Numerical Rating Scale [25, 26]</td>
<td>37</td>
<td>39</td>
<td>76</td>
<td>9.2</td>
<td>8.7</td>
</tr>
<tr>
<td>Patient Rated Global Assessment (TQ) [27]*</td>
<td>10</td>
<td>53</td>
<td>63</td>
<td>2.5</td>
<td>2.5</td>
</tr>
</tbody>
</table>

PROM = Patient-reported outcome measure  
* Example reference only  
† Of proportional use as primary outcome measure  

doi:10.1371/journal.pone.0164573.t002
smoother intersects the x-axis at 1982, the year of its publication, it was not recorded as used in any included trials until 1987. As such, these Lowess lines should be interpreted as a general impression of trend only. We have included as an additional file (S1 Fig) the plot with a smoother of half the bandwidth, to depict more sensitive trend-lines.

**Reporting methods and analytical approach**

For the two most commonly used outcomes, the VAS-P and the RMDQ, reporting methods include describing mean changes (within groups) or mean differences (between groups), P-values for changes or differences, and standard deviations (Table 3). Confidence intervals were only provided in around 40% of trials. Few authors reported individual improvements. The t test has been the most common analytical approach, followed by regression analyses. The most common covariates used in adjusted models were baseline score, sex, age, study centre, and episode duration, respectively. In 4% of trials utilising the VAS-P or RMDQ, authors present within-group analyses only.

**Discussion**

We discuss the results and consider their implications in three sections. We consider these data and past recommendations for core sets of outcome measures in back pain trials (Respice), the current situation and emerging recommendations for core sets (Adspice), and implications for the future of outcome measure use in back pain trials (Prospice).
The results show that the most commonly used domains in back-specific outcome measures over the whole time-period of interest are domains and PROMs that were included in the millennial recommendations. The millennial recommendations of Deyo et al in 1998, updated by Bombardier et al in 2000, recommend measurement in the domains of pain, function, wellbeing, disability, and work disability and both recommended using the VAS-P and NRS for measuring pain, and the RMDQ or ODI for measuring function (disability). [1, 2] The World Health Organisation made back-specific recommendations to measure pain, disability, and depression, in 1999, recommending the VAS-P and the ODI be used as primary outcome measures in trials. [3] International Classifications of functioning (ICF) categories were later proposed, recommending 78 (comprehensive) or 35 (brief) domains of measurement for LBP in 2004. [28] The brief set, intended in particular for use in clinical studies, included body functions, structures, activities and participation and environmental factors. The recommendations were criticised for having incomplete coverage. [29]

If the millennial recommendations for measurement instruments had impacted practice then notwithstanding the time it takes for recommendations to be implemented, one might expect relative increases in proportional use trajectories sometime after the publication of the recommendations at the turn of the millennium. There is possibly a post-millennial increase in trajectory in the VAS-P and ODI; there may be a decrease in use of the RMDQ. As these data represent a population of trials over the time period, inferential statistics are unnecessary, and the millennial recommendations appear to have had a limited or inconsistent impact on actual use.
Adspice

Towards the end of the 1980–2012 period of interest, researchers suggested outcome measures are inadequate and that their reassessment needed to be prioritised. [4, 7] A second wave of recommendations is now emerging. In 2011, Buchbinder et al identified several domains of measurement not covered by existing core sets, including loss of independence, worry about the future, and negative or discriminatory actions by others. [30] In 2012, Hush et al recommended adding the Patient Generated Index (PGI) and a global back pain recovery scale to the core set, specifically to measure recovery. [31] In 2014, Froud et al showed social components were not adequately represented in measurement. [32] The results in the current study highlight the absence of measurement in social domains; although social factors are covered to some extent within the domain of health-related quality of life.

In 2014, a National Institutes of Health (NIH) Task Force recommended using Patient Reported Outcomes Measurement Information System (PROMIS) measures as a minimum dataset in all NIH-funded LBP research; permitting the RMDQ to substitute the PROMIS physical activity items if more extensive ‘legacy measures’ are required. [14] In 2015, Froud et al extended the Hush recommendation for patient-centric instruments, such as the PGI and TQ, for trial outcome measurement more generally, after demonstrating that people may not attend to thinking about their back pain when completing the RMDQ and that changes in RMDQ score can be independent of changes in back pain. [8, 31] In 2015, Chiarotto et al recommended three domains for inclusion in a core set, including physical functioning, pain intensity and health-related quality of life. [33]
Thus, it may currently be difficult from a trialist’s perspective to decide upon appropriate primary and secondary outcomes in the design-phase of nsLBP trials. We suggest that this requires careful future attention from groups making core set recommendations, and we discuss this in the next section. In the meantime, given the widespread use shown in the results, there may be some value in using either the VAS-P or RMDQ as an outcome measure in nsLBP trials as this facilitates the greatest cross-comparison. Choices of secondary outcomes might include some of the more patient-relevant and coverage-improving domains, such as the PGI and TQ. The results showed that some domains (for example, adverse events and adherence) are only ever measured as secondary outcomes. In our view this is reasonable (unless the research question relates to adverse events) since the primary focus across most trials and core set recommendations has emphasis on pain and function.

