

Identifying back pain subgroups: developing and applying approaches using individual patient data collected within clinical trials

Shilpa Patel, Siew Wan Hee, Dipesh Mistry, Jake Jordan, Sally Brown, Melina Dritsaki, David R Ellard, Tim Friede, Sarah E Lamb, Joanne Lord, Jason Madan, Tom Morris, Nigel Stallard, Colin Tysall, Adrian Willis, Martin Underwood and the Repository Group



**National Institute for
Health Research**

Identifying back pain subgroups: developing and applying approaches using individual patient data collected within clinical trials

Shilpa Patel,^{1*} Siew Wan Hee,¹ Dipesh Mistry,¹
Jake Jordan,^{2,3} Sally Brown,⁴ Melina Dritsaki,^{1,5}
David R Ellard,¹ Tim Friede,⁶ Sarah E Lamb,^{1,5}
Joanne Lord,^{2,7} Jason Madan,¹ Tom Morris,⁸
Nigel Stallard,¹ Colin Tysall,⁴ Adrian Willis,¹
Martin Underwood¹ and the Repository Group[†]

¹Warwick Medical School, University of Warwick, Coventry, UK

²Brunel University, Health Economics Research Group, Uxbridge, UK

³Surrey Health Economic Centre, School of Economics, University of Surrey,
Guildford, UK

⁴Universities/User Teaching and Research Action Partnership (UNTRAP),
University of Warwick, Coventry, UK

⁵Centre for Rehabilitation Research, Nuffield Department of Orthopaedics,
Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

⁶Department of Medical Statistics, University Medical Centre Göttingen,
Göttingen, Germany

⁷Southampton Health Technology Assessments Centre (SHTAC),
University of Southampton, Southampton, UK

⁸Leicester Clinical Trials Unit, Diabetes Research Centre, University of Leicester,
Leicester, UK

*Corresponding author

†Repository Group collaborators are listed in *Acknowledgements*

Declared competing interests of authors: Sarah E Lamb is chairperson of the Health Technology Assessment Clinical Evaluation and Trials (HTA CET) Board. Martin Underwood is a member of the National Institute for Health Research Journals Library Editorial Group.

Published July 2016

DOI: 10.3310/pgfar04100

This report should be referenced as follows:

Patel S, Hee SW, Mistry D, Jordan J, Brown S, Dritsaki M, *et al.* Identifying back pain subgroups: developing and applying approaches using individual patient data collected within clinical trials. *Programme Grants Appl Res* 2016;**4**(10).

Programme Grants for Applied Research

ISSN 2050-4322 (Print)

ISSN 2050-4330 (Online)

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: nihredit@southampton.ac.uk

The full PGfAR archive is freely available to view online at www.journalslibrary.nihr.ac.uk/pgfar. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

Criteria for inclusion in the *Programme Grants for Applied Research* journal

Reports are published in *Programme Grants for Applied Research* (PGfAR) if (1) they have resulted from work for the PGfAR programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Programme Grants for Applied Research programme

The Programme Grants for Applied Research (PGfAR) programme, part of the National Institute for Health Research (NIHR), was set up in 2006 to produce independent research findings that will have practical application for the benefit of patients and the NHS in the relatively near future. The Programme is managed by the NIHR Central Commissioning Facility (CCF) with strategic input from the Programme Director.

The programme is a national response mode funding scheme that aims to provide evidence to improve health outcomes in England through promotion of health, prevention of ill health, and optimal disease management (including safety and quality), with particular emphasis on conditions causing significant disease burden.

For more information about the PGfAR programme please visit the website: <http://www.nihr.ac.uk/funding/programme-grants-for-applied-research.htm>

This report

The research reported in this issue of the journal was funded by PGfAR as project number RP-PG-0608-10076. The contractual start date was in October 2010. The final report began editorial review in October 2014 and was accepted for publication in September 2015. As the funder, the PGfAR programme agreed the research questions and study designs in advance with the investigators. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The PGfAR editors and production house have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the final report document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent research funded by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, CCF, NETSCC, PGfAR or the Department of Health. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, NETSCC, the PGfAR programme or the Department of Health.

© Queen's Printer and Controller of HMSO 2016. This work was produced by Patel et al. under the terms of a commissioning contract issued by the Secretary of State for Health. This issue may be freely reproduced for the purposes of private research and study and extracts (or indeed, the full report) may be included in professional journals provided that suitable acknowledgement is made and the reproduction is not associated with any form of advertising. Applications for commercial reproduction should be addressed to: NIHR Journals Library, National Institute for Health Research, Evaluation, Trials and Studies Coordinating Centre, Alpha House, University of Southampton Science Park, Southampton SO16 7NS, UK.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).

Programme Grants for Applied Research Editor-in-Chief

Professor Paul Little Professor of Primary Care Research, University of Southampton, UK

NIHR Journals Library Editor-in-Chief

Professor Tom Walley Director, NIHR Evaluation, Trials and Studies and Director of the EME Programme, UK

NIHR Journals Library Editors

Professor Ken Stein Chair of HTA Editorial Board and Professor of Public Health, University of Exeter Medical School, UK

Professor Andree Le May Chair of NIHR Journals Library Editorial Group (EME, HS&DR, PGfAR, PHR journals)

Dr Martin Ashton-Key Consultant in Public Health Medicine/Consultant Advisor, NETSCC, UK

Professor Matthias Beck Chair in Public Sector Management and Subject Leader (Management Group), Queen's University Management School, Queen's University Belfast, UK

Professor Aileen Clarke Professor of Public Health and Health Services Research, Warwick Medical School, University of Warwick, UK

Dr Tessa Crilly Director, Crystal Blue Consulting Ltd, UK

Dr Eugenia Cronin Senior Scientific Advisor, Wessex Institute, UK

Ms Tara Lamont Scientific Advisor, NETSCC, UK

Professor Elaine McColl Director, Newcastle Clinical Trials Unit, Institute of Health and Society, Newcastle University, UK

Professor William McGuire Professor of Child Health, Hull York Medical School, University of York, UK

Professor Geoffrey Meads Professor of Health Sciences Research, Health and Wellbeing Research and Development Group, University of Winchester, UK

Professor John Norrie Health Services Research Unit, University of Aberdeen, UK

Professor John Powell Consultant Clinical Adviser, National Institute for Health and Care Excellence (NICE), UK

Professor James Raftery Professor of Health Technology Assessment, Wessex Institute, Faculty of Medicine, University of Southampton, UK

Dr Rob Riemsma Reviews Manager, Kleijnen Systematic Reviews Ltd, UK

Professor Helen Roberts Professor of Child Health Research, UCL Institute of Child Health, UK

Professor Jonathan Ross Professor of Sexual Health and HIV, University Hospital Birmingham, UK

Professor Helen Snooks Professor of Health Services Research, Institute of Life Science, College of Medicine, Swansea University, UK

Professor Jim Thornton Professor of Obstetrics and Gynaecology, Faculty of Medicine and Health Sciences, University of Nottingham, UK

Professor Martin Underwood Director, Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, UK

Please visit the website for a list of members of the NIHR Journals Library Board:
www.journalslibrary.nihr.ac.uk/about/editors

Editorial contact: nihredit@southampton.ac.uk

Abstract

Identifying back pain subgroups: developing and applying approaches using individual patient data collected within clinical trials

Shilpa Patel,^{1*} Siew Wan Hee,¹ Dipesh Mistry,¹ Jake Jordan,^{2,3} Sally Brown,⁴ Melina Dritsaki,^{1,5} David R Ellard,¹ Tim Friede,⁶ Sarah E Lamb,^{1,5} Joanne Lord,^{2,7} Jason Madan,¹ Tom Morris,⁸ Nigel Stallard,¹ Colin Tysall,⁴ Adrian Willis,¹ Martin Underwood¹ and the Repository Group[†]

¹Warwick Medical School, University of Warwick, Coventry, UK

²Brunel University, Health Economics Research Group, Uxbridge, UK

³Surrey Health Economic Centre, School of Economics, University of Surrey, Guildford, UK

⁴Universities/User Teaching and Research Action Partnership (UNTRAP), University of Warwick, Coventry, UK

⁵Centre for Rehabilitation Research, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

⁶Department of Medical Statistics, University Medical Centre Göttingen, Göttingen, Germany

⁷Southampton Health Technology Assessments Centre (SHTAC), University of Southampton, Southampton, UK

⁸Leicester Clinical Trials Unit, Diabetes Research Centre, University of Leicester, Leicester, UK

*Corresponding author Shilpa.Patel@warwick.ac.uk

†Repository Group collaborators are listed in *Acknowledgements*

Background: There is good evidence that therapist-delivered interventions have modest beneficial effects for people with low back pain (LBP). Identification of subgroups of people with LBP who may benefit from these different treatment approaches is an important research priority.

Aim and objectives: To improve the clinical effectiveness and cost-effectiveness of LBP treatment by providing patients, their clinical advisors and health-service purchasers with better information about which participants are most likely to benefit from which treatment choices. Our objectives were to synthesise what is already known about the validity, reliability and predictive value of possible treatment moderators (patient factors that predict response to treatment) for therapist-delivered interventions; develop a repository of individual participant data from randomised controlled trials (RCTs) testing therapist-delivered interventions for LBP; determine which participant characteristics, if any, predict clinical response to different treatments for LBP; and determine which participant characteristics, if any, predict the most cost-effective treatments for LBP. Achieving these objectives required substantial methodological work, including the development and evaluation of some novel statistical approaches. This programme of work was not designed to analyse the main effect of interventions and no such interpretations should be made.

Methods: First, we reviewed the literature on treatment moderators and subgroups. We initially invited investigators of trials of therapist-delivered interventions for LBP with > 179 participants to share their data with us; some further smaller trials that were offered to us were also included. Using these trials we developed a repository of individual participant data of therapist-delivered interventions for LBP. Using this data set we sought to identify which participant characteristics, if any, predict response to different treatments (moderators) for clinical effectiveness and cost-effectiveness outcomes. We undertook an analysis of covariance to identify potential moderators to apply in our main analyses. Subsequently, we developed and applied three methods of subgroup identification: recursive partitioning (interaction trees and subgroup identification based on a differential effect search); adaptive risk group refinement; and an individual participant data indirect network meta-analysis (NWMA) to identify subgroups defined by multiple parameters.

Results: We included data from 19 RCTs with 9328 participants (mean age 49 years, 57% females). Our prespecified analyses using recursive partitioning and adaptive risk group refinement performed well and allowed us to identify some subgroups. The differences in the effect size in the different subgroups were typically small and unlikely to be clinically meaningful. Increasing baseline severity on the outcome of interest was the strongest driver of subgroup identification that we identified. Additionally, we explored the application of Bayesian indirect NWMA. This method produced varying probabilities that a particular treatment choice would be most likely to be effective for a specific patient profile.

Conclusions: These data lack clinical effectiveness or cost-effectiveness justification for the use of baseline characteristics in the development of subgroups for back pain. The methodological developments from this work have the potential to be applied in other clinical areas. The pooled repository database will serve as a valuable resource to the LBP research community.

Funding: The National Institute for Health Research Programme Grants for Applied Research programme. This project benefited 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.

Contents

List of tables	xiii
List of figures	xvii
List of boxes	xxi
Glossary	xxiii
List of abbreviations	xxv
Plain English summary	xxvii
Scientific summary	xxix
Chapter 1 Overview of the programme	1
Background	1
Defining low back pain	1
Economic burden of low back pain	2
Treatment options for low back pain	2
Effectiveness and cost-effectiveness of treatments for low back pain	3
Subgrouping	3
Aim and objectives	4
Structure of this report	4
Chapter 2 Literature reviews	7
Systematic review 1: identification of potential moderators	7
<i>Abstract</i>	7
<i>Background</i>	7
<i>Aims</i>	7
<i>Method</i>	8
<i>Results</i>	10
<i>Discussion and conclusion</i>	17
Systematic review 2: quality of subgroup analyses in low back pain trials	17
<i>Abstract</i>	17
<i>Background</i>	18
<i>Aims</i>	18
<i>Method</i>	18
<i>Results</i>	20
<i>Discussion and conclusion</i>	26
Summary of reviews	27

Chapter 3 Collating data	29
Identification of potential trials	29
Justification of sample size	29
Process for approaching investigators	29
Secure data transfer	30
Final data set obtained	30
Summary of the included trials in the repository	34
Grouping of interventions	41
 Chapter 4 Creating the repository database and data control	 45
Typographical conventions	45
Background	45
System architecture	47
Mapping and transformation	49
<i>Pilot mapping and transformation</i>	49
<i>XML and XSD for mapping and transforming</i>	50
<i>Mapping clinical data</i>	50
<i>Transforming clinical data</i>	50
<i>Mapping and transforming health-care resource-use data</i>	50
Using entity-attribute-value data	56
Extract, transform and load	56
Data validation	57
Storage	57
Future data sharing	57
 Chapter 5 Crosswalking between disability questionnaire scores	 59
Background	59
Data	59
Outcome conversion	60
Correlation	60
Responsiveness	60
Results	60
Conclusion	65
 Chapter 6 Preliminary statistical analyses and results	 67
Background	67
Statistical analysis plan	67
Definitions	67
<i>Treatment arms</i>	67
<i>Follow-up time point</i>	67
<i>Outcome variables</i>	68
Data sets	69
<i>Clinical analysis</i>	70
<i>Health-economic analysis</i>	70
Methods	70
<i>Descriptive summary</i>	70
<i>One-step meta-analysis</i>	70
<i>Moderator identification</i>	71
Results	71
<i>Descriptive</i>	71
<i>One-step meta-analysis</i>	85
<i>Analyses of covariance</i>	88

Chapter 7 Methodology and statistical developments 1: subgroup identification with recursive partitioning	95
Background	95
Individual patient data interaction tree	96
<i>Growing an initial tree</i>	96
<i>Pruning the initial tree</i>	97
<i>Selecting the best tree</i>	97
Individual patient data subgroup identification based on a differential effect search	97
<i>Growing an initial tree</i>	98
<i>Selecting the final candidate subgroups</i>	99
Analyses	99
Results	100
<i>Analysis 1</i>	100
<i>Analysis 2: pairwise comparisons</i>	102
 Chapter 8 Methodology and statistical developments 2: subgroup identification using an adaptive refinement by directed peeling algorithm	 111
Background	111
Adaptive refinement by directed peeling in individual patient data meta-analysis	111
Analyses	112
Results	113
<i>Analysis 1: overall comparison treatment compared with control</i>	113
<i>Analysis 2: pairwise comparisons</i>	123
 Chapter 9 Methodology and statistical developments 3: identification of cost-effective subgroups by directed peeling	 135
Introduction	135
Methods	135
<i>Quality-adjusted life-years</i>	135
<i>Moderator identification</i>	135
<i>Peeling algorithm</i>	136
<i>Cost-effectiveness</i>	136
Results	137
<i>All interventions versus control: moderators – age and physical component score</i>	137
<i>All interventions versus control: moderators – age and Roland–Morris Disability Questionnaire</i>	141
<i>All interventions versus control: moderators – age, physical component score and Roland–Morris Disability Questionnaire</i>	142
<i>Active physical intervention versus control: moderators – age and Roland–Morris Disability Questionnaire</i>	142
<i>Passive physical intervention versus control: moderators – age and physical component score</i>	146
<i>Adaptive refinement by directed peeling in individual patient data meta-analysis</i>	147
<i>directed peel</i>	147
Discussion	148

Chapter 10 Methodology and statistical developments 4: subgroup identification with individual participant data indirect network meta-analysis	149
Background	149
Methods	149
Results	150
<i>Short-term Roland–Morris Disability Questionnaire outcome</i>	150
<i>Short-term Short Form questionnaire-12 items/-36 items physical component summary outcome</i>	154
<i>Short-term Short Form questionnaire-12 items/-36 items mental component score outcome</i>	156
Chapter 11 Discussion	159
Introduction	159
Summary of key findings	159
<i>Systematic reviews (see Chapter 2)</i>	159
<i>Analyses of covariance (see Chapter 6)</i>	160
<i>Recursive partitioning (see Chapter 7)</i>	160
<i>Adaptive refinement by directed peeling in individual patient data meta-analysis (see Chapter 8)</i>	167
<i>Identification of cost-effective subgroups by direct peeling (see Chapter 9)</i>	168
<i>Network meta-analysis (see Chapter 10)</i>	169
<i>Interpretation</i>	169
<i>Methodological development</i>	171
<i>Strengths</i>	173
<i>Limitations</i>	173
Meaning of the results and clinical implications	174
Recommendations for future research	175
Conclusions	175
Acknowledgements	177
References	181
Appendix 1 Review 2: summary of excluded papers	195
Appendix 2 Invitation letter	201
Appendix 3 Information sheet	203
Appendix 4 Sample data sharing agreement	207
Appendix 5 Instruction on secure data transfer	211
Appendix 6 Excluded studies	213
Appendix 7 Trials unavailable	215
Appendix 8 Scatterplots of raw change scores of outcome measures	217
Appendix 9 Statistical analysis plan	219

List of tables

TABLE 1 Between-group differences for the RMDQ outcome	3
TABLE 2 Review 1: included studies	10
TABLE 3 Review 1: results of the risk of bias assessment	11
TABLE 4 Review 1: results of methodological quality assessments	11
TABLE 5 Mean difference (95% CI) of potential moderators with strong evidence ($p < 0.05$) and weak evidence ($p < 0.20 \geq 0.05$)	12
TABLE 6 Summary of included papers in descending order by subgroup quality assessment	22
TABLE 7 Trials excluded and reason for exclusion, $n = 4$	31
TABLE 8 Trials included and associated publications, $n = 19$	31
TABLE 9 Summary of the included trials in the repository	34
TABLE 10 Step 3: Final grouping of treatment arms for analyses	43
TABLE 11 A sample of the repository standard attributes and values	51
TABLE 12 Instruments used and number of participants by trial	61
TABLE 13 Pearson's correlation coefficient and Cohen's kappa agreement for responsiveness of each pairwise comparison of outcome measures by trial	64
TABLE 14 Number of trials (m) and participants (n) for each outcome by follow-up time points and treatment arms	72
TABLE 15 Demographics and clinical characteristics at baseline by treatment arms	78
TABLE 16 One-step meta-analysis: estimated mean change from baseline to short-term follow-up by treatment arms and the estimated difference between treatment arms (95% CI)	88
TABLE 17 Analysis of covariance: analysis for short-term outcomes (change from baseline to short-term follow-up)	89
TABLE 18 Summary of the included trials and variables used for each short-term outcome measure in analysis 1	100
TABLE 19 Candidate subgroups identified by the IPD-SIDES method for the intervention vs. control/placebo comparison	103

TABLE 20 Summary of the trials included and variables used for each change from baseline to short-term outcome measure and the QALY health outcome measure analysed for the different comparisons	106
TABLE 21 Candidate subgroups identified by the IPD-SIDES method for the passive physical vs. non-active usual care comparison	107
TABLE 22 Candidate subgroups identified by the IPD-SIDES method for the psychological vs. non-active usual care comparison	109
TABLE 23 Candidate subgroups identified by the IPD-SIDES method for the sham vs. non-active usual care comparison	109
TABLE 24 Summary of included trials and variables considered to construct a region that predicts the best or worst response to treatment	113
TABLE 25 Thresholds for selected size of subgroup for the short-term average pain as seen in <i>Figure 25</i>	114
TABLE 26 Thresholds for selected size of subgroup for the short-term EQ-5D as seen in <i>Figure 26</i>	116
TABLE 27 Thresholds for selected size of subgroup for the short-term FFbHR outcome as seen in <i>Figure 27</i>	117
TABLE 28 Thresholds for selected size of subgroup for the short-term SF-12/36 MCS as seen in <i>Figure 28</i>	120
TABLE 29 Thresholds for selected size of subgroup for the short-term SF-12/36 PCS as seen in <i>Figure 29</i>	121
TABLE 30 Thresholds for selected size of subgroup for the short-term RMDQ outcome as seen in <i>Figure 30</i>	124
TABLE 31 Summary of included trials and variables considered to construct a region that predicts the best of worst response to treatment for different direct comparisons	125
TABLE 32 Thresholds for selected size of subgroup for the short-term RMDQ outcome as seen in <i>Figure 31</i>	126
TABLE 33 Thresholds for selected size of subgroup for the short-term FFbHR outcome as seen in <i>Figure 32</i>	127
TABLE 34 Thresholds for selected size of subgroup for the short-term SF-12/36 MCS as seen in <i>Figure 33</i>	128
TABLE 35 Thresholds for selected size of subgroup for the short-term SF-12/36 PCS as seen in <i>Figure 34</i>	130
TABLE 36 Thresholds for selected size of subgroup for the short-term RMDQ outcome as seen in <i>Figure 35</i>	131

TABLE 37 Thresholds for selected size of subgroup for the short-term FFbHR outcome as seen in <i>Figure 36</i>	132
TABLE 38 Thresholds for selected size of subgroup for the short-term SF-12/36 MCS as seen in <i>Figure 37</i>	133
TABLE 39 Thresholds for selected size of subgroup for the short-term SF-12/36 PCS as seen in <i>Figure 38</i>	134
TABLE 40 Adaptive refinement by directed peeling in IPD meta-analysis: analyses conducted on economic outcomes	137
TABLE 41 Algorithm output for analysis 9.3.1 (see <i>Table 40</i>)	139
TABLE 42 Algorithm output for analysis 9.3.2 (see <i>Table 40</i>)	143
TABLE 43 Algorithm output for analysis 9.3.3 (see <i>Table 40</i>)	145
TABLE 44 Treatment effect with modification (absolute reduction in the short-term RMDQ outcome, mean and 95% credible interval). Coefficients given for individual aged 50 years, male, RMDQ score = 10, PCS = 40 and MCS = 40 at baseline	151
TABLE 45 Means, 95% credible intervals and BPs (%) for impact of participant characteristics on effect of treatments (vs. control)	152
TABLE 46 Probability that any given treatment is optimal for a range of participant profiles	153
TABLE 47 Treatment effect with modification (absolute increase in short term PCS, mean and 95% credible interval). Coefficients given for individual aged 50 years, male, PCS and MCS = 40, Predicted change in condition without treatment adjusted for age, sex, MCS	154
TABLE 48 Means, 95% credible intervals and BPs (%) for impact of participant characteristics on effect of treatments (vs. control)	155
TABLE 49 Probability that any given treatment is optimal for a range of participant profiles with PCS as outcome of interest	156
TABLE 50 Treatment effect with modification (absolute change in short-term MCS, mean and 95% credible interval). Coefficients given for individual aged 50 years, male, PCS and MCS = 40. Predicted change in condition without treatment adjusted for age, sex, baseline values of SF-12/36 PCS and MCS	157
TABLE 51 Mean, 95% credible intervals and BPs (%) for impact of participant characteristics on effect of treatments in the network	157
TABLE 52 Probability that any given treatment is optimal for a range of participant profiles	158
TABLE 53 Overview of results: intervention vs. control (usual care or sham)	161