Table 3. Reporting methods and statistical analysis: Prevalence of use.

<table>
<thead>
<tr>
<th>Details</th>
<th>VAS-P(%)</th>
<th>RMDQ(%)</th>
<th>Totals(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td>205 (100)</td>
<td>152 (100)</td>
<td>357 (100)</td>
</tr>
<tr>
<td><strong>Reporting method</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean or mean difference</td>
<td>179 (87)</td>
<td>126 (83)</td>
<td>300 (85)</td>
</tr>
<tr>
<td>P-values</td>
<td>162 (79)</td>
<td>109 (72)</td>
<td>271 (76)</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>134 (65)</td>
<td>98 (64)</td>
<td>232 (65)</td>
</tr>
<tr>
<td>Median</td>
<td>22 (11)</td>
<td>14 (9)</td>
<td>36 (10)</td>
</tr>
<tr>
<td>Range or IQR</td>
<td>20 (10)</td>
<td>13 (9)</td>
<td>33 (9)</td>
</tr>
<tr>
<td>Standard error</td>
<td>18 (9)</td>
<td>18 (12)</td>
<td>36 (10)</td>
</tr>
<tr>
<td>Confidence intervals</td>
<td>62 (30)</td>
<td>76 (50)</td>
<td>138 (39)</td>
</tr>
<tr>
<td>Number/proportion improved*</td>
<td>20 (10)</td>
<td>8 (5)</td>
<td>28 (8)</td>
</tr>
<tr>
<td>Number needed to treat</td>
<td>2 (1)</td>
<td>5 (3)</td>
<td>7 (2)</td>
</tr>
<tr>
<td>Odds ratio (improvement)</td>
<td>3 (1)</td>
<td>9 (6)</td>
<td>12 (3)</td>
</tr>
<tr>
<td>Relative risk (improvement)</td>
<td>1 (0)</td>
<td>6 (4)</td>
<td>7 (2)</td>
</tr>
<tr>
<td>Percentage change score</td>
<td>25 (12)</td>
<td>15 (10)</td>
<td>40 (11)</td>
</tr>
<tr>
<td>Standardised mean difference</td>
<td>9 (4)</td>
<td>11 (7)</td>
<td>20 (6)</td>
</tr>
</tbody>
</table>

**Graphs and tables**

<table>
<thead>
<tr>
<th>Details</th>
<th>VAS-P(%)</th>
<th>RMDQ(%)</th>
<th>Totals(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage</td>
<td>174 (85)</td>
<td>119 (78)</td>
<td>293 (82)</td>
</tr>
<tr>
<td>Reporting method</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Table</td>
<td>174 (85)</td>
<td>119 (78)</td>
<td>293 (82)</td>
</tr>
<tr>
<td>Line chart</td>
<td>60 (29)</td>
<td>39 (26)</td>
<td>99 (28)</td>
</tr>
<tr>
<td>Bar chart</td>
<td>29 (14)</td>
<td>9 (6)</td>
<td>38 (11)</td>
</tr>
<tr>
<td>Other</td>
<td>8 (4)</td>
<td>9 (6)</td>
<td>17 (5)</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANCOVA regression / mixed model</td>
<td>53 (25)</td>
<td>64 (42)</td>
<td>117 (33)</td>
</tr>
<tr>
<td>ANOVA/MANOVA regression / mixed model</td>
<td>74 (36)</td>
<td>39 (26)</td>
<td>113 (32)</td>
</tr>
<tr>
<td>t-test</td>
<td>93 (45)</td>
<td>43 (28)</td>
<td>136 (38)</td>
</tr>
<tr>
<td>Non-parametric test†</td>
<td>68 (33)</td>
<td>37 (24)</td>
<td>105 (29)</td>
</tr>
<tr>
<td>Other</td>
<td>18 (9)</td>
<td>9 (6)</td>
<td>27 (8)</td>
</tr>
<tr>
<td>Within-group analysis only</td>
<td>11 (5)</td>
<td>3 (2)</td>
<td>14 (4)</td>
</tr>
<tr>
<td>Test not described</td>
<td>10 (5)</td>
<td>10 (7)</td>
<td>20 (6)</td>
</tr>
</tbody>
</table>