TABLE 54 Overview of results: active physical vs. control (usual care)	162
TABLE 55 Overview of results: passive physical vs. usual care control	163
TABLE 56 Overview of results: psychological vs. usual care control	164
TABLE 57 Overview of results: sham vs. control	165

List of figures

FIGURE 1 The structure of the current report	5
FIGURE 2 Review 1: Quorum statement flow diagram	10
FIGURE 3 Review 2: Quorum statement flow diagram	21
FIGURE 4 Quorum statement flow diagram for database identification	30
FIGURE 5 Step 1: classification of trials into core groups	42
FIGURE 6 Step 2: classification of trials with indication of number of trials and participants for direct and indirect comparisons	43
FIGURE 7 (a) A sample of original tabular format data; and (b) normalised relational interpretation of the original tabular data	46
FIGURE 8 The entity–relationship diagram for the hybrid repository database depicting the fixed schema with the subschema EAV tables	48
FIGURE 9 (a) A sample of original tabular format clinical data; and (b) the XML mapping and transformation instructions; and (c) the sample data represented as EAV	49
FIGURE 10 (a) A sample of questions in a CRF at 3-month follow-up; (b) a sample of original tabular format health-care resource-use data; and (c) a sample of how the health-care resource-use data populate the repository standard	52
FIGURE 11 The XML mapping and transformation instructions for the sample data in <i>Figure 10</i>	54
FIGURE 12 Scatterplots of standardised change scores for (a) PCS vs. CPG ($n = 2451$); and (b) PCS vs. FFbHR ($n = 3620$) outcome measures	61
FIGURE 13 Scatterplots of standardised change scores for (a) PCS vs. RMDQ ($n = 1694$); and (b) PCS vs. ODI ($n = 206$) outcome measures	62
FIGURE 14 Scatterplots of standardised change scores for (a) PCS vs. PSFS ($n = 158$); and (b) CPG vs. FFbHR ($n = 1110$) outcome measures	62
FIGURE 15 Scatterplots of standardised change scores for (a) CPG vs. RMDQ ($n = 1661$); and (b) PSFS vs. RMDQ ($n = 625$) outcome measures	63
FIGURE 16 Scatterplots of standardised change scores for (a) PDI vs. PCS ($n = 281$), and (b) FFbHR vs. and PDI ($n = 284$) outcome measures	63
FIGURE 17 The estimated efficacy between control (non-active usual care and sham) and intervention treatments from one-step meta-analysis for (a) FFbHR; and (b) RMDQ score	85

FIGURE 18 The estimated efficacy between control (non-active usual care and sham) and intervention treatments from one-step meta-analysis for (a) average pain (based on VAS); and (b) PCS of SF-12/36	86
FIGURE 19 The estimated efficacy between control (non-active usual care and sham) and intervention treatments from one-step meta-analysis for (a) MCS of SF-12/36; and (b) EQ-5D	87
FIGURE 20 Example of a tree structure	95
FIGURE 21 Candidate subgroups identified (shaded green) by the IPD-SIDES method when applied to change from baseline to short-term FFbHR (range 0–100; lower score implies greater disability) outcome for the intervention against control/placebo comparison	104
FIGURE 22 Candidate subgroup identified (shaded green) by the IPD-SIDES method when applied to change from baseline to short-term SF-12/36 MCS outcome (range 0–100; lower score implies worse mental functioning) for the intervention against control/placebo comparison	104
FIGURE 23 Candidate subgroups identified (shaded green) by the IPD-SIDES method when applied to change from baseline to short-term SF-12/36 PCS outcome (range 0–100; lower score implies worse physical functioning) for the intervention against control/placebo comparison	105
FIGURE 24 Schematic diagram of the ARDP algorithm to identify subgroups of treatment responders	112
FIGURE 25 Trajectory plot for the treatment effect against the size of the constructed region for the average pain short-term outcome	114
FIGURE 26 Trajectory plot for the treatment effect against the size of the constructed region for the EQ-5D short-term outcome	115
FIGURE 27 Trajectory plot for the treatment effect against the size of the constructed region for the FFbHR short-term outcome	117
FIGURE 28 Trajectory plot for the treatment effect against the size of the constructed region for the SF-12/36 MCS short-term outcome	119
FIGURE 29 Trajectory plot for the treatment effect against the size of the constructed region for the SF-12/36 PCS short-term outcome	121
FIGURE 30 Trajectory plot for the treatment effect against the size of the constructed region for the RMDQ short-term outcome	123
FIGURE 31 Trajectory plot for the treatment effect between active physical and non-active usual care against the size of the constructed region for the RMDQ short-term outcome	126
FIGURE 32 Trajectory plot for the treatment effect between passive physical and non-active usual care against the size of the constructed region for the FFbHR short-term outcome	127

FIGURE 33 Trajectory plot for the treatment effect between passive physical and non-active usual care against the size of the constructed region for the SF-12/36 MCS short-term outcome	128
FIGURE 34 Trajectory plot for the treatment effect between passive physical and non-active usual care against the size of the constructed region for the SF-12/36 PCS short-term outcome	129
FIGURE 35 The size of the constructed region for the RMDQ short-term outcome	130
FIGURE 36 The size of the constructed region for the FFbHR short-term outcome	131
FIGURE 37 Trajectory plot for the treatment effect between sham and non-active usual care against the size of the constructed region for the SF-12/36 MCS short-term outcome	132
FIGURE 38 Trajectory plot for the treatment effect between sham and non-active usual care against the size of the constructed region for the SF-12/36 PCS short-term outcome	133
FIGURE 39 Mean treatment effect in subgroup	138
FIGURE 40 Weighted mean treatment effect across treatment subgroup and non-treatment subgroup	138
FIGURE 41 Mean treatment effect in subgroup	141
FIGURE 42 Weighted mean treatment effect across treatment subgroup and non-treatment subgroup	141
FIGURE 43 Mean treatment effect in subgroup	144
FIGURE 44 Weighted mean treatment effect across treatment subgroup and non-treatment subgroup	144
FIGURE 45 Mean treatment effect in subgroup	146
FIGURE 46 Mean treatment effect in subgroup	146
FIGURE 47 Mean treatment effect in subgroup	147
FIGURE 48 Network of evidence for the short-term RMDQ outcome	151
FIGURE 49 Roland–Morris Disability Questionnaire outcome: optimal treatment as a function of RMDQ score at baseline and age for men with MCS = PCS = 40, with proportion of male trial participants whose baseline RMDQ score and age fit into each zone ($n = 721$)	153
FIGURE 50 Roland–Morris Disability Questionnaire outcome: optimal treatment as a function of RMDQ score at baseline and age for women with MCS = PCS = 40, with proportion of female trial participants whose baseline RMDQ score and age fit into each zone ($n = 1054$)	153

FIGURE 51 Network of evidence for short-term PCS	154
FIGURE 52 Physical component score outcome: optimal treatment as a function of MCS and PCS at baseline for men aged 50 years, with proportion of male participants whose MCSs and PCSs at baseline fit into each zone ($n = 2296$)	155
FIGURE 53 Physical component score outcome: optimal treatment as a function of MCS and PCS at baseline for women aged 50 years, with proportion of female participants whose MCSs and PCSs at baseline fit into each zone ($n = 3278$)	156
FIGURE 54 Mental component score outcome: optimal treatment as a function of MCS and PCS at baseline for men aged 50 years, with proportion of male participants whose MCSs and PCSs at baseline fit into each zone ($n = 2296$)	158
FIGURE 55 Mental component score outcome: optimal treatment as a function of MCS and PCS at baseline for women aged 50 years, with proportion of female participants whose MCSs and PCSs at baseline fit into each zone ($n = 3278$)	158

List of boxes

BOX 1 Review 1: inclusion and exclusion criteria	9
BOX 2 Review 2: inclusion and exclusion criteria	19
BOX 3 Key recommendations in the area of subgroup analyses	20

Glossary

Adaptive refinement A method to identify subgroups of participants, defined by cut-offs for the selected covariates, resulting in box-shaped subgroups.

Crosswalking A method of mapping multiple participant-reported outcome measures that measure the same domain, to a common scale.

Moderators Factors measured prior to randomisation that subsequently influence the effect of the treatment.

Recursive partitioning A technique that searches all possible binary splits of covariates to identify subgroups of participants.

Standardised mean difference The score divided by the standard deviation of the baseline score of all participants.

List of abbreviations

ANCOVA	analysis of covariance	IT	interaction tree
APT	active physical therapy	LBP	low back pain
ARDP	adaptive refinement by directed peeling	MCS	mental component score
ARDP-MA	adaptive refinement by directed peeling in individual patient data meta-analysis	NICE	National Institute for Health and Care Excellence
AUC	area under the curve	NMB	net monetary benefit
BeST	Back Skills Training (trial)	NSLBP	non-specific low back pain
BP	Bayesian Probabilities of effect modification	NWMA	network meta-analysis
CI	confidence interval	ODI	Oswestry Disability Index
CNSLBP	chronic non-specific low back pain	PCS	physical component score
CPG	Chronic Pain Grade Scale	PDI	Pain Disability Index
CPG-DS	Chronic Pain Grade Scale disability score	PI	principal investigator
CPG-PS	Chronic Pain Grade Scale pain intensity score	PROM	participant-reported outcome measure
CRF	case report form	PSFS	Patient-Specific Functional Scale
EAV	entity-attribute-value	QALY	quality-adjusted life-year
EQ-5D	European Quality of Life-5 Dimensions	RCT	randomised controlled trial
ETL	extract, transform and load	RMDQ	Roland–Morris Disability Questionnaire
FFbHR	Hannover Functional Ability Questionnaire for measuring back pain-related functional limitations (Funktionsbeeinträchtigung durch Rückenschmerzen)	SD	standard deviation
FU	follow-up (table)	SF-12	Short Form questionnaire-12 items
GP	general practitioner	SF-36	Short Form questionnaire-36 items
IPD	individual patient data	SF-6D	Short Form questionnaire-6 Dimensions
IPD-IT	individual patient data interaction tree	SIDES	subgroup identification based on a differential effect search
IPD-SIDES	individual patient data subgroup identification based on a differential effect search	SMD	standardised mean difference
		TENS	transcutaneous electrical nerve stimulation
		UK BEAM	UK Back pain Exercise And Manipulation
		VAS	visual analogue scale
		XML	extensible mark-up language

Plain English summary

Low back pain is a common and costly disorder for both the patient and the health service, which can be managed using different treatment approaches, some of which are delivered in a physiotherapy department. The benefits of treatments delivered by therapists are small, on average, that is, they get small improvements. If we could predict which patients would be most likely to benefit from different treatments it would be possible to improve the overall effectiveness of treatments and potentially make better use of NHS resources. To address this we pooled together data from 19 back pain trials from around the world. This provided us with a data set of 9328 patients. We developed novel statistical methods to identify subpopulations (groups of people with similar characteristics) that would be likely to benefit from certain treatments. Of the three methods developed, two allowed us to identify subpopulations. The additional benefits for individuals in the subpopulations were modest and unlikely to be of clinical importance. Our third method was exploratory and allowed us to identify the chance of a particular treatment choice being effective for a particular patient.

Overall, we did not find any subpopulations that would benefit from treatment. Neither did we find that such an approach to identifying patients would be cost-effective. We have developed new ways of identifying subpopulations and would recommend the application of these methods to other clinical conditions. We have also developed, from prior trials, a data pool that will now become a resource for back pain researchers to help them answer other questions in the field.

Scientific summary

Background

Identifying subgroups of people living with low back pain (LBP) who may do better, or worse, with different treatment choices is a high research priority internationally. Many randomised controlled trials (RCTs) could be designed to address individual components of this problem. High-quality trials in this area are very costly and time-consuming (typically requiring a minimum of 700 participants, at a cost of £1M to £2M, and taking at least 6 years from design to implementation); each will address only one small part of this complex problem.

Alternative methods can provide complementary information that could add value to our knowledge. Approaches, that make the best possible use of existing data might produce timely answers to a range of important research questions and provide substantial added value to the money that is already invested in this area.

We present a programme of work – using systematic reviews, methodological development, and secondary analyses of existing data sets – to identify strategies to improve outcomes for people who are seeking treatment for back pain by improving how patients, clinicians and purchasers choose treatments. Our programme of work ensures that the maximum information is gleaned from existing substantial trial data sets. The analysis plan for these data and the modelling of clinical effectiveness and cost-effectiveness are informed by our literature reviews.

Aims and objectives

The overall aim of this programme grant was to improve the clinical effectiveness and cost-effectiveness of therapist-delivered treatments for LBP by providing patients, their clinical advisors and health service purchasers with better information about which patients are most likely to benefit from which treatment choices. Our objectives were to:

1. synthesise what is already known about the validity, reliability and predictive value of possible treatment moderators (patient factors that predict response to treatment)
2. develop a repository of individual participant data from RCTs testing therapist-delivered interventions for LBP
3. determine which participant characteristics, if any, predict clinical response to different treatments for LBP
4. determine which participant characteristics, if any, predict the most cost-effective treatments for LBP.

Seeking to achieve these objectives required substantial methodological work, including the development and evaluation of some novel statistical approaches. This programme of work was not designed to analyse the main effect of interventions and no such interpretations should be made.

Method and results

To synthesise what is already known about the validity, reliability and predictive value of possible treatment moderators

We carried out two systematic reviews: one to identify potential moderators of treatment effect from studies of therapist-delivered interventions to inform our analyses, and, the second, to review the quality of subgroup analyses in LBP trials.

As the purpose of moderator identification was for future application in our analyses, we identified potential moderators with strong evidence ($p < 0.05$) and potential moderators with weaker evidence in one or more studies ($0.05 < p \leq 0.20$). Data from four trials were included in the review. Potential moderators with strong evidence included age, employment status and type, back pain status, narcotic medication use, treatment expectations and education. Potential moderators with weaker evidence included gender, psychological distress, pain/disability and quality of life. Although the overall data were weak and lacking in rigour to inform clinical practice, they provided a starting point for application in our analyses.

The second review looked at the quality and reporting of subgroup analyses in LBP. Thirty-nine papers were included in the final review. The majority of papers provided only exploratory or insufficient findings. Only three trials provided confirmatory findings (i.e. subgroup analyses were hypothesis driven and grounded in existing theory or empirical data). The overall quality of reporting was poor and, generally, the subgroup analyses have been severely underpowered. We concluded the need to develop new approaches to subgroup identification to identify multiple participant characteristics or clusters of moderators that would identify who is most or least likely to benefit.

To develop a repository of individual participant data from randomised controlled trials testing therapist-delivered interventions for low back pain

To allow the identification of subgroups in appropriately powered data sets, we developed a repository of data from completed trials. We used a systematic approach in identifying trials and approached chief investigators for their data. Our pool of potential trials came from the search results that were generated in our review of moderators. As a starting point, we were interested only in RCTs of therapist-delivered interventions with a sample size of > 179 . We were offered data from three smaller trials, which we also included.

The final repository comprises 19 trials, with 9328 participants. No two trials had identical interventions or controls. Despite the large initial sample, we had to broadly pool interventions into groups for our analyses in order to draw any meaningful comparisons. As a first step, we identified the control interventions and classified these as either usual care or as a sham control; furthermore, we have specified the type of sham, as there may be qualitative differences between sham treatments. To cluster the interventions we first classified them into core groups (individual physiotherapy, exercise, manipulation, advice/education, psychological therapy, graded activity, acupuncture, combination therapy, mock transcutaneous electrical nerve stimulation, sham acupuncture and control). We later looked at the data to explore the scope for direct and indirect comparisons, and the data available for these comparisons. This indicated that, without grouping these interventions, it would be difficult to make any meaningful comparisons; therefore, the collaborative team decided on broader categories: active physical (exercise and graded activity), passive physical (individual physiotherapy, manipulation and acupuncture) and psychological (advice/education and psychological therapy). In this programme of work we are not seeking to estimate the true effect size of any individual intervention. Rather, we are seeking to identify predictors of treatment response making it reasonable to pool in this manner.

In addition to the challenges of pooling multiple data sets using multiple interventions, there was careful consideration of how to most accurately map multiple participant-reported outcome measures that measure the same domain, to a common scale. We concluded that, because of the lack of correlation and responsiveness in outcomes from two measures in the same individual, it would not be appropriate to map any physical disability outcome measures to another.

To determine which participant characteristics, if any, predict clinical response to different treatments for low back pain

We undertook analysis of covariance analyses comparing all of the intervention groups with all controls to identify potential moderators to take forward for our main analyses. We were able to take forward the Hannover Functional Ability score, the Roland–Morris Disability questionnaire (RMDQ), the Short Form questionnaire-12 items (SF-12)/ Short Form questionnaire-36 items (SF-36) physical and mental component scores, age, gender, pain, fear avoidance and coping as variables, with a possible signal in one or more analysis.

In this programme grant we have explored, in considerable detail, new and novel methods for subgroup identification. We have presented three core methods in this report: recursive partitioning (interaction trees and subgroup identification based on a differential effect search), adaptive risk group refinement and individual participant data indirect network meta-analysis (NWMA).

Our prespecified analytical approaches – recursive partitioning and adaptive risk group refinement – produced identifiable subgroups, the parameter definitions of which were grounded in the data. The differences in effect sizes, between groups, however, were small, and unlikely to be clinically meaningful. The effect sizes in the groups who did less well would still justify the use of these interventions. The overall results point to larger treatment responses in those with higher levels of the outcome of interest at baseline. The results also suggest that those with greater psychological distress, as measured by the SF-12/36 mental component score, do not have a greater treatment effect on physical outcomes from any of the therapist-delivered interventions tested. Targeting low-intensity interventions at those with higher levels of psychological distress for treatment might not be justified.

We undertook a post hoc exploratory individual participant data indirect NWMA to identify subgroups. This does not identify subgroups in the traditional manner but rather uses the available data to work out the probability that a particular treatment choice is most likely to be effective. The outputs from this method have the potential to inform clinical decision-making but requires further testing and application.

To determine which participant characteristics, if any, predict the most cost-effective treatments for low back pain

We applied the directed peeling algorithm to the economic and resource-use data. When exploring interventions compared with control, subgroups were identified. These subgroups comprised patients who were older, with relatively worse physical functioning at baseline. The gain in treatment effect for the subgroup was small, therefore, given the relatively low cost of the intervention treatment it is likely to be cost-effective for the whole patient group. No convincing subgroups were found for active and passive physical treatment. This may be as a result of lack of power or simply that there is no subgroup to be found.

Age, SF-12/36 physical component score and RMDQ score were the three potential moderators identified from the economic analysis. However, the relationship of the quality-adjusted life-years (QALYs) with the moderators differed in some cases to that of the clinical outcome measures. Subgroups were identified only in the comparison of treatment with control. Our interpretation is that those who are older, with worse RMDQ score and SF12/36 physical component score are likely to gain a greater benefit on QALY outcomes from treatment. Doing this, however, will not improve overall QALY gain and is very unlikely to be seen as a cost-effective choice if the National Institute for Health and Care Excellence threshold of £20,000–30,000 per QALY is used to inform treatment choices.

Conclusions and recommendations

In this programme of work we have developed advances in methodological developments for subgroup analyses. We have developed different approaches to the identification of differential subgroup effects that provide considerable added value compared with conventional analyses that simply test for interactions between single baseline parameters and treatment allocation. In addition, we have developed advanced systems for pooling and storing large data sets, highlighted that it is not possible to map different outcome measures for a meta-analysis, and, finally, we have developed an important resource for back pain researchers who wish to undertake further analyses on data from multiple trials.

Clinically, the application of the different frequentist methods (recursive partitioning and adaptive design) has not allowed us to identify subgroups of patients who might benefit from different back pain treatments. Some of the core outputs and recommendations from this work include:

- application of these methods for the identification of subgroups in other clinical areas
- reanalysis of existing meta-analyses of back pain treatments to separate out results from trials with different outcome measures
- further development of methods and application to the data that we already have
- making the data set available to other researchers
- adding additional trial data sets to the repository
- developing and testing a web portal to help inform choice of treatments based on our NWMA.

Overall, our results do not provide sufficient clinical effectiveness or cost-effectiveness justification for the use of baseline characteristics in the development of subgroups for LBP. We would, however, suggest that such methods should be applied in other clinical areas where subgroups may be important. The exploratory outputs from our Bayesian NWMA provide some scope for deciding on optimal therapies. This, however, would need empirical testing before clinical recommendation.

Funding

Funding for this study was provided by the Programme Grants for Applied Research programme of the National Institute for Health Research. This project benefited 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.

Chapter 1 Overview of the programme

In this chapter we have provided the background and rationale for our programme to improve the clinical effectiveness and cost-effectiveness of low back pain (LBP) treatment by identifying groups that may gain maximum benefit from therapist-delivered treatments.

Background

Chronic non-specific LBP (CNSLBP) is a common problem affecting a large proportion of the population.^{1–4} In the UK, around 70–80% of adults will experience back pain at some point in their life.⁵ Some argue that episodic LBP is a universal part of human experience.^{6,7} Half of the adult population in the UK (49%) report LBP lasting at least 24 hours in a 1-year period.⁵ The 2010 Global Burden of Disease study⁸ identified LBP as the leading cause of years lived with disability internationally. LBP affects around one-third of the world's population.⁸

Most episodes of back pain are short lived, resolving without the need for any specific treatment. It is the minority of episodes that develop into CNSLBP which create the greatest health need. The natural history of LBP is untidy; around 70% of those affected will experience at least one recurrent episode within a 12-month period.⁹

The true prevalence of CNSLBP is difficult to estimate, as definitions and populations vary between studies and countries. However, a review of prevalence studies, reported, between 1966 and 1998, a 12–33% point prevalence; 22–65% 1-year prevalence and up to 84% lifetime prevalence.¹⁰

Since this review, further reviews on the prevalence, focusing on older people and adolescents, have been published.^{3,11} A 2012 systematic review synthesised the global prevalence of LBP in studies published between 1980 and 2009. The greatest prevalence was in females aged 40–80 years. After adjusting for methodological variations the point prevalence of back pain lasting for > 1 day was 11.9% [95% confidence interval (CI) 7.98% to 15.82%] and 1-month period prevalence was estimated at 23.2% (95% CI 17.52% to 28.88%).¹²

Defining low back pain

The International Association of the Study of Pain defines pain as 'an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage'.¹³ The British Pain Society defines *acute pain* as 'short term lasting less than 12 weeks' duration', whereas *chronic pain* is defined as 'long-term pain of more than 12 weeks or after the time that healing would have been thought to have occurred in pain after trauma or surgery'.¹⁴

Low back pain is diagnosed based on the presence of pain and discomfort in the lumbosacral area.¹⁵ Some people also experience pain in the upper leg as a result of LBP. In the majority of cases it is difficult to identify a single cause for back pain. A 2013 systematic review¹⁶ of studies of new presentations of LBP found a combined prevalence of 1.5% for fracture and malignancy in primary care; in secondary and tertiary care, prevalence was 6.5%. Once specific causes for LBP have been excluded [malignancy, fracture, infection, inflammatory disorders (such as ankylosing spondylitis)] then a diagnosis of non-specific LBP (NSLBP) is made. This recognises the difficulty in producing robust classification criteria to identify different populations of people affected by chronic LBP.

There is no evidence for a reduction in the population burden of LBP over time. Between 1990 and 2010, in the UK, the number of disability-adjusted life-years attributable to LBP increased by 3.7% from 2231 (95% CI 1555 to 3015) of 100,000 to 2313 (95% CI 1574 to 3113) of 100,000 of the age-standardised population.¹⁷

Economic burden of low back pain

Low back pain is a costly condition to society, health care and the individual. It is the leading cause of sickness absence and health-care use.¹⁸⁻²¹ In the UK, the direct health-care cost of back pain in 1998 was £163M. However, the larger burden is that of the indirect costs related to lost production and informal care, which were estimated to be at least £5018M.²² More up-to-date UK estimates are not available. The current cost is likely to be substantially larger. It is difficult to make direct comparisons of the cost of LBP internationally because of varying health and social care systems.²³

Low back pain results in approximately 4% of the UK population taking time off work. This translates to around 90 million working days lost and between 8 and 12 million general practitioner (GP) consultations per year.^{22,24} In 2013 the Office for National Statistics reported 131 million lost working days due to sickness absences in that year in the UK; 30.6 million of these (23%) were lost because of musculoskeletal conditions including back and neck pain.²⁵

Treatment options for low back pain

People experiencing LBP will often seek medical and drug therapies, as well as therapist-delivered complementary therapies, such as acupuncture, chiropractic or osteopathy, to help relieve pain.²⁶ Until comparatively recently there were few robust trials of treatments for LBP, and no convincing evidence for the effectiveness of any back-pain treatments. Guidance on the management of LBP was based largely on expert opinion, custom and practice. Since the mid-1990s, there has been a substantial investment in high-quality randomised controlled trials (RCTs) of different treatments for NSLBP. We now have good evidence to show that several therapist-delivered treatment approaches are effective, and for some of these there is also evidence that they are cost-effective.^{15,27} By 'therapist-delivered interventions' we mean non-drug, non-surgical approaches to the treatment of LBP. Typically, these are delivered by physiotherapists or health/clinical psychologists, but they may be delivered by doctors, health trainers, statutorily regulated complementary practitioners (such as osteopaths or chiropractors), or independently registered professionals providing treatments such as acupuncture or the Alexander technique. The types of interventions offered include acupuncture, manual treatments, exercise regimens, cognitive behavioural approaches or combinations of these.

A number of therapist-delivered interventions are superior to 'treatment as usual' (GP care) for participants with chronic LBP. There are numerous treatment options for LBP and several guidelines recommending treatment, including the National Institute for Health and Care Excellence (NICE), the European Corporation in Science and Technology, and the American College of Physicians and American Pain Society guidelines. Such guidance is typically framed as examining independent treatment modalities. Any recommendation for a treatment modality is, inevitably, recommending a package of care including both the non-specific effects of the therapist encounter and the specific effects of the treatment modality in question.

In 2009, NICE guidance¹⁵ advised that all people with persistent LBP should be given advice and encouraged to self-manage. As part of this advice, people are encouraged to remain physically active and to engage in daily activity. Subsequently, those affected should be offered a course of acupuncture, exercise or manual therapy.¹⁵ The decision on which treatment to select should be a collaborative decision, taking into account the patient's treatment preferences. If the selected treatment option is not effective then the patient should be offered another option from the remaining recommended treatments. If the patient is still troubled by back pain then he/she should be considered for an intense physical and psychological intervention. NICE is currently revising its LBP guidelines.

Effectiveness and cost-effectiveness of treatments for low back pain

Although the effectiveness of adding a range of therapist-delivered interventions to best usual care or to no treatment has been well established, the typical mean effect sizes are, at best, modest. By way of illustration, the minimally important (within-person) change in the Roland–Morris Disability Questionnaire (RMDQ) score,²⁸ the most commonly used outcome measure in back pain trials, has been established as 5 points.^{29,30} Typical between-group differences in high-quality RCTs are in the order of 1–2 points on the RMDQ, although a few studies have found larger effect sizes (*Table 1*). These modest mean differences probably translate into ‘numbers needed to treat’ in the order of 5–10.^{29,33} These are similar to the numbers needed to treat that are found with antidepressant or antiepileptic drugs which are used to treat chronic painful disorders.³⁶

The cost per quality-adjusted life-year (QALY) for some of these treatments is well within cost-effectiveness thresholds that are usually used by NICE. Despite this, evidence of access to such treatments within the UK NHS remains patchy. The guideline-endorsed treatments of interdisciplinary rehabilitation, exercise, acupuncture, spinal manipulation and cognitive–behavioural therapy for subacute or chronic LBP have been shown to be cost-effective, but evidence for other endorsed treatments for NSLBP do not yield conclusive or consistent evidence about their relative cost-effectiveness.³⁷ The scarcity of economic evaluations for some guideline-endorsed treatments means that well-conducted economic evaluations are required to strengthen the evidence base of treatments for LBP.

Subgrouping

Identifying which participants are likely to gain the greatest benefit from different treatments for LBP is an identified high research priority internationally and was one of the key recommendations for future research in the 2009 NICE guidelines for the management of persistent LBP. Current research does not provide any robust data on how to match back pain treatments to participants to maximise effects on outcomes relevant to the participant and cost-effectiveness for the health service.

As different treatment options are argued to work in very different ways, it is a reasonable hypothesis that matching people with LBP to those treatments that are more likely to be effective for their back pain will be a more efficient use of health-care resources and will improve patient outcomes. One might expect that

TABLE 1 Between-group differences for the RMDQ outcome

Study	Control	Intervention	Mean difference in RMDQ score (95% CI); SMD	
			3 months	12 months
UK BEAM ³¹	GP care	Exercise	1.36 (0.63 to 2.10); 0.34	0.39 (–0.41 to 1.19); 0.10
		Manipulation	1.57 (0.82 to 2.32); 0.39	1.01 (0.22 to 1.81); 0.25
		Manipulation plus exercise	1.87 (1.15 to 2.60); 0.47	1.30 (0.54 to 2.07); 0.33
A-TEAM ³²	Usual care	Massage	1.96 (0.74 to 3.18); 0.39	0.58 (0.77 to 1.94); 0.12
		Alexander technique (six sessions)	1.71 (0.47 to 2.95); 0.34	1.40 (0.03 to 2.77); 0.28
		Alexander technique (12 sessions)	2.91 (1.66 to 4.16); 0.58	3.40 (2.03 to 4.76); 0.68
BeST ^{33,34}	Advice only	Cognitive–behavioural therapy	1.10 (0.38 to 1.71); 0.22	1.30 (0.56 to 2.06); 0.27
York Yoga ³⁵	Usual care	Yoga	2.17 (1.03 to 3.31); 0.50	1.57 (0.42 to 2.71); 0.36

A-TEAM, Alexander technique lessons, exercise, and massage; BeST, Back Skills Training Trial; SMD, standardised mean difference; UK BEAM, UK Back pain Exercise And Manipulation.

people with high levels of psychological distress that is related to their back pain may gain greater benefit from a psychologically orientated intervention, such as cognitive-behavioural therapy; those with marked loss of physical fitness to benefit most from an exercise intervention; or those with poor back function to benefit most from manual therapy interventions. Developing an evidence base to inform the development of such a stratified care approach has great potential to improve outcomes for people with LBP.

We are aware of one trial of a stratified care approach, published after this programme of work started. The STarT Back trial³⁸ successfully demonstrated that a combination of using a stratification tool and enhanced physiotherapy packages for selected participants improves outcomes and reduces costs when compared with usual physiotherapy care. This study³⁸ does not, however, allow the performance of the stratification tool to identify subgroups to be assessed.

There is a myriad of RCTs that could be designed to address individual components of this problem. High-quality trials in this area are very costly and time-consuming, and can address only one small part of this complex problem. Alternative approaches, which make the best possible use of existing data, can produce timely answers to a range of important research questions and provide substantial added value to the money that is already invested in this area.

We present a programme of work – using systematic reviews, methodological development and secondary analyses of existing data sets – to identify strategies to improve outcomes for people seeking treatment for back pain, by improving how participants, clinicians and purchasers choose treatments. Our programme of work ensures that the maximum information is gleaned from existing substantial trial data sets. The analysis plan for these data and modelling of clinical effectiveness and cost-effectiveness are informed by our literature reviews.

Aim and objectives

The overall aim was to improve the clinical effectiveness and cost-effectiveness of LBP treatment by providing participants, their clinical advisors and health service purchasers with better information about which participants are most likely to benefit from which treatment choices. To achieve this, our objectives were to:

1. synthesise what is already known about the validity, reliability and predictive value of possible treatment moderators
2. develop a repository of individual participant data from RCTs testing therapist-delivered interventions for LBP
3. determine which participant characteristics, if any, predict clinical response to different treatments for LBP
4. determine which participant characteristics, if any, predict the most cost-effective treatments for LBP.

We have defined a therapist as a person trained in administering any of the available recommended treatments, excluding drug interventions and surgical interventions, for the management of LBP.

Structure of this report

This report has been structured as shown in *Figure 1*. In this report we use some specific terminology that needs additional definition to aid understanding. We have defined these in the *Glossary* at the start of this report and in more detail at relevant points in the report.

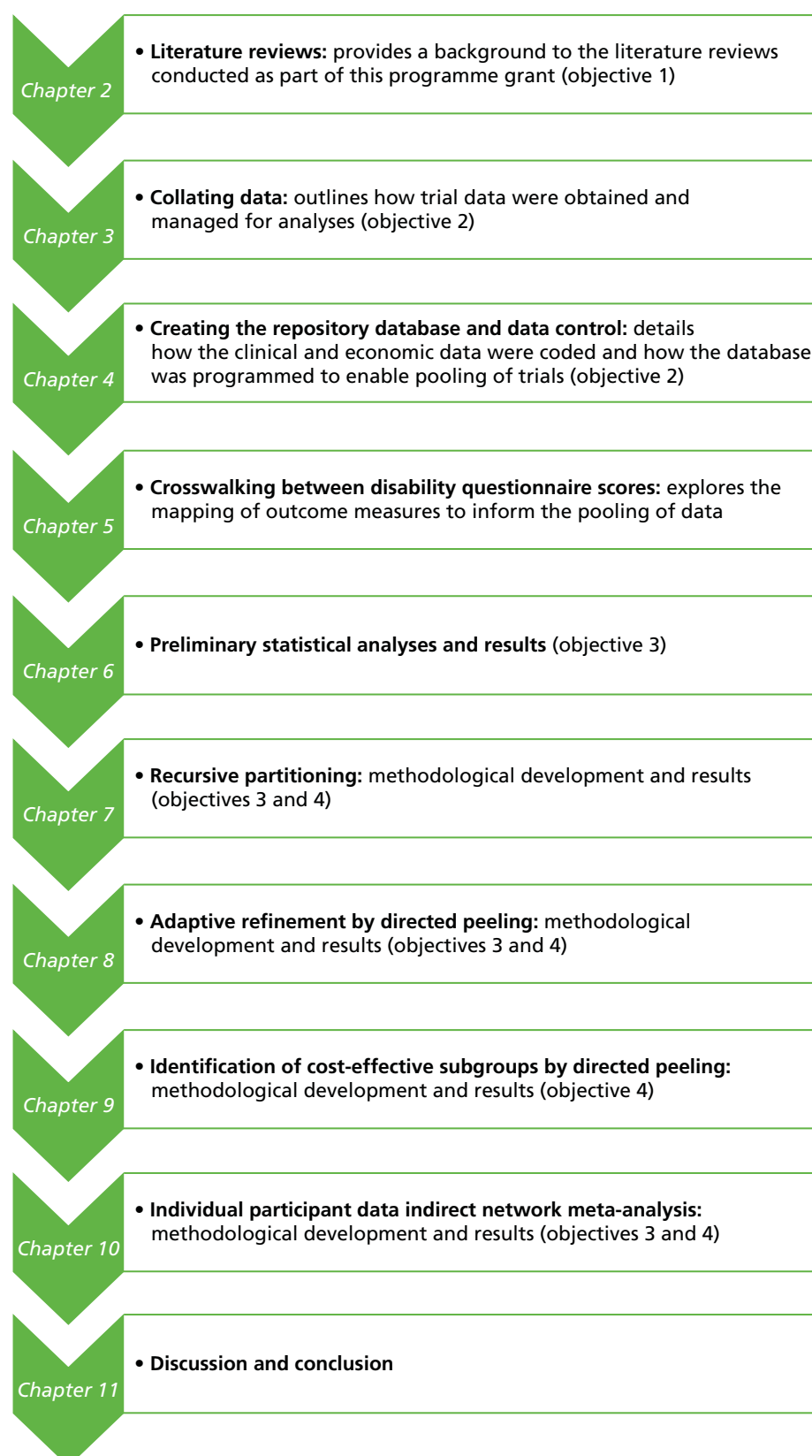


FIGURE 1 The structure of the current report.

Chapter 2 Literature reviews

As part of this programme of work, we carried out two systematic reviews. In this chapter, we have presented the details and results of each review, followed by an overall summary.

Systematic review 1: identification of potential moderators

This review has been published in *Physiotherapy* under the terms of the Creative Commons Attribution – NonCommercial – NoDerivs (CC BY-NC-ND 4.0) Licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>). Here we present a summary of the paper.³⁹

Abstract

Background: in RCTs, moderators are baseline characteristics that predict whether or not an intervention will be more or less effective for an individual in the trial. For our final individual participant data meta-analyses selected potential moderators grounded in existing data to inform our selection.

Aim: to identify potential moderators from existing studies of therapist-delivered interventions for LBP to apply to our data set.

Methods: we developed a review protocol detailing the inclusion and exclusion criteria, search strategy, data extraction process and quality assessment method. We conducted electronic searches in MEDLINE, EMBASE, Web of Science (Science Citation Index and Social Science Citation Index) and Cochrane Central Register of Controlled Trials (CENTRAL) databases for studies reporting moderator analyses. Two researchers independently screened the titles and abstracts. Additionally, we searched the reference lists of relevant articles for any further potential references. We included RCTs with ≥ 500 participants, and cohort studies of ≥ 1000 participants. We classified potential moderators into those with strong evidence ($p < 0.05$) or weaker evidence ($p < 0.20$, $p \geq 0.05$).

Results: we identified 914 potential citations. We selected 64 papers for detailed evaluation. Four papers, all RCTs, were included. We identified potential moderators with strong evidence ($p < 0.05$) in one or more studies as age, employment status and type, back pain status, narcotic medication use, treatment expectations and education. Potential moderators with weaker evidence ($0.05 < p \leq 0.20$) include gender, psychological distress, pain/disability and quality of life.

Conclusion: the overall data obtained from this review were weak and lacked the rigour to inform clinical practice. However, this review has helped us to identify potential moderators of treatment effect with some weak evidence to inform our further analyses.

Background

The ability to identify which patients are likely to gain the greatest benefit from a treatment would have significant implications in clinical practice. To explore this it is crucial to identify moderators of treatment response. These are factors measured prior to randomisation and subsequently influence the effect of the treatment.⁴⁰ To identify such moderators, large data sets are required to provide sufficient statistical power to detect any interaction between the moderator and treatment.⁴¹

Aims

The purpose of this review was to identify potential moderators which we could test in our individual participant data pooled repository.

Method

Originally this review was conducted up until September 2011. Searches were updated in July 2014. Electronic searches were conducted using the following databases:

- MEDLINE
- Ovid MEDLINE® In-Process & Other Non-Indexed Citations
- EMBASE
- Web of Science
- Citation Index and CENTRAL.

To ensure that we had not overlooked useful data identifying possible treatment moderators, we searched for both RCTs and observational studies that had tested for effect modification.

Search strategy

We started our searches using the terms 'low back pain' combined with keywords including 'subgroup', 'effect modifier' and 'moderator'. The results from this preliminary search allowed identification only of publications that used the term 'subgroup' in the title and/or the abstract – it failed to pick up papers that used the term in the main body of the text. We therefore re-ran searches using keywords ('trial' for RCTs and ('Observational', 'Cohort', 'Prospective studies') for non-RCTs or observational studies separately and then combining them with terms 'low back pain'. Hand-searching and screening of included studies were carried out for additional studies.

Minimum sample size for included studies

To allow us to identify meaningful interactions it was critical to select research based on an adequate sample size. We made the following assumptions to determine the sample size criterion:

- the outcome of interest is continuous and normally distributed
- there are two treatment arms (intervention and control)
- the potential moderator is binary.

To determine the minimum sample needed to test for an interaction we used a model proposed by Lachenbruch.⁴² To test for a long-term (12 months) moderate standardised effect size [between-group difference/baseline standard deviation (SD)] of 0.5 for the interaction at a 0.05 level of significance and 80% power for the primary outcome, a minimum data set of 503 participants was needed. Recognising the inherent risk of bias in observational studies we set a higher threshold of 1000 participants for any observational studies included.

A priori we estimated that we needed to include RCTs with at least 500 participants to identify a moderate standardised mean difference (SMD; between-group difference/baseline SD) of 0.5 for the interaction at a 0.05 level of significance and 80% power. The SMDs in high-quality RCTs of therapist-delivered interventions for LBP are typically in the range of 0.1–0.7 (see *Table 1*). Smaller trials would be able to detect treatment moderation, at this level, only if the moderation effect was substantially larger than the main treatment effect. Thus, even having set quite a large entry criterion by size we would run the risk of failing to consider potential treatment effect moderators that did not reach the conventional level of statistical significance. Therefore, any variables identified as moderators of treatment effect at $p < 0.05$ were classed as potential moderators with strong evidence and those at $0.05 < p \leq 0.20$ as potential moderators with weak evidence. For our final analyses we considered potential moderators with both strong and weak evidence to be worth exploring further.

Inclusion and exclusion criteria

Box 1 provides an outline of the inclusion and exclusion criteria for this review.

BOX 1 Review 1: inclusion and exclusion criteria**Inclusion criteria**

- Aged ≥ 18 years.
- NSLBP of any duration.
- Therapist-delivered interventions.
- RCTs with sample size of ≥ 500 .
- Non-RCTs and observational studies with sample size of ≥ 1000 .
- English language.
- Primary and secondary analysis seeking to identify predictors of response to treatment using 'a priori' and 'post hoc' subgroups and those looking for interaction between baseline variable and treatment.

Exclusion criteria

- Studies with no comparison between two treatment groups.
- Studies that did not report effect sizes for treatment by using moderator interactions.

Screening and data extraction

At all stages two researchers (Dr Tara Gurung and DE) worked independently to screen titles and abstracts based on the inclusion criteria. All agreed full papers were obtained for data extraction. Data were extracted on to a standardised extraction form and any discrepancies were resolved using a third reviewer (DM). As no relevant observational studies were identified we do not address further methodological considerations related to observational studies.

Risk of bias and quality assessment

Both reviewers independently assessed risk of bias for the between-group comparison using the Cochrane Collaboration risk-of-bias tool.⁴³ From this tool the criteria used were:

- method of randomisation
- allocation concealment
- incomplete outcome data
- selective outcome reporting
- other sources of bias.

To assess quality we used the criteria developed by Pincus *et al.*,⁴⁴ whereby the answers to the five questions presented below allowed evidence to be classified as 'confirmatory' or 'exploratory':

1. Was the subgroup analysis specified a priori?
2. Was the selection of subgroup factors for analysis theory/evidence driven?
3. Were subgroup factors measured prior to randomisation?
4. Was measurement of subgroup factors, measured by adequate (reliable and valid) measurements, appropriate for the target population?
5. Does the analysis contain an explicit test of the interaction between moderator and treatment?

To reduce conflicts of interest, members of the reviewing team who were authors on any included studies did not participate in the quality assessment exercises.

Results

Our initial electronic searches generated 7208 hits; 6294 were removed based on title, abstract and duplicates. We obtained 64 papers for detailed review; of these, 60 papers were excluded (*Figure 2*). Four studies^{45–48} were included in this review (*Table 2*). All four trials^{45–48} were RCTs, constituting a total sample of $n = 5514$.

Once we had identified these papers we revisited our search results to include any studies with a sample size of ≥ 300 in a two-group comparison because the trial by Cherkin *et al.*⁴⁹ was a four-arm trial with a sample of $n = 638$, whereas our sample size calculation of ≥ 500 was based on a two-arm trial. As this paper⁴⁸ generated some useful moderators for our exploratory work we decided to include it. We did not identify any additional relevant studies with between 300 and 499 participants.

Although the Witt *et al.*⁴⁷ paper provided insufficient data to judge the quality of its exploratory analysis, it did include a specific test for interaction. The data presented did not allow for any pooling of moderator analyses across studies testing similar interventions.