VAS = Visual analogue scale
RMDQ = Roland-Morris Disability Questionnaire
IQR = Interquartile range
* Where ‘improvement’ is defined by the change in score of a specified magnitude
† e.g. Wilcoxon, Friedman’s, Mann Whitney U, or Kruskal Wallis

doi:10.1371/journal.pone.0164573.t003
Prospice

There is a risk that overlaps or contractions between and within the millennial and current recommendations may continue to give a discordant message to trialists. For example, the millennial Deyo/Bombardier recommendations suggested use of VAS-P or NRS (pain) and RMDQ or ODI (function) whereas the World Health Organisation recommended only VAS-P and ODI for measuring these domains. Currently, researchers are taking slightly different directions with their recommendations for the future (see above). Future divergences in recommendations may benefit from clear justification and argument. One driving issue may be that researchers disagree upon fundamental clinical measurement properties of instruments. For example, although the RMDQ is the second most commonly used outcome in nsLBP trials, it has gained both criticism and support of its clinimetric properties. [8, 27, 34] An underlying issue here may be disagreement on and heterogeneity in how key clinimetric assessments, such as test-retest studies and responsiveness studies, are conducted. [35–37] The Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) checklist may go some way towards helping with this, and it may be that new instruments, with a focus on performance, coverage, and patient-relevance need to be developed prior to future recommendations being made. [38]

The historical data in the current study suggest that the current wave of recommendations may be set to have limited or inconsistent impact on practice. In future, consideration may need to be given to mechanisms for maximising impact of recommendations and affecting change. We suggest that to maximise impact it may be worth considering (1) the consistency of recommendations, which might be facilitated through (2) inclusivity—a more widespread and comprehensive collaboration between methodologists working on outcome measures, along with clinimetric/psychometric experts, and domain-specific experts where required, in addition to clinical experts. One approach to achieving this may be Delphi technique, which is supported by the Core Outcome Measures in Effectiveness Trials (COMET) initiative, with consideration given to fully reporting panel disagreement and lack of consensus. [39] Finally, generating sufficient (3) participation or ‘buy-in’ from trialists with acquisition of support, and a clear and implementable transition strategy to manages the large lag-times from trial conception to publication, may help to avoid fragmentation in outcome measure use. It may be that amendments to well-adopted check-lists would also help. The Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) check list strongly encourages trialists to explain the clinical relevance of chosen outcomes in trial protocols (item 12). [40] However, for trial reporting, the Consolidated Standards of Reporting Trials (CONSORT) statement (in particular item 6a), might be modified to encourage authors to discuss any divergences from recommended core sets or justify outcome measurement selection, which over time, may help to homogenise outcome measurement across similar trials. [41] Additionally, it may be that public and private funders, as stakeholders with a vital interest in maximising the use of clinical trial data, are well-placed to encourage trial teams to carefully consider outcome selection prior to their awarding a grant for a trial.

Reporting methods and analysis

Specific recommendations for including individual improvements when reporting the outcomes of back pain trials were not made until 2011. [13] These were reiterated by the NIH task force in 2014. [14] The 1980 to 2012 data in the current study are not consistent with the recommendations, but the results of the current study will allow future monitoring of the impact of these or future recommendations.
It is encouraging that the majority of authors report between-group differences and a reasonable proportion (33%) report adjusted models. Adjusted models improve the precision of estimates by taking account of imbalances that exist between groups notwithstanding randomisation. [42] Few authors have used inappropriate analytical methods, such as analysing only within-group changes; meaning it is not possible to differentiate treatment effect from regression to the mean.

Comparisons to existing research

Litcher-Kelly et al report, from a systematic review of musculoskeletal clinical trials, that the most frequently used instruments were the VAS (60%), and the NRS (12%), with an RMDQ prevalence of 14%. [43] These data are not directly comparable to ours since the study population and publication years differ. In 2004, Müller et al identified 84 different back-specific PROMs and showed that the Oswestry Disability Questionnaire and the RMDQ were most commonly cited. [9] In 2011, Chapman et al also reviewed which outcome measures had been cited in back pain trials between 2006 and 2011. [10] They found that the most cited functional measure was the ODI and that the most cited pain measure was the NRS. Counting citations may not reflect actual use. In some cases, citations may not be given and for long-running outcome measures, such as the VAS, references may not be uniform. For example, Huskisson is commonly credited with developing the VAS, in 1974, but he is not uniformly credited with it. [22] There is evidence that the VAS was being used at least as far back as 1921. [44] Also, some citations will reference validatory work.