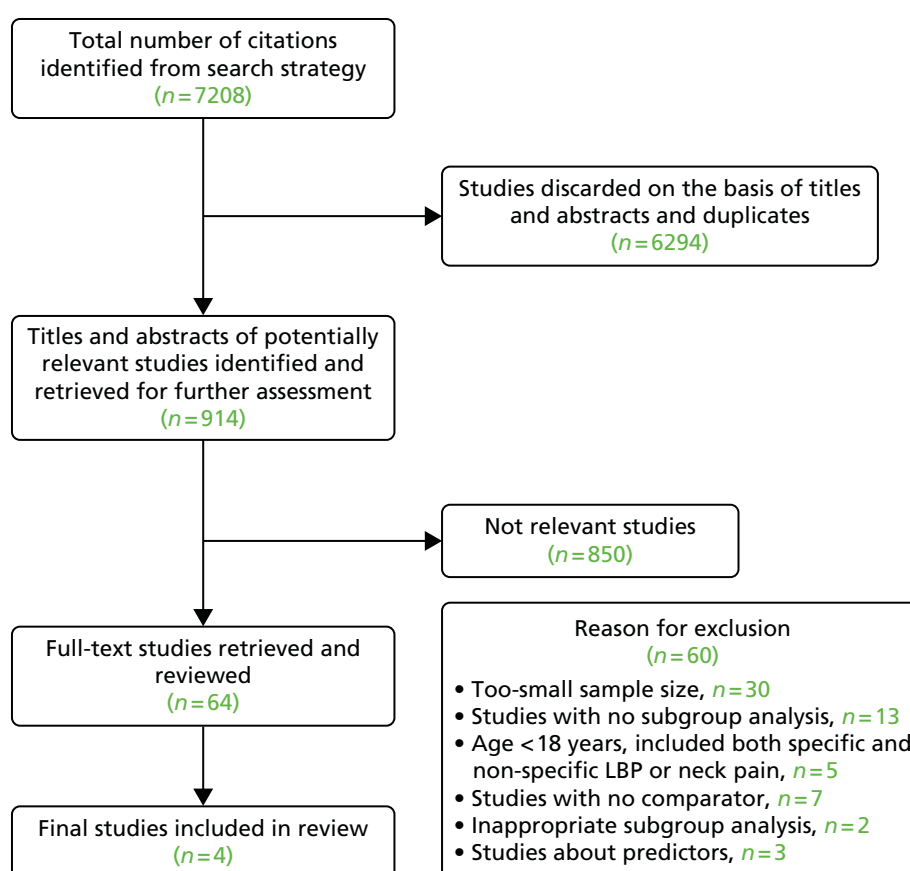


FIGURE 2 Review 1: Quorum statement flow diagram.

TABLE 2 Review 1: included studies

Study	Country	Sample	Interventions
UK BEAM ⁴⁵	UK	1334	Group exercise, manual therapy and combination therapy
BeST ⁴⁶	UK	701	Group cognitive-behavioural approach
Witt ⁴⁷	Germany	2841	Acupuncture
Cherkin ⁴⁹	USA	638	Acupuncture

BeST, Back Skills Training Trial; UK BEAM, UK Back pain Exercise And Manipulation.

Risk of bias and methodological quality for subgroups

To assess risk of bias and quality of subgroups we used both the original main trial papers and the associated secondary papers where appropriate (*Tables 3 and 4*).

Table 5 presents the potential moderators with strong and/or weak evidence from the four included trials.^{45–48} The many interactions tested that were not statistically significant are not reported here.

TABLE 3 Review 1: results of the risk of bias assessment

Quality of the study based on main trial paper(s)	UK BEAM ³¹	BeST ^{33,34}	Witt ⁴⁷	Cherkin ^{48,49}
Random sequence generation	L	L	L	L
Allocation concealment	L	L	L	L
Blinding of participants and personnel	H	H	H	H
Blinding of outcome assessment	L	L	H	L
Incomplete outcome data	L	L	U	L
Selective reporting	L	L	U	L
Generalisability	L	L	L	L
Sample size calculation	L	L	U	L
Conflict of interest	L	L	H	L
Source of funding	MRC	NIHR HTA	Social Health Fund Providers	National Institutes of Health

BeST, Back Skills Training Trial; H, high risk of bias; L, low risk of bias; MRC, Medical Research Council; NIHR HTA, National Institute for Health Research Health Technology Assessment; U, unclear; UK BEAM, UK Back pain Exercise And Manipulation.

TABLE 4 Review 1: results of methodological quality assessments

Quality of the moderator analyses based on subgroup paper(s)	UK BEAM ³¹	BeST ^{33,34}	Witt ⁴⁷	Cherkin ^{48,49}
Was the subgroup analysis specified a priori?	N	Y	N	N
Was the selection of subgroup factors for analysis theory/evidence driven?	N	Y	N	N
Were subgroup factors measured prior to randomisation?	Y	Y	U	Y
Was measurement of subgroup factors measured by adequate (reliable and valid) measurements, appropriate for the target population?	Y	Y	N	Y
Does the analysis contain an explicit test of the interaction between moderator and treatment?	Y	Y	U	Y
Strength of evidence	EE	CE for two potential moderators	IE	EE

BeST, Back Skills Training Trial; CE, confirmatory evidence – fulfils all five criteria for moderator studies; EE, exploratory evidence – fulfils three, four or five criteria for moderator studies; IE, insufficient evidence to judge quality; N, no; U, unclear; UK BEAM, UK Back pain Exercise And Manipulation; Y, yes.

TABLE 5 Mean difference (95% CI) of potential moderators with strong evidence ($p < 0.05$) and weak evidence ($p < 0.20 \geq 0.05$)

Study ID	Potential moderators	Significant interaction on selected outcomes (12 months)		
		RMDQ	MVK pain	MVK disability
BeST ^{3,4,6}	Troublesomeness (very/extremely – moderately)	$p = 0.190$; -1.01 (-2.52 to 0.50)	$p = 0.184$; -5.04 (-12.47 to 2.40)	NS
	Age (≥ 54 years – < 54 years)	$p = 0.035$; -1.58 (-3.05 to -0.12)	NS	NS
	Female – male	$p = 0.102$; -1.27 (-2.79 to 0.25)	NS	NS
	Left FT education (> 16 years of age – ≤ 16 years of age)	$p = 0.098$; 1.29 (-0.24 to 2.82)	NS	NS
	Employed – not employed	$p = 0.011$; 1.89 (0.43 to 3.35)	$p = 0.181$; 5.01 (-2.33 to 12.34)	NS
	HADS – anxiety (≥ 11 – < 11)	$p = 0.195$; -1.12 (-2.83 to 0.58)	NS	NS
	HADS – depression (≥ 11 – < 11)	$p = 0.135$; -2.07 (-4.79 to 0.65)	NS	$p = 0.051$; -14.58 (-29.19 to 0.03)

Significant interactions; outcome, RMDQ						
Study ID	Potential moderators	8 weeks		52 weeks		
		IA ^d	StA ^e	StA ^f	IA	StA
Cherkin ^{48,49}	Age	NS	$p=0.08$; 0.08 (-0.02 to 0.18)	NS	NS	$p=0.15$; 0.07 (-0.03 to 0.17)
	Self-efficacy	$p=0.04$; -6.17 (-12.01 to -0.33)	NS	NS	NS	NS
	RMDQ (B/L)	$p<0.0001$; -0.48 (-0.72 to -0.24)	$p=0.004$; -0.37 (-0.62 to -0.12)	$p=0.001$; -0.41 (-0.66 to -0.16)	$p=0.07$; -0.23 (-0.48 to 0.02)	$p=0.07$; -0.24 (-0.49 to 0.01)
	Bothersomeness score (B/L)	NS	$p=0.10$; 0.47 (-0.10 to -1.04)	NS	NS	NS
	Heavy lifting	$p=0.03$; 4.29 (0.43 to 8.15)	$p=0.13$; 3.00 (-0.86 to 6.86)	$p=0.18$; 2.73 (-1.27 to 6.73)	$p=0.01$; 5.19 (1.17 to 9.21)	$p=0.15$; 3.03 (-1.05 to 7.11)
	Sedentary	NS	NS	NS	$p=0.12$; 2.73 (-0.72 to 6.18)	$p=0.15$; 2.47 (-0.90 to 5.84)
	Use of narcotic medication	$p=0.08$; 3.52 (-0.38 to 7.42)	NS	$p=0.01$; 4.81 (0.97 to 8.65)	NS	$p=0.04$; 4.06 (0.18, 7.94)
	Acupuncture expectation (top tertile)	$p=0.05$; -2.65 (-5.28 to -0.02)	NS	NS	NS	$p=0.17$; -1.9 (-4.60 to 0.80)
						$p=0.19$; 2.71 (-1.31 to 6.73)
						$p=0.03$; -2.91 (-5.56 to -0.26)
continued						

TABLE 5 Mean difference (95% CI) of potential moderators with strong evidence ($p < 0.05$) and weak evidence ($p < 0.20 \leq 0.05$) (continued)

Study ID	Potential moderators	Significant interactions; outcome, bothersomeness score		
		8 weeks	52 weeks	
Cherkin ^{48,49}	Age	NS	NS	$p = 0.15$; 0.04 (-0.02 to 0.10)
	Self-efficacy	$p = 0.14$; -2.21 (-5.13 to 0.71)	NS	NS
	Baseline RMDQ score	$p = 0.01$; -0.15 (-0.27 to -0.03)	$p = 0.0005$; -0.22 (-0.34 to -0.10)	NS
	Heavy lifting	$p = 0.05$; 1.97 (0.03 to 3.91)	$p = 0.04$; 2.10 (0.10 to 4.10)	NS
	Light/medium lifting	NS	$p = 0.12$; 1.35 (-0.36 to 3.06)	NS
	Sedentary	NS	$p = 0.19$; 1.20 (-0.58 to 2.98)	NS
	Acupuncture expectation (top tertile)	$p = 0.10$; -1.10 (-2.41 to 0.21)	$p = 0.051$; -1.44 (-2.87 to -0.01)	$p = 0.06$; -1.29 (-2.64 to 0.06)

Study ID	Potential moderators	3 months for RMDQ outcome, combined treatment	12 months for RMDQ outcome, combined treatment
UK BEAM ^{31,45}	Quality of life	$p = 0.174$; -0.1 (-0.26 to 1.43)	NS
	Treatment expectation (helpful)	$p = 0.073$; -3.2 (-6.74 to 0.30)	$p = 0.038$; -3.8 (-7.39 to -0.20)
	Treatment expectation (very helpful)	$p = 0.192$; -2.2 (-5.49 to 1.11)	$p = 0.019$; -4.0 (-7.38 to -0.67)
	Manipulation		
	Beliefs	$p = 0.07$; -0.8 (-1.62 to 0.06)	NS
	Quality of life	$p = 0.118$; 1.4 (-0.35 to 3.07)	NS
	Pain/disability	$p = 0.176$; -1.9 (-4.61 to 0.85)	$p = 0.143$; -2.2 (-5.16 to 0.75)
	Treatment expectation (helpful)	NS	$p = 0.083$; -0.1 (-0.16 to 0.01)
	Treatment expectation (very helpful)	$p = 0.113$; 1.6 (-0.38 to 3.60)	NS
Study ID	Potential moderators	Outcome, FFbHR	
Witt ⁴⁷	Worse initial back function	$p < 0.001$	
		Back function and pain improvement at 3 months with acupuncture treatment	
	Younger	$p < 0.001$	
	> 10 years of schooling	$p = 0.01$	
BeST, Back Skills Training Trial; B/L, baseline; FFbHR, Hannover Functional Ability Questionnaire for measuring back pain-related functional limitations (Funktionsbeeinträchtigung durch Rückenschmerzen); HADS, Hospital Anxiety and Depression Scale; IA, individualised acupuncture; MVK, Modified von Korff; NS, no significant interaction found; SiA, simulation acupuncture; SiA, standardised acupuncture; UK BEAM, UK Back pain Exercise And Manipulation.			

Moderator variables identified

Potential moderators with strong evidence ($p < 0.05$) in one or more studies include age (younger participants may gain more benefit), employment status and type (those employed or in sedentary occupations may gain greater benefit), back pain status (those who are worse may gain greater benefit), narcotic medication use (users may benefit less), treatment expectations (those with a greater positive expectation gained more benefit) and education (those with > 10 years of schooling gained a greater benefit). Potential moderators with weaker evidence ($0.05 < p \leq 0.20$) include gender (female participants may gain greater benefit), psychological distress (those with anxiety and depressive symptoms may benefit more), pain/disability (those with greater pain/disability at baseline may benefit more) and quality of life (those with a better quality of life may benefit more). It should be noted that these findings might just be a chance finding, particularly as these conclusions come from different studies.

Age: the BeST (Back Skills Training Trial), Cherkin and Witt trials^{46,49,50} found an interaction with age. In the BeST trial,⁴⁶ younger participants gained more benefit from cognitive behavioural therapy than older participants on the RMDQ score. The treatment difference was -1.58 ($p = 0.035$; 95% CI -3.05 to -0.12). As the p -value was < 0.05 , the interactions provided strong evidence. Witt *et al.*⁵⁰ found a statistically significant additional benefit from acupuncture treatment in younger participants ($p < 0.001$).

Gender: the BeST trial⁴⁶ found that gender had a moderating effect on treatment. In this trial, females had comparatively greater improvement following group cognitive behavioural therapy than males. The treatment difference between male and female was -1.27 ($p = 0.102$; 95% CI -2.79 to 0.25) for the RMDQ score. As the p -value was $0.05 < p \leq 0.20$, the interaction provides weak evidence.

Employment status: employment was found to be one of the positive moderating factors. In the BeST trial,⁴⁶ the authors found that employed participants gained additional benefit from a cognitive behavioural approach compared with those who were unemployed. The treatment difference between employed and unemployed was 1.89 ($p = 0.011$; 95% CI 0.43 to 3.35) and 5.01 ($p = 0.181$; 95% CI -2.33 to 12.34) for the RMDQ and Modified von Korff (MVK) pain scores, respectively. The interaction effect in the analysis of the MVK pain score was weak.⁴⁶ The Cherkin trial^{48,49} found some moderating effect according to types of employment status. The participants in this trial⁴⁸ received acupuncture therapy. Those participants whose job involved heavy lifting showed positive moderating effect against back-related dysfunction score at 8 weeks ($p = 0.03$ to 0.18) and 52 weeks ($p = 0.01$ to 0.04). Those participants doing medium/light lifting at work showed positive moderating effect in terms of the bothersomeness score ($p = 0.12$) at 8 and 52 weeks; however, the interaction was weak. Finally, those participants with sedentary work showed positive moderating effect at 52 weeks ($p = 0.12$ to 0.19). The interaction was generally weak.

Education: the BeST trial⁴⁶ found that participants who had left full-time education after the age of 16 years had better improvement from cognitive behavioural therapy than participants who left full-time education aged ≤ 16 years. The treatment difference was 1.29 ($p = 0.098$; 95% CI -0.24 to 2.82) for the RMDQ score. The interaction effect was > 0.05 and, therefore, this provides weak evidence. Witt *et al.*⁵⁰ found that those participants who have had > 10 years of schooling gained a greater benefit from acupuncture ($p = 0.01$).

Back pain status: In the Cherkin and Witt trials⁴⁸⁻⁵⁰ participants with a worse initial back pain status (baseline RMDQ score) gained an increased benefit from acupuncture compared with those with a better back pain status at baseline (p -values ranged from < 0.001 to 0.16). The extent to which LBP inconveniences participants – how troublesome or bothersome it is – was found to be a moderator in two trials, with a greater benefit from treatment in those with a more troublesome/bothersome condition. The interaction was weak, with the p -values being > 0.05 . In the Cherkin trial,^{48,49} the p -value was 0.10 , whereas in the BeST trial⁴⁶ the treatment difference for the RMDQ score was -1.01 ($p = 0.190$; 95% CI -2.52 to 0.50) and -5.04 ($p = 1.184$; 95% CI -12.47 to 2.40) for MVK pain score.

Pain/disability: similarly, those participants with greater pain/disability at baseline seemed to benefit more at 3 months ($p = 0.176$) and 12 months ($p = 0.143$) for the RMDQ score with manipulation treatment [UK Back pain Exercise And Manipulation (UK BEAM⁴⁵)] (see *Table 5*). The p -values are > 0.05 and < 0.2 , therefore providing weak evidence.⁴⁵

Narcotic: Cherkin *et al.*^{48,49} found that use of medication such as narcotics had a negative moderating effect in those receiving acupuncture. The p -value for this interaction ranged from 0.01 to 0.19, demonstrating a spectrum of strong to weak evidence.

Treatment expectations: having better expectations about the treatment was found to be a moderating factor in two trials.^{45,48,49} The p -values ranged between 0.03 and 0.192, demonstrating a spectrum of strong to weak evidence for the interactions.^{48,49} Cherkin *et al.*^{48,49} found that participants with higher expectation of acupuncture treatment helpfulness gained more benefit in the back-related dysfunction score ($p = 0.03$ – 0.17) and bothersomeness score ($p = 0.05$ – 0.10).^{48,49} In the UK BEAM trial,⁴⁵ manipulation at 3 months ($p = 0.113$) and 12 months ($p = 0.083$), or a combined treatment of manipulation and exercise ($p = 0.03$ to 0.192) at both 3 and 12 months, showed positive moderating effect, as was demonstrated by the RMDQ score. Overall, the interactions were found to range between a spectrum of strong to weak evidence.

Quality of life: good quality of life showed weak evidence for a moderating effect on treatment outcome for both manipulation treatment ($p = 0.118$) and a combined manipulation and exercise treatment ($p = 0.174$).⁴⁵

Psychosocial status: in the BeST trial,⁴⁶ psychosocial status moderated treatment effect. The trial⁴⁶ investigated whether psychological status moderated better outcome from a cognitive behavioural therapy. Participants with higher levels of anxiety at baseline gained more benefit from treatment in terms of the RMDQ score. The treatment difference was found to be -1.12 ($p = 0.195$; 95% CI -2.83 to 0.58), demonstrating a weak interaction. Similarly, those participants who were depressed considerably gained more benefit from the treatment than those who were less depressed as was found in the RMDQ and MVK disability scores. The treatment difference was found to be -2.07 ($p = 0.135$; 95% CI -4.79 to 0.65) and -14.58 ($p = 0.051$; 95% CI -29.19 to 0.03) for the RMDQ and MVK disability scores, respectively.

Discussion and conclusion

In this review we aimed to identify potential moderators of treatment effect to test in our repository of data. Only four trials were included. We considered any variables that were identified as moderators of treatment effect at $p < 0.05$ as potential moderators with strong evidence, and those at $p < 0.20$ and $p \geq 0.05$ as potential moderators with weak evidence. Only for two comparisons, in one study,⁴⁶ were any confirmatory analyses performed. Any apparently positive findings need to be interpreted with considerable caution. We have set the threshold for potential moderation with weak evidence at $p = 0.02$, and the included studies included many comparisons, meaning that any positive results may well be no more than chance findings. Nevertheless, we have identified some domains for which there is some weak evidence of moderation that is worth exploring further.

Systematic review 2: quality of subgroup analyses in low back pain trials

This review has been published in *Spine*.⁵¹ Here we present a summary of the paper.

Abstract

Background: trials of back pain interventions have generally shown small to moderate positive effects. Therefore, identifying subgroups in this population is a research priority. This review evaluates the quality, conduct and reporting of subgroup analyses performed in the NSLBP literature.

Aim: to evaluate the quality, conduct and reporting of subgroup analyses performed in RCTs of therapist-delivered interventions for NSLBP.

Method: electronic databases were searched for RCTs of therapist-delivered interventions for NSLBP. We included papers reporting only subgroup analyses (confirmatory or exploratory). The quality of subgroup analyses and quality of conduct and reporting were also evaluated.

Results: thirty-nine papers were included in the final review. Of these, only three (8%) tested hypotheses about moderators (confirmatory findings); 18 (46%) generated hypotheses about moderators to inform future research (exploratory findings) and 18 (46%) provided insufficient findings. The appropriate statistical test for interaction was performed in 27 of the papers, of which 10 papers reported results from interaction tests, four papers incorrectly reported results within individual subgroups and the remaining papers either reported *p*-values or nothing at all.

Conclusions: subgroup analyses performed in NSLBP trials have been severely underpowered, are able to provide only exploratory or insufficient findings and have rather poor quality of reporting. Using current approaches, few definitive trials of subgrouping in back pain are very likely to be performed. There is a need to develop new approaches to subgroup identification in back pain research.

Background

The identification of subgroups that gain the most benefit from interventions for the management of LBP is an important research priority internationally.^{15,52–54} Although several trials claim to have performed subgroup analyses, the quality, conduct and reporting of the analyses performed has not been critically reviewed. There is some confusion in the papers between investigating ‘subgroup effects’ and investigating ‘differential subgroup effects’, where the former investigates a specific subset or subpopulation of the entire sample for a main effect and the latter investigates treatment effect heterogeneity using an interaction test between subgroups defined by factors measured prior to treatment.⁵⁵

Aims

The objective of this literature review is to first identify RCTs of therapist-delivered interventions for NSLBP, which have performed secondary analyses in the form of subgroup analyses. All identified literature was assessed using a set of methodological criteria to evaluate the quality of subgroup analyses. Furthermore, the conduct and reporting of subgroup analyses were also assessed.

Method

This literature review work was carried out as part of the PhD studentship funded in this programme of work.

The same search strategy described above in our previous review was used in this review to identify potential papers of RCTs looking at therapist-delivered interventions for LBP. Originally, the following databases were searched until September 2011. Searches were updated in July 2014. Electronic searches were conducted using the following databases:

- MEDLINE
- Ovid MEDLINE In-Process & Other Non-Indexed Citations
- EMBASE
- Web of Science
- Citation Index and CENTRAL.

Search strategy

As described above we started our searches using the terms 'low back pain' combined with keywords including 'subgroup', 'effect modifier' and 'moderator'. This only yielded publications which used the term 'subgroup' in the title and/or the abstract, it failed to pick up papers that used the term in the main body of the text. Therefore, we reran searches to identify all 'low back pain' and 'RCTs' which we filtered for therapist-delivered interventions.

Inclusion and exclusion criteria

Box 2 outlines the inclusion and exclusion criteria for this review.

Screening and data extraction

We screened titles and abstracts based on the predetermined inclusion criteria. We selected all papers potentially reporting subgroup analysis for further investigation. All agreed full papers were obtained for data extraction. Data were extracted on to a standardised extraction form and any discrepancies were resolved using a second reviewer.

Quality assessment of subgroup analysis

We used the same Pincus *et al.*⁴⁴ criteria described in the previous review (see *Risk of bias and quality assessment*, above) the review above to assess the quality of subgroups. Three independent reviewers (DM, SP and SWH) assessed the quality of the identified papers. All discrepancies were addressed and resolved through discussion.

To reduce conflicts of interest, members of the reviewing team who were authors on any included studies did not participate in the quality assessment exercises.

BOX 2 Review 2: inclusion and exclusion criteria

Inclusion criteria

- Randomised controlled trials.
- Participants aged 18 years or more with history of NSLBP.
- Therapist delivered interventions for NSLBP (including psychological interventions and intensive rehabilitation programmes).
- Primary or secondary analysis of RCTs reporting that a subgroup analysis had been conducted.

Exclusion criteria

- LBP with known likely cause (fracture, infection, malignancy specific cause, ankylosing spondylitis and other inflammatory disorders).
- Studies investigating disorders additional to NSLBP, e.g. NSLBP and neck pain.
- Outcome not a valid clinical measure of NSLBP, e.g. number of days sick leave.
- Testing a clinical prediction rule.
- Treatment effect modification over time, i.e. treatment × moderator × time.
- Pooled datasets of similar trials.

Reproduced from Mistry D, Patel S, Hee SW, Stallard N, Underwood M. Evaluating the quality of subgroup analyses in randomized controlled trials of therapist-delivered interventions for nonspecific low back pain: a systematic review. *Spine* 2014;**39**:618–29; with permission from Lippincott Williams & Wilkins.

Analysis

To assess the conduct and reporting of subgroup analysis we referred to existing authoritative reviews.^{56,57} Papers were assessed for:

- design and methods – for all papers
 - results
 - interpretation and discussion.
- } Only for those papers that used interaction tests for subgroup analyses.

Each paper was examined to see if it conformed to four key recommendations in the area of subgroup analyses (*Box 3*).

Results

Our initial search identified 5581 papers. All titles and abstracts were screened to identify potential papers reporting results of RCTs of therapist-delivered interventions for LBP. We excluded 5521 papers during the screening process. The full text for the remaining 60 papers was then thoroughly examined to look for subgroup analyses, of which 21 were excluded as they either did not meet the inclusion criteria or they met one or more of the exclusion criteria. We included 39 papers in the final review (*Figure 3*).

BOX 3 Key recommendations in the area of subgroup analyses

Key recommendations

- Exact subgroup definitions should be given beforehand for continuous and categorical variables, along with some justification to avoid post-hoc data dependent definitions of subgroups.
- Subgroup analyses should be performed on the primary outcome in the study. This is simply because trials are designed to detect differences in the primary outcome only; therefore, performing subgroup analyses on any other outcome measure will substantially reduce the power.
- A differential subgroup effect should be formally evaluated using a statistical test for interaction and the interaction effect reported. Performing tests within individual subgroups and then comparing the results is an incorrect approach to subgroup analyses as it does not directly evaluate the subgroup effect.
- The number of subgroup analyses to be performed should be kept to a minimum. This is to avoid the issue of false-positive discovery (type I error inflation) due to multiple testing; a well-known issue if there are several subgroups of interest. Any concerns regarding multiplicity should be acknowledged and addressed appropriately, e.g. applying a Bonferroni or Sidak correction.

Reproduced from Mistry D, Patel S, Hee SW, Stallard N, Underwood M. Evaluating the quality of subgroup analyses in randomized controlled trials of therapist-delivered interventions for nonspecific low back pain: a systematic review. *Spine* 2014;**39**:618–29; with permission from Lippincott Williams & Wilkins.

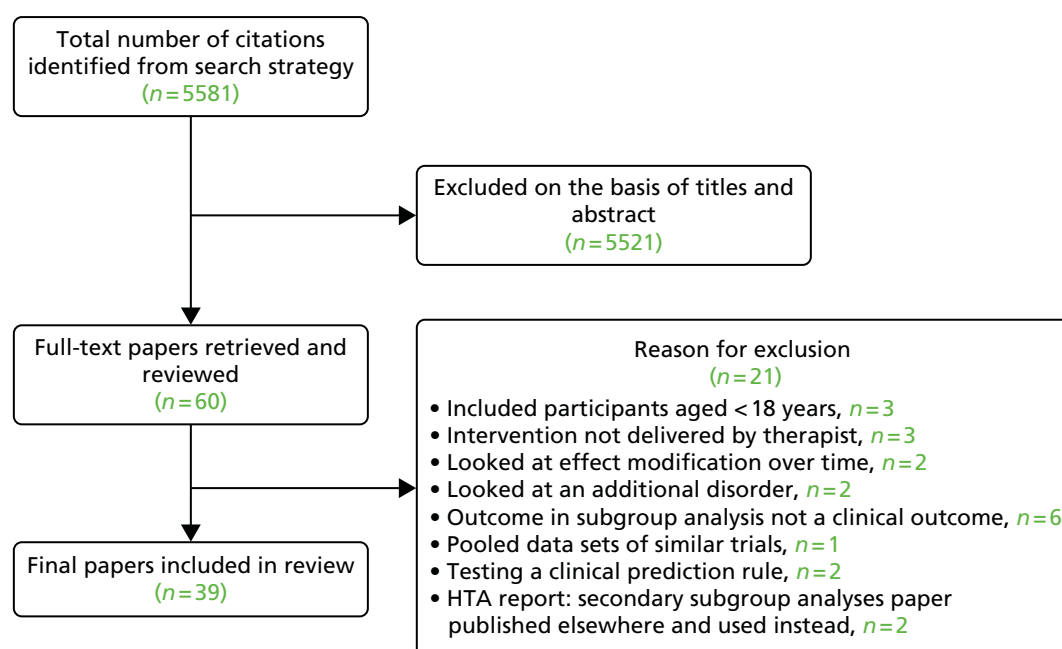


FIGURE 3 Review 2: Quorum statement flow diagram. Reproduced from Mistry D, Patel S, Hee SW, Stallard N, Underwood M. Evaluating the quality of subgroup analyses in randomized controlled trials of therapist-delivered interventions for nonspecific low back pain: a systematic review. *Spine* 2014;**39**:618–29; with permission from Lippincott Williams & Wilkins.

A summary of the included studies is given in *Table 6* and a summary of excluded studies can be found in *Appendix 1*. A total of 63% of the included papers were from the Netherlands, the UK or the USA. The median study size was 223, ranging from 100 to 3093.

Methodological quality of subgroup analyses

The methodological quality of the subgroup analyses performed in the identified papers was assessed to determine the strength of evidence that they provide. Of the 39 papers:^{35,45,46,48–50,58–90}

- Three (8%) papers^{46,53,54,58,59} met all five criteria and therefore provided confirmatory evidence. Two of these papers^{58,59} were too small to anticipate finding any important interaction if it were present ($n = 148$ and 259).
- Eighteen (46%) papers provided exploratory evidence, that is, they met criteria 3, 4 and 5 (see *Table 6*).
- Eighteen (46%) papers provided insufficient evidence (see *Table 6*).

Assessment of conduct and reporting of subgroups

We examined the conduct and reporting of subgroups in terms of design and methods and found that:

- One study⁵⁰ had sufficient power to detect an interaction; however, subgroups of interest were not prespecified a priori.
- Thirty-one (79%) studies^{35,45,48–50,60–63,66–74,76–78,80,81,83–90} did not prespecify subgroups of interest.
- Eight studies^{46,58,59,64,65,75,79,82} reported prespecified subgroups for confirmatory analyses; six of these studies also carried out exploratory analyses without clear distinction between analysis types.
- Sometimes it was not clear from the methods that subgroup analyses were going to be performed; they were just presented in the results.^{62,69,74,80}
- All papers measured subgroups of interest prior to randomisation, with most using adequate measurements.
- Prior to performing analyses, only one paper⁵⁸ reported the expected size and direction of the subgroup effect. A further three papers^{46,59,85} predicted the direction of the subgroup effect.

TABLE 6 Summary of included papers in descending order by subgroup quality assessment

Subgroup quality assessment	Author	Date of publication	Country	Study size	Interventions compared	Outcome measure and follow-up	Subgroups identified (interaction test only)
Confirmatory findings	Sheets ⁵⁸	2012	Australia	148	First-line care group vs. McKenzie group	Pain measured at 1 week and 3 weeks	None
	Smeets ⁵⁹	2009	Australia and New Zealand	259	Exercise and advice vs. exercise and sham advice vs. sham exercise and advice vs. sham exercise and sham advice	GPE at 3 weeks Pain intensity (11-point scale) and patient-specific function scale (0–10 scale) measured at baseline 6 weeks and 52 weeks	None
Exploratory findings	Underwood ⁴⁶	2011	UK	701	Advice plus cognitive-behavioural intervention vs. advice only	RMDQ and MVK scores measured at baseline and 3, 6 and 12 months	Age and employment
	Becker ⁶⁰	2008	Germany	1378	Multifaceted GI vs. GI plus MC vs. postal dissemination of guideline (control)	FFbHR measured at baseline and 6 months	None
	Cecchi ⁶¹	2012	Italy	210	Back school vs. individual physiotherapy vs. spinal manipulation	RMDQ score measured at baseline and 3, 6 and 12 months	None
	Cherkin ⁶²	1998	USA	321	Physical therapy vs. chiropractic manipulation vs. educational booklet	Bothersomeness of symptoms and RMDQ score measured at baseline, 4 weeks and 12 weeks	Mental health
	Cherkin ⁶³	2001	USA	262	Chinese acupuncture vs. therapeutic massage vs. self-care education	Bothersomeness of symptoms and RMDQ score measured at baseline, 4 weeks, 10 weeks and 1 year	None
	Cherkin ⁴⁹	2009	USA	638	Individualised acupuncture vs. StA vs. SiA vs. usual care	Bothersomeness of symptoms and RMDQ score measured at baseline, 8 weeks, 26 weeks and 1 year	None
	Hansen ⁶⁴	1993	Denmark	180	Intensive dynamic back-muscle exercise vs. conventional physiotherapy vs. placebo control (semi-hot packs and light traction)	Pain level (10-point scale) measured at baseline, 4 weeks, 6 weeks and 1 year	None
	Hay ⁶⁵	2005	UK	402	Brief pain management vs. manual physiotherapy	RMDQ score measured at baseline, 3 months and 12 months	None
	Junj ⁶⁶	2009	Switzerland	104	Standard care alone vs. standard care plus SMT	Pain intensity (11-point scale) and analgesic use measured at baseline, days 1–14 and 6 months	None

Subgroup quality assessment	Author	Date of publication	Country	Study size	Interventions compared	Outcome measure and follow-up	Subgroups identified (Interaction test only)
	Karjalainen ⁶⁷	2004	Finland	170	Mini-intervention group vs. worksite visit group vs. usual care group	Pain intensity (11-point scale) measured at baseline, 3 months, 6 months, 1 year and 2 years	Perceived risk for not recovering and type of occupation (comparing Mini-intervention vs. usual care and worksite visit vs. usual care)
	Kole-Snijders ⁶⁸	1999	Netherlands	159	OPCO vs. OPDI vs. WLC	Main outcome unclear	None
	Roche ⁶⁹	2007	France	132	AIP vs. FRP	Outcomes measured post treatment and at 6 months and 1 year Main outcome unclear	Sorenson score
	Sherman ⁴⁸	2009	USA	638	Individualised acupuncture vs. StA vs. SiA vs. usual care	Outcomes measured at baseline and 5 weeks	Baseline RMDQ score
	Smeets ⁷⁰	2006	Netherlands	223	APT vs. CBT vs. Combined APT and CBT (CTrt) vs. WL	Bothersomeness of symptoms and RMDQ score measured at baseline, 8 weeks, 26 weeks and 1 year	Baseline RMDQ
	Smeets ⁷¹	2008	Netherlands	223	ATP vs. GAP vs. CTrt vs. WL	RMDQ score measured at baseline, 10 weeks, 6 months and 12 months	None
	Tilbrook ³⁵	2011	UK	313	Yoga vs. usual care	RMDQ score measured at baseline, 10 weeks, 6 months and 12 months	None
	Underwood ⁴⁵	2007	UK	1334	Control (best care in general practice) vs. exercise programme vs. spinal manipulation vs. combined treatment (manipulation and exercise)	RMDQ score measured at baseline, 3, 6 and 12 months	Expectation
	Van der Hulst ⁷²	2008	Netherlands	163	RRP vs. usual care	RMDQ score measured at baseline, 1 week after treatment and 4 months after treatment	Pain intensity and depression
	Witt ⁵⁰	2006	Germany	3093	Acupuncture vs. control (delayed acupuncture treatment 3 months later)	FFbHR (0–100 scale) measured at baseline and 3 and 6 months	Initial back pain, age and years of schooling

continued

TABLE 6 Summary of included papers in descending order by subgroup quality assessment (continued)

Subgroup quality assessment	Author	Date of publication	Country	Study size	Interventions compared	Outcome measure and follow-up	Subgroups identified (interaction test only)
Insufficient findings	Bendix ⁷³	1998	Denmark	816	FRP programme vs. outpatients programme (control)	Main outcome unclear Outcomes measured at baseline and 1 year	
	Beurskens ⁷⁴	1995	Netherlands	151	Traction vs. sham traction	GPE and severity measured on VAS at baseline and 5 weeks	
	Bishop ⁷⁵	2011	USA	112	Supine thrust technique vs. side-lying thrust vs. non-thrust technique	ODI measured at 1 week, 4 weeks and 6 months	None
	Carl ⁷⁶	2005	UK	237	Group exercise programme vs. individual physiotherapy	RMDQ score measured at baseline, 3 months and 6 months	
	Ferreira ⁷⁷	2009	Australia	191	General exercise vs. motor control exercise vs. SMT	GPE (11-point scale), Patient specific functional status, RMDQ score, Pain intensity (10-point scale) and spinal stiffness measured at baseline and 8 weeks	None
	Glazov ⁷⁸	2009	Australia	100	Laser acupuncture vs. sham acupuncture (control)	Pain (VAS) measured at baseline, immediately after treatment, 6 weeks and 6 months	
	Gudavalli ⁷⁹	2006	USA	235	FD vs. ATEP	Perceived pain (VAS), RMDQ score and SF-36 measured at baseline, 4 weeks, 3 months, 6 months and 1 year	
	Hsieh ⁸⁰	2004	China	146	Acupressure vs. physical therapy	Short-form pain questionnaire measured at baseline, 4 weeks and 6 months	
	Jellema ⁸¹	2005	Netherlands	314	MIS vs. usual care	RMDQ score, perceived recovery (7-point scale) and sick leave measured at baseline; 6, 13 and 26 weeks; and 1 year	
	Johnson ⁸²	2007	UK	234	Group exercise and education using a cognitive behavioural approach vs. usual care	Pain (VAS) and RMDQ score measured at baseline and 3, 9 and 15 months	Patient preference
	Kalaokalan ⁸³	2001	USA	166	Acupuncture vs. massage (subanalysis of Cherkin 2001 paper)	RMDQ score measured at baseline, 4 weeks, 10 weeks and 1 year	Patient expectations

Subgroup quality assessment	Author	Date of publication	Country	Study size	Interventions compared	Outcome measure and follow-up	Subgroups identified (interaction test only)
	Melin ⁸⁴	1989	Finland	456	Inpatient treatment vs. outpatient treatment vs. control (advice)	LBP disability index (scale 0–45) measured at baseline and 3 months	
	Klaser Moffett ⁸⁵	2004	UK	187	Exercise vs. usual care	RMDQ score measured at baseline, 6 weeks, 6 months and 1 year	
	Myers ⁸⁶	2008	USA	444	Usual care vs. usual care plus patient choice of acupuncture, chiropractic or massage	RMDQ score measured at baseline, 5 weeks and 12 weeks	None
	Seferlis ⁸⁷	1998	Sweden	180	MTP vs. ITP vs. GPP	Main outcome unclear	
	Thomas ⁸⁸	2006	UK	241	Traditional acupuncture vs. usual care	Outcomes measured at baseline and 1, 3 and 12 months	Expectation
	Van der Roel ⁸⁹	2008	Netherlands	114	Intensive group training protocol vs. guideline group	Bodily pain dimension of the SF-36 (0–100 scale) measured at baseline and 3, 12 and 24 months	
	Vollenbroek-Hutten ⁹⁰	2004	Netherlands	163	RRP vs. usual care	RMDQ score measured at baseline and 6, 13, 26 and 52 weeks	
						RMDQ score measured at baseline, 1 week after treatment and 4 months after treatment	

AIP, active individual therapy; APT, active physical therapy; ATEP, active trunk exercise protocol; CBT, cognitive-behavioural treatment; CTrt, combination treatment; FD, flexion distraction; Hannover Functional Ability Questionnaire for measuring back pain-related functional limitations; FRP, functional restoration programme; GAP, graded activity with problem solving training; GI, guideline implementation; GPE, global perceived effect; GPP, general practitioner programme; ITP, intensive training programme; MC, motivational counselling; MIS, minimal intervention strategy; MTP, manual therapy programme; ODI, Oswestry Disability Index; OPCO, operant behavioural treatment with cognitive coping skills training; OPDI, operant behavioural treatment with group discussion; RRP, Roessing back rehabilitation; SF-36, Short Form questionnaire-36 items; SIA, simulation acupuncture; SMT, spinal manipulative therapy; StA, standardised acupuncture; VAS, visual analogue scale; WL, waiting list; WLC, waiting list control.

Reproduced from Mistry D, Patel S, Hee SW, Stallard N, Underwood M. Evaluating the quality of subgroup analyses in randomized controlled trials of therapist-delivered interventions for nonspecific low back pain: a systematic review. *Spine* 2014;**39**:618–29; with permission from Lippincott Williams & Wilkins.

- One-third (13/39) of the papers^{45,46,48,58,59,72,77,79,83,85–87,90} provided some justification regarding the choice of subgroups to be analysed.
- In two papers^{45,59} around 60 interaction tests were conducted, substantially increasing the chances of detecting false-positive findings. Of the three papers^{46,58,59} that provided confirmatory findings, only one of them⁴⁶ adjusted for multiplicity. The authors applied a Bonferroni correction to their confirmatory subgroup analyses.
- Twelve (31%) of the papers^{64,73,74,76,79–81,84,85,87,89,90} did not use a statistical test for interaction to assess for treatment effect modification. Of these, two of the papers^{74,87} did not give any indication as to what statistical method they used. Two papers^{73,84} looked at correlations between individual subgroups and outcomes within each treatment arm separately. Two papers^{79,80} used *t*-tests between treatment groups within individual subgroups. Five papers^{76,81,85,89,90} used either multiple linear regression or multiple logistic regression for each individual subgroup. One paper⁶⁴ compared the medians across three trial arms within individual subgroups using Kruskal–Wallis tests.

We examined the conduct and reporting of subgroups in terms of reporting of results and found that:

- A statistical test for interaction was reported to have been used in 27 (69%) of the papers.^{35,45,46,48–50,58–72,75,77,82,83,86,88}
- Six studies^{45,48,61,72,75,77} reported both the interaction effect sizes with CIs and the corresponding *p*-values.
- Four studies^{46,58,59,82} reported only the interaction effect sizes with CIs.
- Eight studies^{35,50,66,67,69,83,86,88} reported only the *p*-values.
- Nine papers^{49,60,62–65,68,70,71} did not report the interaction effect sizes, CIs or *p*-values.
- Four studies^{60,66,70,88} reported subgroup analyses within individual subgroups rather than between-group interaction.

We examined the conduct and reporting of subgroups in terms of reporting of interpretation and discussion and found that:

- Four^{60,66,70,88} out of 27 papers that performed interaction tests reported subgroup analyses within individual subgroups and thus based the interpretations and discussion on this as well.
- Reference to other relevant studies (supporting or contradicting) were made in around one-third of the papers.
- The limitations of subgroup analyses were reported in 12 papers.^{45,46,48,58–61,65,76,79,86,90}

Discussion and conclusion

Subgroup analyses have been attempted in several papers; however, there is confusion between investigating 'subgroup effects' and 'differential subgroup effects'.⁵⁵ The overall quality of the subgroups is poor, with most papers providing only exploratory or insufficient findings. The overall reporting in papers for subgroups is generally of poor standard. The sample sizes of the trials have been small and thus underpowered to detect interactions. Only one trial⁵⁰ was appropriately powered for the analysis; however, the authors failed to specify the subgroups a priori. The recommended guidelines should be used when performing subgroup analyses to ensure that they are reliable and of a good standard.^{56,91} The current approaches are not suitable to address the research question. New methods to perform subgroup analyses are required to address the methodological concerns highlighted.

Summary of reviews

Both reviews conducted during this programme of work have been informative in developing our understanding of subgrouping in LBP.

Review 1 looked at identifying potential moderators to be tested within the back pain repository. The literature on moderators is weak and, subsequently, lacking in rigour to inform clinical practice. Despite this, the review has helped us to identify some potential moderators of treatment effect, including age, educational attainment, employment status, symptoms of anxiety or depression, longer history of back pain and treatment expectations in at least one trial. We used these variables in our later analyses within our repository of data.

Review 2 looked at the quality of subgroup analyses conducted in the LBP literature. This review concluded that the overall quality was poor. A trial that is sufficiently powered to detect subgroups would need to be approximately four times larger than a traditional trial powered to detect a main effect of the same magnitude.⁹² This would be a timely and costly undertaking, for which care would also need to be taken to select moderators that were clinically relevant and applicable.

In addition to these reviews we have previously published a systematic review⁹³ that summarised findings from RCTs testing the effects of a clinical prediction rule for NSLBP. Clinical prediction rules have been developed and are being used in clinical practice to help clinicians to make decisions on treatment; however, the overall effect of such tools is unclear. Multicomponent clinical prediction rules have the potential to be much more powerful tools for targeting treatments than single-component measures. We identified 1821 potential citations after all duplications had been removed. Two reviewers independently screened the titles and abstracts, and consensus was reached on obtaining 35 papers for full detailed evaluation. Of these, only three papers^{94–96} were included in the review. The results from the available trials do not convincingly support the use of clinical prediction rules in the management of NSLBP. We concluded that the existing RCTs looking to validate clinical prediction rules in LBP are limited. Methodologies for the validation of these rules lack clarity and, subsequently, the evidence for, and development of, the existing prediction rules in LBP is generally weak.