In other fields, Araújo et al have shown that recommendation of core sets of measurement in gout have not impacted practice. [45] Page et al showed in 2015 that shoulder pain trials suffer from having no core measurement sets recommended. [46] Relatively the situation in back pain research may be similar or better, respectively.

Nuovo et al examined the prevalence of reporting of absolute risk reduction (ARR) and number needed to treat (NNT) in RCTs published in five mainstream medical journals. [47] Evaluating publications in 1989, 1992, 1995, and 1998 they found that amongst 359 articles that ARR was reported in 18 (5.0%) reports, and that NNT was used in eight (2.2%). This is similar to our study, in which 8% reported the number or proportion of improvements (a coding in our review which subsumed ARR), and 2% of trials reported NNT.

Strengths and limitations

A strength of our research is provision of prevalence data based on actual use, rather than using citations as a proxy. Other authors have reviewed what outcome measures exist for LBP, but have not estimated prevalence of use. [48–50]

We acknowledge that it is preferable to conduct a systematic review using a singular search pattern. As the EGLBP search was comprehensive, included all interventions, and we had previously used it to identify all nsLBP trials, we reasoned it had good cross-coverage with our later search. Also, we did not search PsycINFO and HTA reports, due to good cross-coverage from the other databases. This fragmented search strategy may be viewed as a limitation. However, in comparison to Castellini et al, who reviewed back pain trials published after 1968 and included 222 trials, we judge our search to have been comprehensive. [51]

We adopted a 20% validation approach rather than full independent reviewer extraction for all variables. We used full independent abstraction when abstracting details of reporting methods and analysis as validation revealed inadequate agreement only on these variables. While full independent abstraction on all variables would have improved validity, as our focus was
methodological and not on estimating a treatment effect, we considered the approach to be reasonable.

Our sensitivity analysis, using only the first two of four criteria for judging a primary outcome measure, showed similar results for estimates of prevalence of use for all but the RMDQ. Few additional (only eight instances) RMDQ primary outcomes were identified using criteria 3 or 4. It may be that, unlike users of other primary outcome measures, those who use the RMDQ as a primary are relatively more likely to explicitly declare it as a primary, or perform a sample size calculation based on detecting a difference in RMDQ score between groups. We note that the difference between estimating prevalence of use using only the first two criteria, or all four criteria, does not change the relative ranking of most common usage of primary outcome measures and that, using either method, the RMDQ is the second most commonly used primary outcome measure.

Conclusions

The Visual Analogue Scale of pain intensity and the Roland Morris Disability Questionnaire have been most commonly used in back pain trials. Recommendations for standardising outcomes may have had a limited or inconsistent effect on practice. Analytical and reporting practice is encouraging, although there is still room for improvement. Research groups planning to make further recommendations on core outcome measures for back pain may have more impact if they consider better ways of generating trialist buy-in.

Supporting Information

S1 Text. Typical search strategy. A typical search strategy used in the systematic review.
(PDF)

S1 Table. Characteristics of included trials. A table showing the characteristics of included trials and their references.
(PDF)

S2 Table. Characteristics of excluded trials. A table showing the characteristics of excluded trials and their references.
(PDF)

S1 Fig. Most commonly used measures over time, standardised by the number of annual publications. The figure shows the most commonly used measures over time, standardised by the number of annual publications, using a Lowess smoother of half the bandwidth of that shown in Fig 5, in case more sensitive trend-lines are preferred.
(TIF)

S1 Checklist. PRISMA checklist. A completed PRISMA checklist for the systematic review.
(PDF)

Acknowledgments

Thanks are due to Hoyskole Kristiania, Warwick Medical School, Cabrini Institute, and Barts and the London Charity for funding this work. Three of the authors (SP, RB, and MU) are employed by one of the funders. RF is employed by two of the funders.

Author Contributions

Conceptualization: RF MU SE RB.
Data curation: RF PB DR TB.

Formal analysis: RF.

Funding acquisition: RF MU SE RB.

Investigation: RF SP TB PB DR SE RB MU.

Methodology: RF MU SE RB.

Project administration: RF.

Software: RF.

Supervision: RF SE RB MU.

Validation: RF SP TB PB DR SE.

Visualization: RF DR MU.

Writing – original draft: RF.

Writing – review & editing: RF SP TB PB DR SE RB MU.

References


44. Hayes M, Patterson D. Experimental development of the graphic rating method. Psychol Bull. 1921; 18(98).