Current approaches have failed to provide the data needed to target treatments for LBP. There is therefore a need to look at alternative methods to address this problem. We propose three recommendations:

1. To develop new and novel methods to identify multiple participant characteristics or clusters of moderators that would identify who is most or least likely to benefit.^{97–99}
2. To apply individual participant data meta-analysis to homogeneous pooled data sets, as this would improve statistical power.
3. To develop subgroups, and suggested interventions, based on clinical reasoning, and test these within trials to determine if the targeted intervention produces a larger average effect size than existing non-specific interventions.

In this programme we address points 1 and 2, leaving point 3 for others within the back pain research community to consider and address.

Chapter 3 Collating data

In this chapter we detail the process of identifying and approaching chief investigators and/or data custodians for trial data for inclusion in our repository of back pain trials.

Identification of potential trials

We used the search results generated from review 1 (Identification of potential moderators, described in *Chapter 2*) as a starting point for identifying trials of interest. In the first instance we were interested in only:

- RCTs
- trials of therapist-delivered interventions
- trials with a sample size of > 179 participants.

Based on these criteria we filtered the original search output to identify 658 citations. These were systematically screened by two members of the team independently (see *Figure 2*). Additionally, we also obtained further data through snowballing; essentially, we were offered data (from researchers aware of the project) from trials that were not on our original list. Although some of the trials obtained through the snowballing process are smaller in sample size than our target studies, we decided to include these to add power to our analysis.

Justification of sample size

We started with an original lower limit of 200 for the sample size. Allowing for some loss to follow-up, a trial of 200 participants would have 90% statistical power to identify a SMD of 0.5 between two treatment groups. Any individual trials smaller than this are likely to be seriously underpowered for their primary outcome. Upon screening the trials there were many that obtained a final sample size of just fewer than 200 participants; typically these were studies aiming for around 200 participants, which fell short of the final target. We therefore revised our inclusion to more than 179 participants. From a practical perspective of approaching trial investigators, this yielded a manageable number of trials to approach; large trials (those with thousands of participants) and small trials (fewer than 100 participants) each create a similar amount of work to collate.

Process for approaching investigators

We identified 42 trials^{33,49,50,62,63,65,70,76,82,84,94,100–130} that fitted our inclusion criteria. For these trials we identified the chief investigator and the best e-mail contacts for them. Between 2011 and 2012 each investigator was sent an e-mail to invite him/her to participate in the repository. Each e-mail included the following attachments:

- formal invitation letter (see *Appendix 2*)
- information sheet (see *Appendix 3*)
- sample data sharing agreement (see *Appendix 4*).

If a response was not received within a 6- to 8-week period, a reminder e-mail was then sent. If a response was received indicating an interest in sharing data then the data sharing agreement was personalised and sent back to the investigator for review and signature. Once the signed document was received by the university, the investigator was provided with details on how to securely send the data to us. We used the University of Warwick secure file transfer service.

Secure data transfer

We requested all data from a trial. Investigators were advised that any data sets being sent to us needed to be anonymised and encrypted using an open-source compression software programme such as 7-Zip 9.20 © Igor Pavlov (www.7-zip.org/). Investigators were then provided with details on how to securely transfer this data to the University of Warwick (see *Appendix 5*) using an upload system that was set up for the project (available at <https://files.warwick.ac.uk/repository/lbpdata/sendto>).

Once these data were received it was the responsibility of the team's statisticians and/or health economists to transform the original data to the repository standard. To aid this process we requested all trial-specific information, including the protocol and questionnaires if they were available.

Final data set obtained

We obtained 14 (33%) trial data sets^{31,33,50,65,70,76,101–107,131} from the original 42 trials^{33,49,50,62,63,65,70,76,82,84,94,100–130} we approached. A further five trials^{132–136} were obtained through snowballing, resulting in a total of 19 data sets (*Figure 4*). We were unsuccessful in getting a response from 15 (36%) investigators and a further six (14%) data sets were not available for data sharing. We still have seven (17%) data sets in negotiation, for which we were unable to agree on the data sharing before starting our formal analysis; therefore, these trials have not been included in this report.

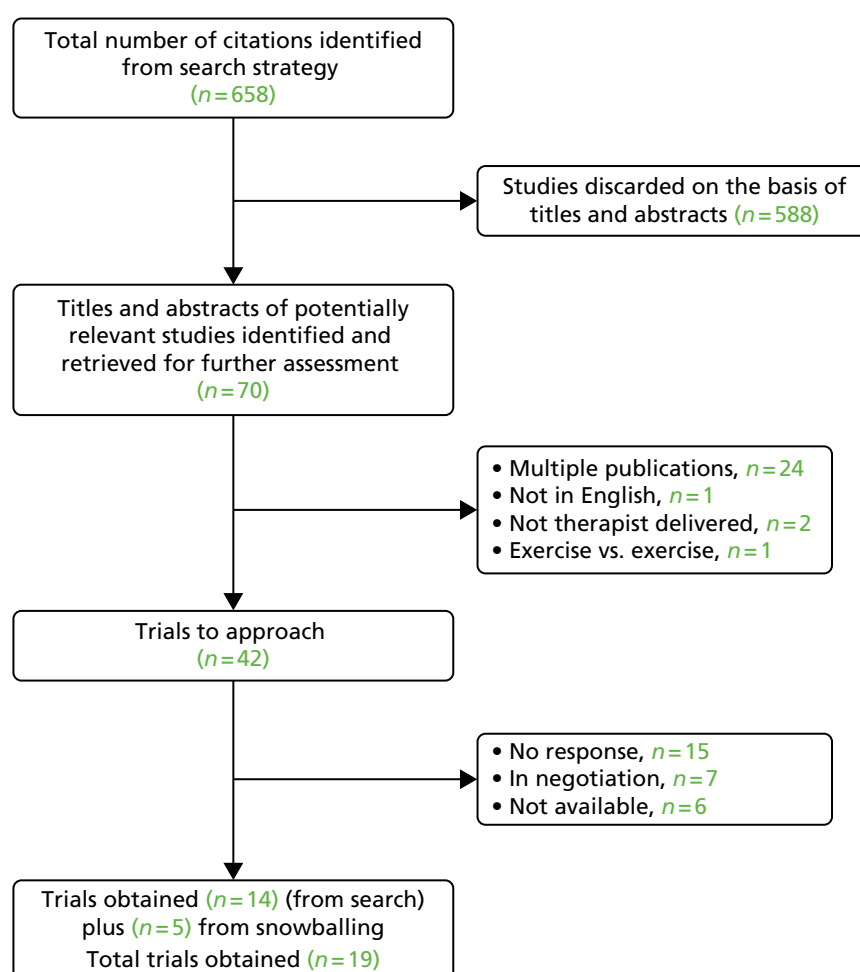


FIGURE 4 Quorum statement flow diagram for database identification.

Through the process of snowballing, further smaller data sets were offered to be included in the repository. The offer of these trials was carefully considered by the research team, and it was decided that any additional data would be helpful in increasing power. Therefore, three (16%) of the 19 trials obtained have a sample size of < 179 participants.

Table 7 shows the trials that were excluded and the reason for the exclusion. Details of papers excluded as a result of multiple publications can be found in *Appendix 6*. A list of trials that were unavailable because of a lack of response from the investigator, data sets not available and those still under negotiation are documented in *Appendix 7*. A final table of included trials and associated papers is presented in *Table 8*.

TABLE 7 Trials excluded and reason for exclusion, *n* = 4

Author	Number of participants	Reason for exclusion
Jellema ¹³⁷	314	Not therapist delivered
Kainz B ¹³⁸	1274	Paper not in English
Long A ¹³⁹	312	Trial of exercise vs. exercise
Von Korff ¹⁴⁰	255	Not therapist delivered

TABLE 8 Trials included and associated publications, *n* = 19

Name of/given name of trial	Corresponding author/chief investigator	Relevant publications related to the trial of interest	Number of participants
Witt	Witt	Witt CM, Jena S, Selim D, Brinkhaus B, Reinhold T, Wruck K, <i>et al.</i> Pragmatic randomized trial evaluating the clinical and economic effectiveness of acupuncture for chronic low back pain. <i>Am J Epidemiol</i> 2006; 164 :487–96 ⁵⁰	3093
UK BEAM	Underwood	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. <i>BMJ</i> 2004; 329 :1377 ³¹ Underwood MR, Morton V, Farrin A. Do baseline characteristics predict response to treatment for low back pain? Secondary analysis of the UK BEAM data set [ISRCTN32683578]. <i>Rheumatology</i> (Oxford) 2007; 46 :1297–302 ⁴⁵	1334
Haake	Haake	Haake M, Müller HH, Schade-Brittinger C, Basler HD, Schäfer H, Maier C, <i>et al.</i> Acupuncture Trials (GERAC) for chronic low back pain: randomized, multicenter, blinded, parallel-group trial with 3 groups. <i>Arch Intern Med</i> 2007; 167 :1892–8 ¹³²	1163
BeST	Lamb	Lamb SE, Hansen Z, Lall R, Castelnuovo E, Withers EJ, Nichols V, <i>et al.</i> Group cognitive behavioural treatment for low-back pain in primary care: a randomised controlled trial and cost-effectiveness analysis. <i>Lancet</i> 2010; 375 :916–23 ³³ Lamb SE, Lall R, Hansen Z, Castelnuovo E, Withers EJ, Nichols V, <i>et al.</i> A multicentred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain. The Back Skills Training (BeST) trial. <i>Health Technol Assess</i> 2010; 14 (41) ³⁴	701

continued

TABLE 8 Trials included and associated publications, *n* = 19 (continued)

Name of/given name of trial	Corresponding author/chief investigator	Relevant publications related to the trial of interest	Number of participants
Keele	Hay	Hay EM, Mullis R, Lewis M, Vohora K, Main CJ, Watson P, <i>et al.</i> Comparison of physical treatments versus a brief pain-management programme for back pain in primary care: a randomised clinical trial in physiotherapy practice. <i>Lancet</i> 2005; 365 :2024–30 ⁶⁵	402
		Whitehurst DG, Lewis M, Yao GL, Bryan S, Raftery JP, Mullis R, <i>et al.</i> A brief pain management program compared with physical therapy for low back pain: results from an economic analysis alongside a randomized clinical trial. <i>Arthritis Rheum</i> 2007; 57 :466–73 ¹⁴¹	
Brinkhaus	Brinkhaus	Brinkhaus B, Witt CM, Jena S, Linde K, Streng A, Wagenpfeil S, <i>et al.</i> Acupuncture in patients with chronic low back pain: a randomized controlled trial. <i>Arch Intern Med</i> 2006; 166 :450–7 ¹⁰¹	298
Dufour	Dufour	Dufour N, Thamsborg G, Oefeldt A, Lundsgaard C, Stender S. Treatment of chronic low back pain: a randomized, clinical trial comparing group-based multidisciplinary biopsychosocial rehabilitation and intensive individual therapist-assisted back muscle strengthening exercises. <i>Spine</i> 2010; 35 :469–76 ¹⁰²	286
Pengel	Pengel	Pengel LH, Refshauge KM, Maher CG, Nicholas MK, Herbert RD, McNair P. Physiotherapist-directed exercise, advice, or both for subacute low back pain: a randomized trial. <i>Ann Intern Med</i> 2007; 146 :787–96 ¹⁰³	260
		Smeets RJ, Maher CG, Nicholas MK, Refshauge KM, Herbert RD. Do psychological characteristics predict response to exercise and advice for subacute low back pain? <i>Arthritis Rheum</i> 2009; 61 :1202–9 ⁵⁹	
YACBAC	Thomas	Thomas KJ, MacPherson H, Thorpe L, Brazier J, Fitter M, Campbell MJ, <i>et al.</i> Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain. <i>BMJ</i> 2006; 333 :623 ⁸⁸	241
		Ratcliffe J, Thomas KJ, MacPherson H, Brazier J. A randomised controlled trial of acupuncture care for persistent low back pain: cost effectiveness analysis. <i>BMJ</i> 2006; 333 :626 ¹⁴²	
		Thomas KJ, MacPherson H, Ratcliffe J, Thorpe L, Brazier J, Campbell M, <i>et al.</i> Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain. <i>Health Technol Assess</i> 2005; 9 (32) ¹⁰⁷	
Hancock	Hancock	Hancock MJ, Maher CG, Latimer J, Herbert RD, McAuley JH. Independent evaluation of a clinical prediction rule for spinal manipulative therapy: a randomised controlled trial. <i>Eur Spine J</i> 2008; 17 :936–43 ⁹⁴	240
		Hancock MJ, Maher CG, Latimer J, Herbert RD, McAuley JH. Can rate of recovery be predicted in patients with acute low back pain? Development of a clinical prediction rule. <i>Eur J Pain</i> 2009; 13 :51–5 ¹⁴³	
		Hancock MJ, Maher CG, Latimer J, McLachlan AJ, Cooper CW, Day RO, <i>et al.</i> 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. <i>Lancet</i> 2007; 370 :1638–43 ¹³¹	

TABLE 8 Trials included and associated publications, *n* = 19 (continued)

Name of/given name of trial	Corresponding author/chief investigator	Relevant publications related to the trial of interest	Number of participants
VKBIA	Von Korff	Von Korff M, Balderson BH, Saunders K, Miglioretti DL, Lin EH, Berry S, <i>et al.</i> A trial of an activating intervention for chronic back pain in primary care and physical therapy settings. <i>Pain</i> 2005; 113 :323–30 ¹⁰⁴	240
HullExPro	Carr	Carr JL, Klaber MJA, Howarth E, Richmond SJ, Torgerson DJ, Jackson DA, <i>et al.</i> A randomized trial comparing a group exercise programme for back pain patients with individual physiotherapy in a severely deprived area. <i>Disabil Rehabil</i> 2005; 27 :929–37 ⁷⁶	237
VKSC2	Moore	Moore JE, von Korff M, Cherkin D, Saunders K, Lorig K. A randomized trial of a cognitive-behavioural program for enhancing back pain self care in a primary care setting. <i>Pain</i> 2000; 88 :145–53 ¹⁰⁵	226
Smeets	Smeets	Smeets RJ, Vlaeyen JW, Hidding A, Kester AD, van der Heijden GJ, van Geel AC, <i>et al.</i> Active rehabilitation for chronic low back pain: cognitive-behavioural, physical, or both? First direct post-treatment results from a randomized controlled trial [ISRCTN22714229]. <i>BMC Musculoskelet Disord</i> 2006; 7 :5 ⁷⁰	223
Cecchi	Cecchi	Cecchi F, Molino-Lova R, Chiti M, Pasquini G, Paperini A, Conti AA, <i>et al.</i> Spinal manipulation compared with back school and with individually delivered physiotherapy for the treatment of chronic low back pain: a randomized trial with one-year follow-up. <i>Clin Rehabil</i> 2010; 24 :26–36 ¹⁰⁶	210
York BP	Torgerson	Moffett JK, Torgerson D, Bell-Syer S, Jackson D, Llewellyn-Phillips H, Farrin A, <i>et al.</i> Randomised controlled trial of exercise for low back pain: clinical outcomes, costs, and preferences. <i>BMJ</i> 1999; 319 :279–83 ¹³³	187
Macedo	Macedo	Macedo LG, Latimer J, Maher CG, Hodges PW, McAuley JH, Nicholas MK, <i>et al.</i> Effect of motor control exercises versus graded activity in patients with chronic nonspecific low back pain: a randomized controlled trial. <i>Phys Ther</i> 2012; 92 :363–77 ¹³⁴	172
Carlsson	Carlsson	Carlsson CP, Sjölund BH. Acupuncture for chronic low back pain: a randomized placebo-controlled study with long-term follow-up. <i>Clin J Pain</i> 2001; 17 :296–305 ¹³⁵	50
Kennedy	Kennedy	Kennedy S, Baxter GD, Kerr DP, Bradbury I, Park J, McDonough SM. Acupuncture for acute non-specific low back pain: a pilot randomised non-penetrating sham controlled trial. <i>Complement Ther Med</i> 2008; 16 :139–46 ¹³⁶	48
VKBIA, von Korff BIA; VKSC2, von Korff SC2; YACBAC, York Acupuncture Back Pain Trial.			

Summary of the included trials in the repository

The agreed and included trials in this repository are detailed in *Table 9*.

TABLE 9 Summary of the included trials in the repository

Name of/given name of trial	Witt, <i>n</i> = 3093 ⁵⁰
Country	Germany
Interventions	In the RCT part of study there were two arms <ul style="list-style-type: none"> • Acupuncture • Control – received acupuncture after 3 months
Recruitment	Patients consulting a physician for LBP that were insured by one of the participating social health insurance funds were recruited. Details of the study were provided to those patients requesting acupuncture or when the physician considered acupuncture to be a suitable treatment option
Inclusion criteria	Age ≥ 18 years, with the ability to provide informed consent. A diagnosis of CLBP with a duration of more than 6 months
Exclusion criteria	Disc prolapse/protrusion of with concurrent neurological symptoms, previous back surgery, infectious spondylopathy, LBP caused by inflammatory, malignant or autoimmune disease, congenital deformation fracture caused by osteoporosis, spinal stenosis, and spondylolysis or spondylolisthesis
Name of/given name of trial	UK BEAM including feasibility study, <i>n</i> = 1334 ^{45,100}
Country	<ul style="list-style-type: none"> • UK
Interventions	<ul style="list-style-type: none"> • Exercise programme: group exercise, including cognitive behavioural principles, delivered over up to eight 60-minute sessions over 4–8 weeks. A refresher session was provided 12 weeks after randomisation • Spinal manipulation: a package of care was developed by chiropractors, osteopathic and physiotherapy professions in the UK. Patients were randomised to private or NHS manipulation. Up to eight 20-minute sessions were provided over 12 weeks • Combined treatment: provision of eight sessions of manipulation over 6 weeks plus eight sessions of exercise over the next 6 weeks plus a final refresher session at 12 weeks • Best care in general practice: patients were advised to remain active and provided with a copy of <i>The Back Book</i>¹⁴⁴
Recruitment	Recruited from GP practices after searching computerised records for potential eligible participants
Inclusion criteria	Aged between 18 and 65 years, consulted with LBP, score of ≥ 4 on RMDQ at randomisation, pain experienced every day for the 28 days before randomisation or 21 out of 28, agreement to avoid other physical treatments during the treatment period
Exclusion criteria	Aged ≥ 65 years, potential spinal disorder, including malignancy, osteoporosis, ankylosing spondylitis, cauda equina compression, and infection, pain primarily below the knee, previous spinal surgery, another musculoskeletal disorder reported to be more troublesome than the back pain, a previous referral or attendance at a pain management clinic, a severe psychiatric or psychological disorder, other medical condition that could interfere with therapy, moderate to severe hypertension, intake of anticoagulants or long-term steroids, inability to walk 100 m when free of back pain, inability to get up off the floor unaided, receipt of physical therapy in the preceding 3 months, RMDQ score of ≤ 3 on the day of randomisation, inability to read and write English fluently

TABLE 9 Summary of the included trials in the repository (*continued*)

Name of/given name of trial	Haake, <i>n</i> = 1163 ¹³²
Country	Germany
Interventions	<p>All groups received 10 30-minute sessions (two per week). Five additional sessions were offered if after the tenth session patients experienced a 10–50% reduction in pain intensity (von Korff CPG)</p> <ul style="list-style-type: none"> • Verum acupuncture: sterile disposable needles used to needle fixed points plus additional points from a prespecified list; 14–20 needles used and manual stimulation to elicit ‘de qi’ (an irradiating feeling) • Sham acupuncture: number and type of needles were the same as verum acupuncture. Needling of verum points or meridians was avoided and needles were inserted superficially and without stimulation • Conventional therapy: this was a multimodal treatment programme by which patients received 10 sessions with a physician or physiotherapist who administered physiotherapy and exercise
Recruitment	Patients were recruited through advertising in newspapers, magazines, radio and television
Inclusion criteria	<p>Aged ≥ 18 years with a clinical diagnosis of CLBP of ≥ 6 months, no previous experience of acupuncture for LBP</p> <p>Mean von Korff CPG score of ≥ 1 and a FFbHR score of $< 70\%$</p>
Exclusion criteria	Any previous spinal surgery or fractures, infectious or tumorous spondylopathy, and chronic pain caused by other diseases
Name of/given name of trial	BeST, <i>n</i> = 701 ^{33,34}
Country	<ul style="list-style-type: none"> • UK
Interventions	<ul style="list-style-type: none"> • Intervention arm: participants received an initial 15-minute advice session and were provided with <i>The Back Book</i>. Subsequently they attended six 1.5-hour group sessions, which covered cognitive behavioural topics • Control arm: participants received a 15-minute advice session and were provided with <i>The Back Book</i>
Recruitment	Recruited from GP practices after being identified from patient records or from consultation with the GP or practice nurse
Inclusion criteria	Aged ≥ 18 years, with at least moderately troublesome subacute or chronic LBP, with a minimum of 6 weeks’ duration, consultation with the GP for LBP within the preceding 6 months
Exclusion criteria	LBP related to a serious cause such as infection, fracture, malignancy, those with severe psychiatric or psychological disorders, and individuals with previous experience of a cognitive-behavioural intervention for LBP
Name of/given name of trial	Keele, <i>n</i> = 402 ^{65,141}
Country	UK
Interventions	<ul style="list-style-type: none"> • Brief pain management programme: patients were encouraged to return to normal activity using functional goal setting and strategies to overcome psychosocial barriers. A management plan was developed covering psychological, physical and functional topics. Exercises were undertaken both at the session and home • Manual physiotherapy: this was aimed at spinal manual therapy techniques. The aim was to diagnose and treat biomechanical dysfunction of the spine using manual therapy methods and exercises. An individualised home exercise programme was also provided
Recruitment	Recruited from GP practices
Inclusion criteria	Adults aged 18–64 years consulting with NSLBP of < 12 weeks’ duration for the first or second time, able to give informed consent
Exclusion criteria	Those with signs of red flags, sick leave of > 12 weeks, diagnosed with osteoporosis or inflammatory arthritis, taking systemic steroids for > 12 weeks, pregnant, previous fracture or hip/back surgery, any abdominal surgery in the preceding 3 months, receipt of treatment by any other professional for the current episode of back pain

continued

TABLE 9 Summary of the included trials in the repository (*continued*)

Name of/given name of trial	Brinkhaus, <i>n</i> = 298 ¹⁰¹
Country	Germany
Interventions	<p>The acupuncture and minimal acupuncture treatments consisted of 12 30-minute sessions delivered over 8 weeks</p> <ul style="list-style-type: none"> • Acupuncture treatment: this was semistandardised. Single-use sterile disposable needles were used. Physicians were instructed to achieve 'de qi' if possible. Manual stimulation of needles at least once during each session • Minimal acupuncture: therapist were advised to needle at least 6 of 10 predefined non-acupuncture points using a superficial insertion with fine needles. None of the points was in the area of the lower back; 'de qi' and manual stimulation of the needles were avoided • WL group: patients received acupuncture 8 weeks after randomisation. At this point they received 12 sessions as per the acupuncture treatment group
Recruitment	Primary recruitment method was via advertisement in local newspapers and subsequent snowballing
Inclusion criteria	Aged between 40 and 75 years, with a clinical diagnosis of chronic LBP present for > 6 months, a VAS of ≥ 40 for average pain intensity over the previous 7 days and the use of only oral NSAIDs in the four weeks preceding treatment
Exclusion criteria	Disc prolapse/protrusion with concurrent neurological symptoms; radicular pain, previous back surgery; infectious spondylopathy; LBP caused by inflammation, malignancy or autoimmune disease; congenital spine problems excluding minor lordosis or scoliosis; compression fracture caused by osteoporosis; spinal stenosis; spondylolysis or spondylolisthesis; those with diagnoses with Chinese medicine warranting treatment with moxibustion and receipt of acupuncture treatment in the preceding 12 months
Name of/given name of trial	Dufour, <i>n</i> = 286 ¹⁰²
Country	Denmark
Interventions	<ul style="list-style-type: none"> • Multidisciplinary biopsychosocial rehabilitation: 12-week programme split into three periods of 4 weeks <ul style="list-style-type: none"> ○ Period 1: exercise was performed three times per week in 2-hour sessions. Exercise comprised warm-up, stretching, aerobic training, and training to strengthen the muscles. Machines and circuit training were used. Biweekly session on anatomy, postural techniques and pain management was provided by a physiotherapist, and back care and lifting techniques by an occupational therapist ○ Period 2: twice-weekly 2-hour exercise sessions at the study site and once per week at home or a fitness centre ○ Period 3: three times per week, 2-hour exercise sessions at home or in a fitness centre • Individual strength training exercises encouraged by a specially trained therapist. Sessions ran for 1 hour, twice per week, for 12 weeks. The therapist initially supported the patient and then, over time, reduced the amount of assistance
Recruitment	Rheumatologists and GPs referred patients
Inclusion criteria	Patients aged 18–60 years with LBP of > 12 weeks with or without pain radiating into the leg(s). The lumbar spine was assessed through radiography, CT or MRI scans. Physical examinations were also used
Exclusion criteria	Those with symptoms of spinal pathology, including malignancy, osteoporosis, vertebral fracture and spinal stenosis, clinical symptoms of an acute herniated disc accompanied by nerve root entrapment, unstable spondylolisthesis, spondylitis, other health conditions preventing engagement in exercise and language problems

TABLE 9 Summary of the included trials in the repository (*continued*)

Name of/given name of trial	Pengel, <i>n</i> = 260 ^{59,103}
Country	Australia
Interventions	<ul style="list-style-type: none"> Exercise: individualised exercise programme using principles of cognitive behavioural therapy Sham exercise: sham pulsed ultrasonography and sham pulsed short-wave diathermy (neither provided output but acted as though it did) Advice: to address unhelpful beliefs and fear avoidance, and encourage return to normal activities Sham advice: in this session the participant was free to talk about his/her back pain and any other problems. The physiotherapist was emphatic but did not give advice
Recruitment	Recruited by referral to trial from health-care professional, invitation to those on a WL for physiotherapy and advert in newspaper
Inclusion criteria	Those aged 18–80 years, NSLBP lasting for at least 6 weeks but no longer than 12 weeks
Exclusion criteria	Those who have had spinal surgery in the past 12 months, any serious spinal abnormality, pregnancy, nerve root compromise, limited understanding of English and a contraindication to exercise
Name of/given name of trial	YACBAC, <i>n</i> = 241 ^{88,142}
Country	UK
Interventions	<ul style="list-style-type: none"> Traditional acupuncture: up to 10 sessions over 3 months Usual care: this group received treatment as usual determined by the GP
Recruitment	Recruited from GP practices
Inclusion criteria	18–65 years with non-specific LBP of 4–52 weeks' duration
Exclusion criteria	Patients currently having acupuncture, those with possible spinal disease, motor weakness, prolapsed central disc, past spinal surgery, bleeding disorders or pending litigation
Name of/given name of trial	Hancock, <i>n</i> = 240 ^{94,131,143}
Country	Australia
Interventions	<ul style="list-style-type: none"> Spinal manipulation: patients in this arm received two to three sessions of treatment per week, limited to a maximum of 12 treatments over 4 weeks. Manipulation was provided as per a protocol Placebo spinal manipulation: detuned pulsed ultrasound was used Both active and placebo manipulative therapy sessions were matched in time (30–40 minutes for initial session, followed by 20-minute follow-up sessions) <p>Four arms in the trial</p> <ul style="list-style-type: none"> SMT group (placebo drug and active SMT) SMT and NSAIDs group (diclofenac and active SMT) NSAIDs group (diclofenac and placebo spinal manipulation) control group (placebo drug and placebo SMT)
Recruitment	Recruited from GP practices
Inclusion criteria	Pain present in the region between the twelfth rib and buttock crease, causing moderate pain and moderate disability
Exclusion criteria	Present episode of pain not preceded by a pain-free period of at least 1 month, suspected or known serious spinal pathology; nerve root compromise; presently taking NSAIDs or undergoing spinal manipulation; any spinal surgery within the preceding 6 months; and contraindication to paracetamol/diclofenac or SMT

continued

TABLE 9 Summary of the included trials in the repository (*continued*)

Name of/given name of trial	VKBIA, <i>n</i> = 240 ¹⁰⁴
Country	USA
Interventions	<ul style="list-style-type: none"> Brief individualised programme: aimed to reduce fear and increase activity levels. This was delivered over four sessions, the first lasting 90 minutes with a psychologist, the second 60 minutes with a physiotherapist, the third lasting 30 minutes with a physiotherapist, and the final visit lasting 30 minutes with a psychologist. Intervention patients also received up to three bonus visits, a book on back pain self-management and video on back pain self-care Usual care: as provided to patients who were not participating in a trial. This care varied but included the use of medication, primary care consultations and secondary care referrals
Recruitment	Invitations were sent to patients who had consulted in primary care for their back pain and who were enrolled in the Group Health Cooperative
Inclusion criteria	Patients with back pain, aged 25–65 years, with a RMDQ score of ≥ 7 on a 23-item scale
Exclusion criteria	Those waiting for back surgery, seeing a physical therapist or psychologist, patients planning to unenroll from the Group Health Cooperative
Name of/given name of trial	HullExPro, <i>n</i> = 237 ⁷⁶
Country	UK
Interventions	<ul style="list-style-type: none"> Back to fitness exercise programme: patients were invited to attend eight one hour sessions aimed at increasing activity over a 4-week period. There was an underpinning cognitive behavioural approach Individual physiotherapy: treatments were provided at the discretion of the therapist
Recruitment	Physiotherapy departments at acute hospitals
Inclusion criteria	Those with mechanical LBP lasting at least 6 weeks
Exclusion criteria	Those with sciatica, recent significant surgery, the presence of a neurological or systemic condition, psychiatric illness or pregnancy; individuals who have had spinal surgery, in receipt of physiotherapy in the 6 weeks prior
Name of/given name of trial	VKSC2, <i>n</i> = 226 ¹⁰⁵
Country	USA
Interventions	<ul style="list-style-type: none"> Self-care arm: this was a group intervention of between 12 and 16 patients delivered over two, 2-hour sessions led by a psychologists covering a range of topics. Each patient had an individual 45-minute session with the psychologist to develop a personal self-care plan. Patients also received one brief follow-up telephone call to encourage continued action on the self-care plan. Patients were also provided with book on managing back pain, 40-minute videotape on back pain self-care and a 25-minute videotape demonstrating exercises Usual care group: received usual care plus a book on back pain
Recruitment	Patients were recruited from primary care by mail 6–8 weeks after a back pain visit to a Group Health primary care physician
Inclusion criteria	Patients with back pain, aged 25–70 years; patients who had been enrolled into Group Health for at least 1 year
Exclusion criteria	Those being considered for surgery

TABLE 9 Summary of the included trials in the repository (*continued*)

Name of/given name of trial	
Smeets, <i>n</i> = 223 ⁷⁰	
Country	The Netherlands
Interventions	<ul style="list-style-type: none"> • APT: this consisted of aerobic and strength training. This was delivered by two physiotherapists in a maximum group of four. Sessions were delivered three times per week lasting 1 hour and 45 minutes • CBT: this aimed to help patients reach their goals, manage beliefs and increase activity levels. Therapists used graded activity and problem-solving training • APT: aimed at increasing aerobic capacity and muscle conditioning • CBT: aimed at helping individuals reach their goals to increase activity levels and manage beliefs. Graded activity was used to encourage gradual increase or pacing of activities important to them. The frequency of the sessions gradually decreased from three sessions to one session per week. In total, 11 half hours of treatment • CTTr: aim was to improve functioning by increasing fitness, behaviour change and management of beliefs. CTTr consisted of APT together with problem-solving training • WL: patients needed to wait 10 weeks before they were offered individual rehabilitation treatment. While on the WL patients were unable to have diagnostic or therapeutic procedures because of their CLBP
Recruitment	Patients referred for the first time to a rehabilitation centre by their GP or other medical professional were invited to the study
Inclusion criteria	Aged 18–65 years with CLBP of ≥ 3 months with or without radiation to leg, a RMDQ score of > 3 and ability to walk at least 100 m without interruption
Exclusion criteria	Vertebral fracture, spinal inflammatory disease, spinal infections or malignancy, current nerve root pathology, spondylolysis or spondylolisthesis, lumbar spondylodesis. A comorbidity preventing exercise, ongoing treatment or investigation for CLBP at the time of referral or a clear treatment preference. Use of other treatments for back pain except pain medication. Any psychopathology affecting ability to take part. Not proficient in Dutch, being pregnant and having substance abuse
Name of/given name of trial	
Cecchi, <i>n</i> = 210 ¹⁰⁶	
Country	Italy
Interventions	<p>All patients were given an educational booklet on the back</p> <ul style="list-style-type: none"> • Back school: 15 1-hour sessions delivered over 15 days. The first five sessions focused on back physiology and pathology. The remaining 10 sessions looked at relaxation techniques, group and individual exercises. Groups were made up of eight patients and two therapists • Individual physiotherapy: therapists were able to select from exercises in a protocol to suit the patient. There were 15 sessions lasting 60 minutes delivered over 15 days • Spinal manipulation: four to six weekly sessions of 20 minutes each over 4–6 weeks
Recruitment	Rehabilitation outpatient department by psychiatrists
Inclusion criteria	NSLBP over at least the last 6 months reported as present 'often' or 'always'
Exclusion criteria	Neurological signs or symptoms, spondylolisthesis, spinal stenosis, scoliosis of $> 20^\circ$, rheumatoid arthritis/spondylitis, previous vertebral fracture, psychiatric condition, cognitive impairment or pain-related litigation

continued

TABLE 9 Summary of the included trials in the repository (*continued*)

Name of/given name of trial	
York BP, <i>n</i> = 187 ¹³³	
Country	UK
Interventions	<ul style="list-style-type: none"> Exercise programme: delivered as a group intervention of eight 1-hour sessions over a 4-week period. The sessions comprised stretching, low-level aerobic exercises and strengthening. The programme used cognitive behavioural principles and patients were encouraged increase their activity levels Controls: patients received usual care from their GP
Recruitment	Recruited from GP practices
Inclusion criteria	Patients aged between 18 and 60 years with LBP that has lasted at least 4 weeks but < 6 months, who had consulted their GP. Patients had to be deemed fit to be able to undertake exercise
Exclusion criteria	Those with a potentially serious pathology, unable to attend or participate in the classes, and those receiving ongoing physiotherapy
Name of/given name of trial	
Macedo, <i>n</i> = 172 ¹³⁴	
Country	Australia
Interventions	<p>In both arms patients received 12 1-hour sessions over an 8-week period. Home exercises were encouraged in both groups. The home exercises and treatment sessions totalled 20 hours</p> <ul style="list-style-type: none"> Graded activity: the aim of graded activity was to get patients to engage in activities that they found difficult because of back pain. Patients were provided with an individualised progressively increasing exercise programme to address functional problems. A cognitive-behavioural approach was used by the physiotherapist Motor control exercise: the aim is to retain optimal control and coordination of the lumbar spine and pelvis. Stage 1 involves regaining basic control strategies. In stage 2 participants progress through to more complex static and dynamic tasks, and training of functional activities. At all progressions the therapist evaluates and corrects trunk muscle recruitment strategies, posture, movement patterns and breathing
Recruitment	Recruitment via GPs, physiotherapists and public hospitals
Inclusion criteria	Aged 18–80 years with NSLBP of at least 3 months and seeking care. English speaking, living in the study region for the duration of the study, fit to engage in exercise, score of moderate or greater for amount of bodily pain in the past week, and interference of pain with normal activities
Exclusion criteria	Serious spinal pathology suspected or known, patients who have had spinal surgery or who are due to have such surgery during the study period, nerve root compromise, any comorbidities preventing participation in exercise
Name of/given name of trial	
Carlsson, <i>n</i> = 50 ¹³⁵	
Country	Sweden
Interventions	<ul style="list-style-type: none"> Manual acupuncture: needle acupuncture was used in predefined areas. There was a gradual increase in the number of needles from 8 to 14–18 during the first three or four treatments. The 'de qi' feeling was sought. Treatment sessions lasted 20 minutes and needles were stimulated on three occasions during this time Electroacupuncture: the first two or three sessions were manual acupuncture followed by treatments consisting of electrical stimulation of four needles in the low back. A similar number of needles as in the manual acupuncture group were inserted and manually activated Placebo stimulation: this was a mock TENS given by a disconnected stimulator. The area targeted was the most painful area in the low back. During the session patients were able to see a flashing lamp
Recruitment	Patients with CLBP, who were referred to an outpatient pain clinic during a 3-year period, were included
Inclusion criteria	Patients with LBP without radiation below the knee for > 6 months, normal neurological examination function of lumbosacral nerve
Exclusion criteria	Those who have had previous acupuncture treatment, patients with major trauma or systemic disease and pregnancy

TABLE 9 Summary of the included trials in the repository (*continued*)

Name of/given name of trial	Kennedy, <i>n</i> = 48 ¹³⁶
Country	UK
Interventions	<ul style="list-style-type: none"> • Verum acupuncture plus <i>The Back Book</i> – acupuncture was based on a ‘western’ approach. Between 3 and 12 sessions were provided over a 4- to 6-week period. At each session 8–13 needles were inserted and manually stimulated until ‘de qi’ was achieved • Sham acupuncture plus <i>The Back Book</i> – the Park Sham Device was used with acupuncture needles • Control intervention – the Park Sham Device was used with non-penetrating needles which touched the skin but did not penetrate the skin
Recruitment	Patients put on a WL for physiotherapy by their GP
Inclusion criteria	Adults aged 18–70 years, who are able to give informed consent with NSLBP, with or without referred pain, of up to 12 weeks’ duration
Exclusion criteria	Those with red flags, pain that has lasted for > 12 weeks, those with a contraindications to acupuncture or previous acupuncture treatment, any other conflicting or ongoing treatments

APT, active physical therapy; CBT, cognitive–behavioural treatment; CLBP, chronic low back pain; CPG, Chronic Pain Grade Scale; CT, computerised tomography; CTrt, combined treatment; FFbHR, Hannover Functional Ability Questionnaire for measuring back pain-related functional limitations (Funktionsbeeinträchtigung durch Rückenschmerzen); MRI, magnetic resonance imaging; NSAID, non-steroidal anti-inflammatory drug; SMT, spinal manipulative therapy; TENS, transcutaneous electrical nerve stimulation; VAS, visual analogue scale; WL, waiting list; YACBAC, York Acupuncture Back Pain Trial.

Grouping of interventions

Initial examination of the data showed that no two trials studied identical interventions. Even the usual-care arms of included studies are likely to differ according to jurisdiction, site of recruitment and age of the study. Even with our initial large sample size it was clear that, to be able to make meaningful comparisons, we would need to pool interventions into broad groups for our analyses. As a first stage we identified the control interventions and classified these as either usual care or a sham control. There is, for example, evidence from the acupuncture literature that the difference between sham acupuncture and usual care is greater than any difference between sham and verum acupuncture.¹⁴⁵ We therefore opted to separate the sham interventions from the usual care control in our analyses comparing different treatments with control or with each other.

There may be qualitative differences between sham treatments. For example, sham acupuncture, through which the participant has had the sensation of being needled, might have a different effect from a sham educational intervention. In some analyses we have included sham interventions, typically sham acupuncture as a separate category. For this reason we have, where appropriate, specified the nature of the sham intervention considered.

We used the following approach to develop our final grouping of interventions:

1. Careful reading of each trial intervention to decide on core groups [individual physiotherapy, exercise, manipulation, advice/education, psychological therapy, graded activity, acupuncture, combination therapy, mock transcutaneous electrical nerve stimulation (TENS), sham acupuncture and control]. We listed all of the trials contributing to each of the core groups together with the number of participants. Subsequently, links were made between core groups to indicate potential direct and indirect comparisons (*Figure 5*).
2. To explore further the potential direct and indirect comparisons, a second figure was constructed (*Figure 6*). This shows the same groups presented in the first step with the additional information on the number of trials and total number of participants contributing to each of the comparisons.
3. Finally, to allow for any meaningful comparisons, we split the groups mentioned in steps 1 and 2 into three broad categories, namely active physical (exercise and graded activity), passive physical (individual physiotherapy, manipulation and acupuncture) and psychological (advice/education and psychological therapy) (*Table 10*).

All trials

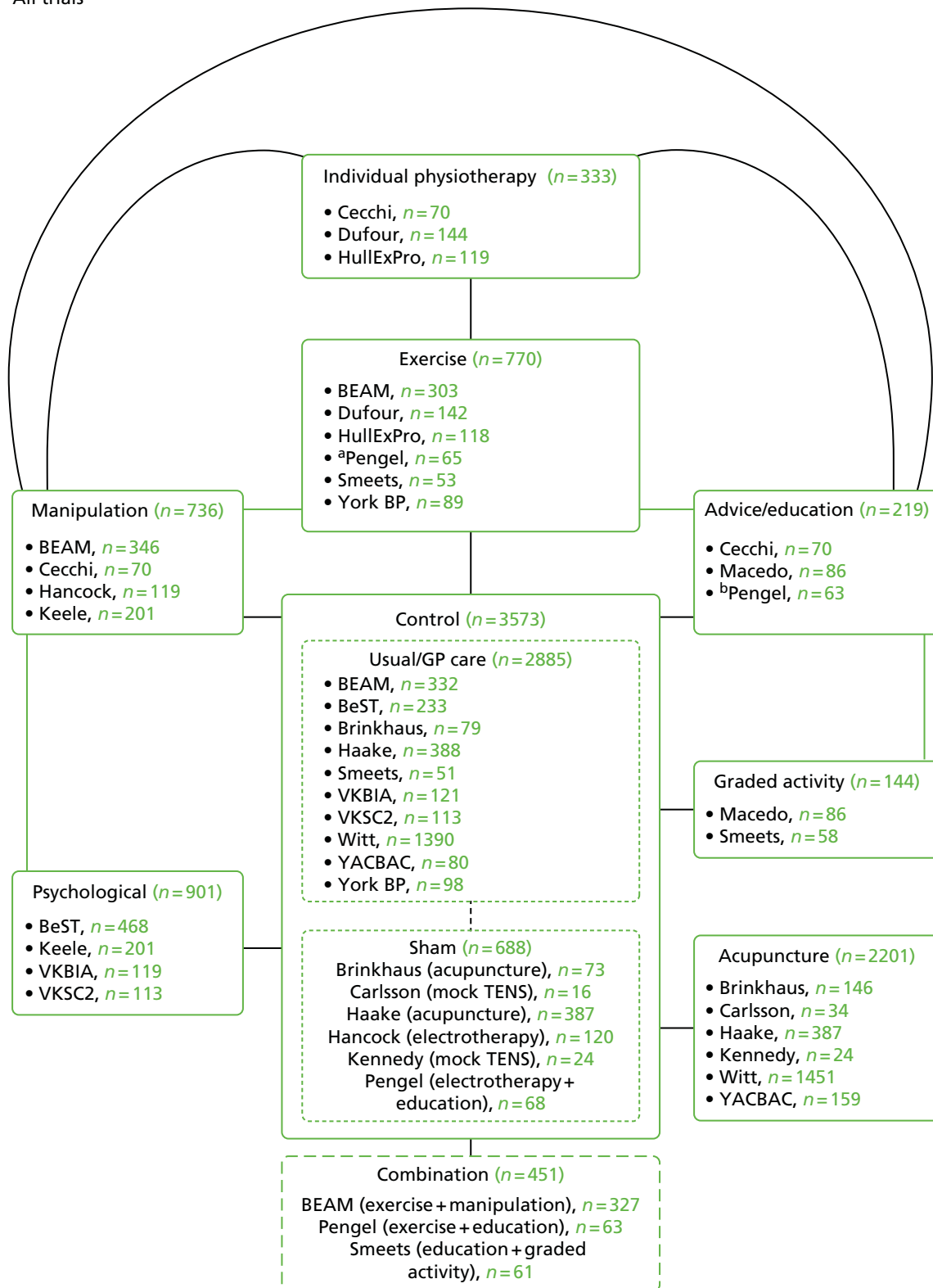
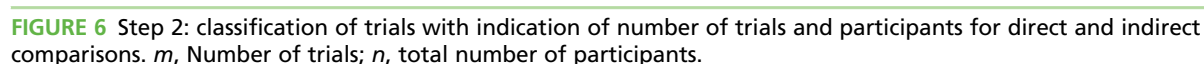


FIGURE 5 Step 1: classification of trials into core groups. a, Plus sham advice/education; b, plus sham electrotherapy.



Parent group	Subgroup	Subtype
Intervention	Active physical	Exercise
		Graded activity
	Passive physical	Acupuncture
		Manual therapy
		Individual physiotherapy
	Psychological	Advice/education
		Psychological (cognitive behavioural approach)
Sham control		Sham acupuncture
		Sham electrotherapy
		Mock TENS
		Sham advice/education
Control (GP/usual care)		GP
		WL

© Queen's Printer and Controller of HMSO 2016. This work was produced by Patel *et al.* under the terms of a commissioning contract issued by the Secretary of State for Health. This issue may be freely reproduced for the purposes of private research and study and extracts (or indeed, the full report) may be included in professional journals provided that suitable acknowledgement is made and the reproduction is not associated with any form of advertising. Applications for commercial reproduction should be addressed to: NIHR Journals Library, National Institute for Health Research, Evaluation, Trials and Studies Coordinating Centre, Alpha House, University of Southampton Science Park, Southampton SO16 7NS, UK.

In this programme of work we were not seeking to estimate the true effect size of any individual intervention. Rather, we were seeking to identify predictors of treatment response. These analyses were constrained by the availability of data on potential moderators that could be pooled across trials. Considering the potential mechanisms through which the potential moderators might affect outcome, the study team concluded that it was reasonable to pool interventions that might under other circumstances appear rather heterogeneous. In particular, the decision to include several superficially different interventions as passive physiotherapy might surprise some readers. Our view, however, is that these are very distinctly different from active exercise-based interventions, or those working through a psychological approach. Essentially, they all consist of an assessment, whatever reassurance and education is provided as part of the treatment session, plus whatever modality is being offered, be it massage/mobilisation/manipulation or needling. We consider these to be conceptually sufficiently close in their mode of action that it is unlikely there will be distinctions in how the potential moderators included in our analyses might affect outcomes. They are, however, distinctly different from their active physical or psychological interventions in how treatment moderation might operate.

In organising the data we also identified combined interventions but there were too few data points for it to be worthwhile pursuing these analyses. For this reason these were excluded from our final analyses.

Chapter 4 Creating the repository database and data control

Typographical conventions

This chapter presents the methods we used to create the repository database. To distinguish database vocabulary and commands from regular texts, different typographical fonts are used. Database object-class vocabulary is printed in sans-serif font [like this] and the command for mapping and transformation procedures is printed in monospaced typewriter font [like this]. In addition, coloured command fonts in the text are for ease of referencing between program commands shown in figures and text explanations.

Background

Clinical trial data sets can be stored in a tabular format, for example Microsoft Excel® or SPSS (Statistical Package for the Social Sciences). A tabular format typically uses each row to represent data from a participant and each column to represent an item from a case report form (CRF).

Tabular formats have the advantage of being intuitive, relatively simple to create and machine readable. However, this format can be susceptible to excessive growth, especially when clinical and non-clinical items are measured across multiple time points. Data collected for withdrawn participants or non-responders would still require columns for all variables irrespective of whether or not they were used. Repeating questions pose a similar problem, whereby storage space must be allocated across the whole domain to accommodate all responses. For example, asking for a participant's medical history of prescribed drugs would require a new column to be added for every drug listed. If only one participant documented a long list of drugs then many columns would have to be created for all participants.

Tabular formats are effective for only the smallest of trials and quickly become inefficient and difficult to maintain when the range of data collected increases. For larger trials, a more robust solution is to use a relational database. The relational database model allows individual tables to be created for each CRF and for repeating sets of questions. Normalisation rules are often applied to define the columns for each table and the logical relationships are used to create table joins.¹⁴⁶

Figure 7 shows sample data in a tabular format and the normalised equivalent in a relational database. The sample data consist of the subject identification, recruitment date, demographic data and the RMDQ scores taken at baseline and at 3-month follow-up. The data are normalised into four tables, namely SUBJECT, DEMOGRAPHICS, RMDQ (for the RMDQ measurement) and FU (follow-up); the last is used to store the time points for each follow-up visit.

Each table has a primary key (PKey) column for storing a unique record identifier that is used as the basis for creating relationships between tables (see Figure 7b). The relationship between SUBJECT and DEMOGRAPHICS is one-to-zero-or-one, that is a subject can have zero or one demographic record. The PKey from the SUBJECT table is copied to the DEMOGRAPHICS tables, thereby creating a join using a shared value.

The relationship between SUBJECT and RMDQ is one-to-zero-or-many, that is, a subject can have zero or many RMDQ completed questionnaires. The FU table is joined to the RMDQ table using a one-to-zero-or-many relationship. This join allows a RMDQ score to be associated with either a baseline or a 3-month follow-up time point.

(a)

Subject ID	Recruitment date	Sex	Age	RMDQ at Baseline	RMDQ at 3 month
1000	23/04/2003	M	50	13	7
1001	29/05/2003	M	28	9	3
1002	16/06/2003	F	43	18	10

(b)

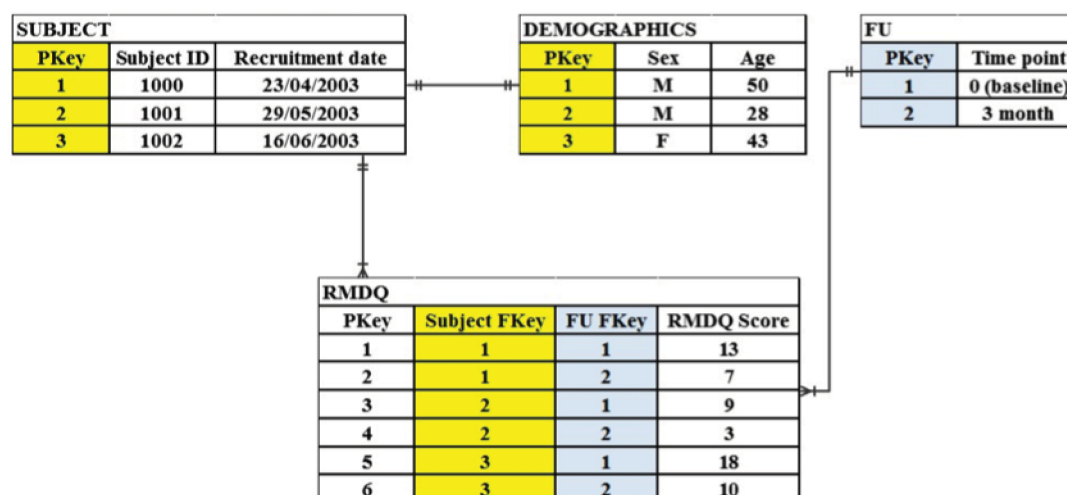


FIGURE 7 (a) A sample of original tabular format data; and (b) normalised relational interpretation of the original tabular data.

To create the relationships to the RMDQ table, the PKeys from both the SUBJECT and FU tables are added as foreign keys (FKeys). This has the result of allowing a subject to have either zero or many RMDQ scores at all time points. A composite unique constraint is applied to the Subject FKey and FU FKey columns to prevent a subject from having duplicate RMDQ scores for the same time point.

The repository differs from a typical clinical trial database in that it is not possible to predetermine requirements by using annotated CRFs. The repository relies on data from multiple trials to be periodically reviewed and classified, and must be frequently altered to accommodate new discoveries. The relational database is not a suitable model for such a scenario because modifications to the schema can be time-consuming and complex, often requiring the expertise of information technology specialists. Thus, the database for this project needs to be flexible so that the end users, namely, statisticians and health economists, can carry out modifications without having to change the database schema.

Our solution is to create a hybrid database that is a cross between an entity-attribute-value (EAV) open schema model and a relational database. This hybrid database has the flexibility of storing sparse heterogeneous data, which allows dynamic changes while enforcing data integrity.

The next section describes the architecture of the hybrid database. The rules used to map and transform the original source data to the repository standard are described below in *Mapping and transformation*. *Using entity-attribute-value data* shows how the repository database is manipulated, such that the data can be viewed in an analysis-friendly format from any statistical program that supports Open Database Connectivity (ODBC). *Extract, transform and load* describes how data from multiple RCTs were extracted, transformed and harmonised to the repository standard and, finally, loaded to the repository database.

System architecture

Tables and columns in a relational database can be represented as classes and attributes in an EAV model.¹⁴⁷ In the subsequent text the terms ‘class’ and ‘attribute’ will be used to conform to the EAV vocabulary. The term ‘entity’ is interchangeable with the term ‘object’ and can be thought of as providing a similar role to a table row but with the significant difference of storing only a pointer to the data and not the actual data itself. The entity–relationship diagram for the hybrid database is shown in *Figure 8*.

We anticipated that there would be some consistent data present in all of the RCTs for describing the trial and for identifying the trial’s subjects. The two tables **Primary Source** and **Subject** were created with fixed schemas to store this data (see *Figure 8*). The **Primary Source** table stores the name of the RCT (prms_TrialName), a brief description of the trial (prms_Description) and the date on which the data were imported into the repository (prms_ImportDate). The **Subject** table stores the original identifier assigned to the trial participant (subj_OriginalID), the date the participant enrolled into the trial (subj_EDate), the date the participant was randomised (subj_RDate) and a unique identifier generated by the system (subj_ID). A foreign key relationship is created to link each subject to the **Primary Source**.

The EAV model uses a subschema consisting of tables for classes, attributes, objects and the EAV data. The **Class** table is used to hold a list of all the identified domains, for example RMDQ and demographics. These domains generally map to a CRF but can also be used to describe a subset of repeating questions, for example repeated medical prescriptions.

The **Attribute** table is used to hold a list of all identified variables that typically map to a CRF question. The **Attribute** table has columns for storing a short name, a verbose name, a reference to the containing class and data type details. The short name is used to store a standardised version of the original CRF question.

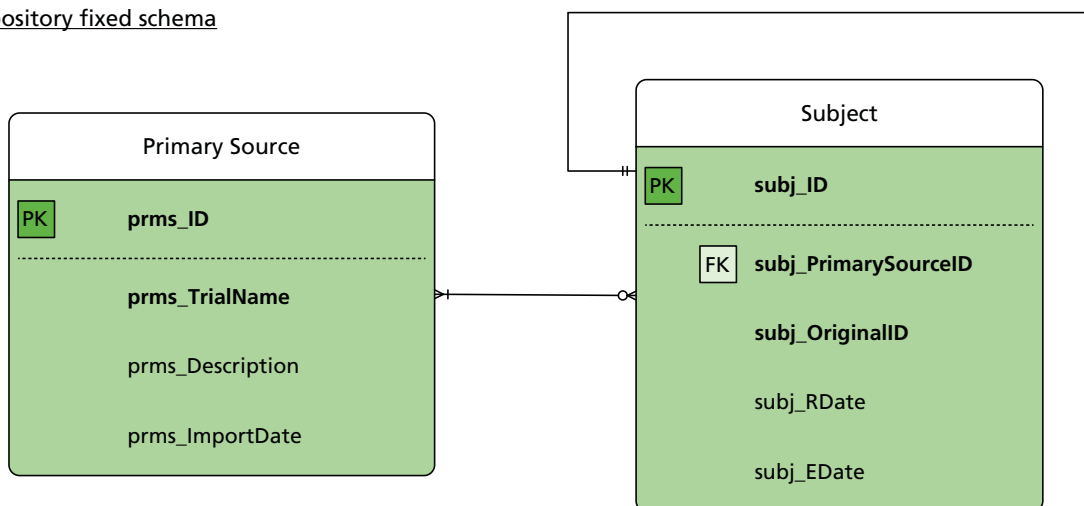
The **Object** table stores a unique identifier for each instance of a class and a reference to the class itself. A foreign key relationship is created to link each **Object** to a **Subject**. This relationship essentially makes the EAV model subject centric, that is, all of the data stored in the **Object** and EAV tables must be directly related to an imported subject. Relationship between objects is possible by using an ‘ancestor column’ to store the unique identifier of a related object. For example, an object used for repeated medical prescriptions will store the unique identifier of the related follow-up object in the ‘ancestor column’.

The EAV data table has three columns and is used to store all of the repository’s RCT data. Two columns hold references to the related objects and attributes, with the other column used for storing the actual value of each object–attribute combination. The references to the objects and attributes take the form of foreign keys to the object and attribute tables. The format of the value is coerced into a string regardless of the intended data type. The intended data type – for example binary data, small integers or strings – details are stored in the related attribute table.

A simplification of how tabular data are represented in an EAV table is shown in *Figure 9*. In this example, the tabular data have one row for each subject (see *Figure 9a*). When the data are shown in the EAV table there are four rows for subject #1000, three rows for subject #1001 and three rows for subject #1002. For each populated cell in the tabular data a row is created in the EAV table. Subject #1000 has all cells populated and, therefore, has a row for each entry. Only three rows are entered for the other subjects because there was no RMDQ baseline score for #1001 and age was not recorded for #1002 (see *Figure 9c*).

In reality the EAV table will use the column Attribute ID to store the unique attribute identifier and not the text value as shown in *Figure 9c*. In addition, the column Object ID stores a reference to the object and not the subject ID. It is the related object that links back to the subject and to the class.

Repository fixed schema



Repository EAV subschema

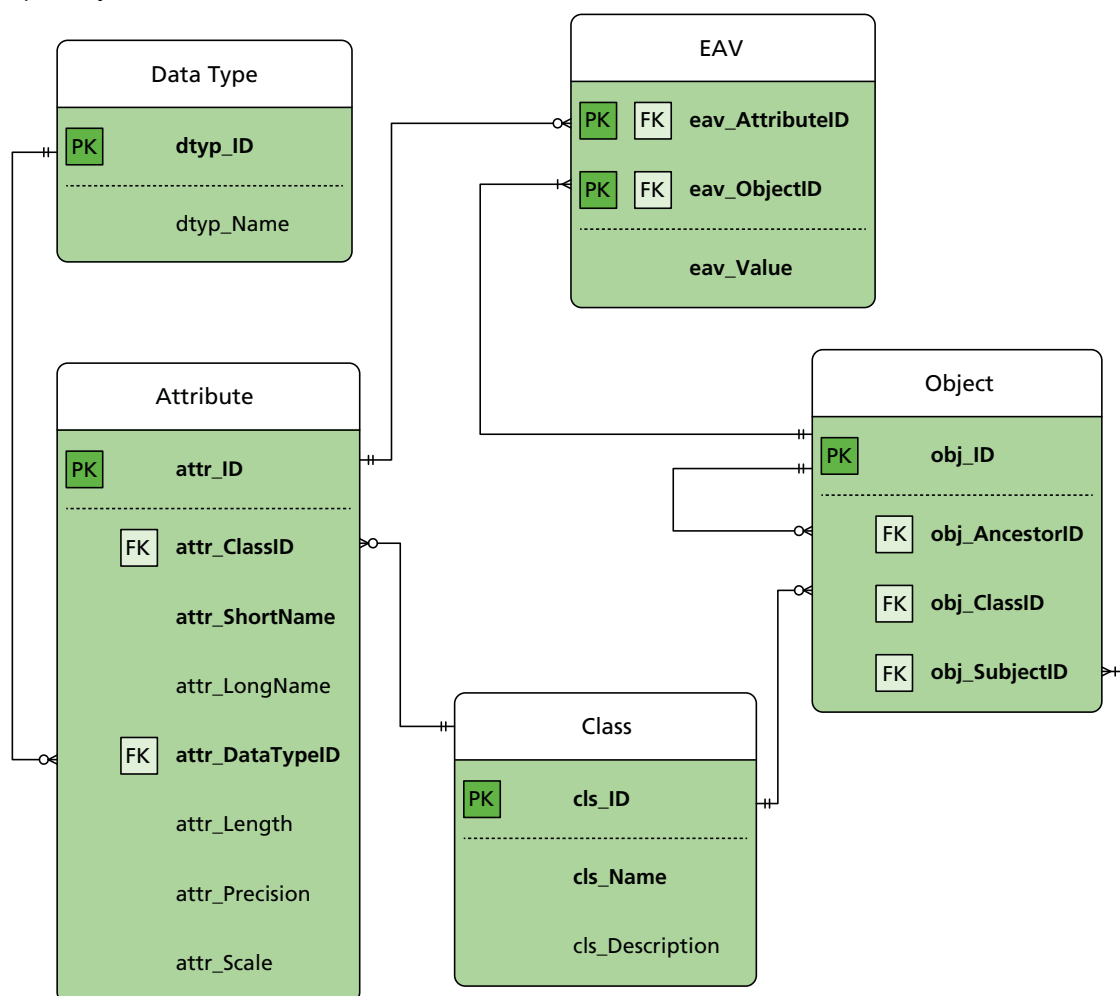


FIGURE 8 The entity-relationship diagram for the hybrid repository database depicting the fixed schema with the subschema EAV tables. Bold text represents required parameters.

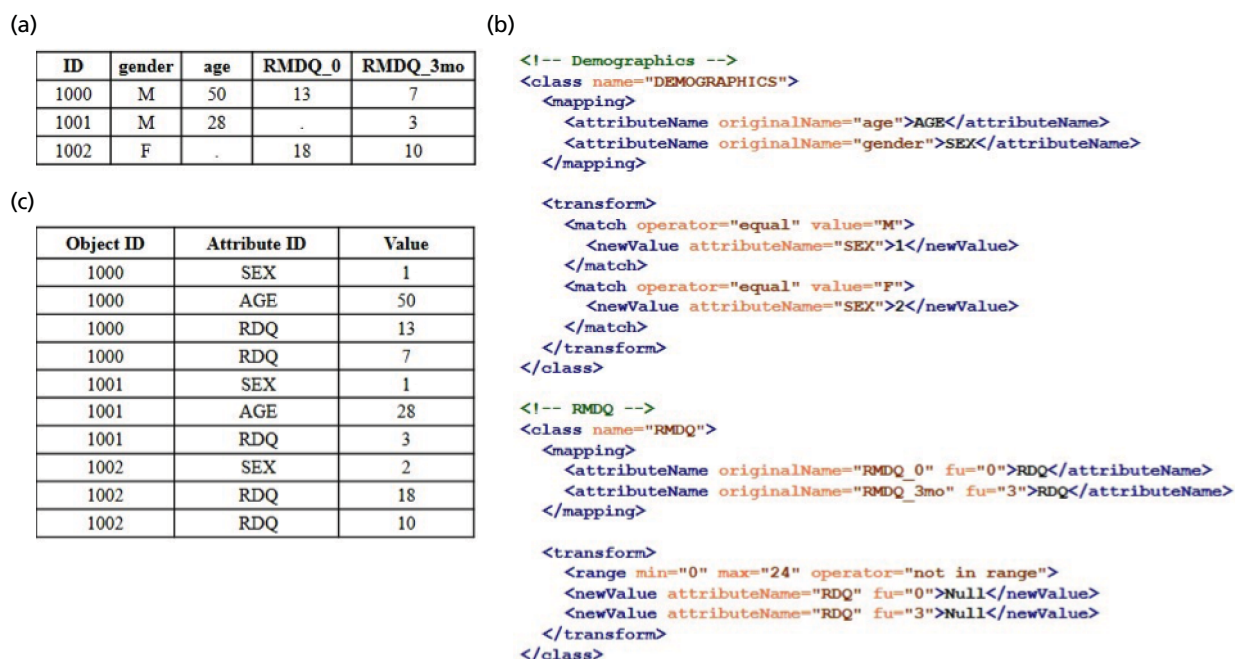


FIGURE 9 (a) A sample of original tabular format clinical data; and (b) the XML mapping and transformation instructions; and (c) the sample data represented as EAV. XML, extensible mark-up language.

Mapping and transformation

Early evaluation of data sets from various RCTs in the project identified large variations between variable naming and coding conventions. For example, the RMDQ was used to measure back pain disability and the participant would tick all of the items that were applicable to him/her on that day. There are 24 items in the questionnaire and the score is the sum of all of the ticked items. One trial might name each column 'rm1', 'rm2' and so on until 'rm24' for all 24 individual items and 'rmscore' as the RMDQ score measured at baseline, 'rm1_3mo', 'rm2_3mo', ..., 'rm24_3mo' and 'rmscore_3mo', for the 3-month follow-up data, and so on. Another trial might name them 'rdq1', 'rdq2', ..., 'rdq24' and 'rdq' for items measured at baseline, 'rdq1fu', 'rdq2fu', ..., 'rdq24fu' and 'rdq1fu' for items measured at the first follow-up, which could have been 1 or 3 months post randomisation, depending on the protocol. In addition, some trials might use the numerical value '1' to represent a tick for that item and '0' if it was not ticked. Other trials might use '1' as ticked and '2' as not.

Pilot mapping and transformation

A system was required to efficiently extract, transform and load (ETL) the original trial data sets into the repository. After evaluating a number of commercial and open-source ETL software packages, a prototype was developed using Microsoft SQL Server Integration Services (SSIS; SQL Server 2005 Enterprise Edition) and spreadsheets for documenting mapping and transformation instructions. The spreadsheet instructions were passed from the statisticians and health economists to the programmer who in turn created the SSIS program.

The pilot was deemed to be an inadequate solution. The versatility of SSIS as a data integration and transformation tool become a hindrance when attempting to customise a solution specifically for the repository. Setting up and configuring SSIS was found to be a laborious task, which was made even more difficult by frequent change requests and the manual interpretation of the mapping and transformation instructions. It became apparent that using SSIS was not viable and a decision was made to develop a bespoke ETL application.

XML and XSD for mapping and transforming

The method used to store mapping and transformation instructions was vastly improved by using extensible mark-up language (XML). XML is a free and open-source standard governed by the World Wide Web Consortium (WC3) and can be used to define a set of rules for encoding documents in a format that is both readable by human and machine.¹⁴⁸ The mapping and transformation XML document is made up of simple and intuitive keywords that both statisticians and health economists can easily interpret and apply. Having non-programmers directly enter the mapping and transformation rules forgoes the requirement to pass these instructions on to a programmer, which, in turn, saves resources and decreases misinterpretation errors.

To ensure that all mapping and transformation rules were specified in the correct format and the correct order, an XML schema definition (XSD) was applied to validate the XML document. The XSD is a separate document that defines the permitted structure of the XML document.

Mapping clinical data

Figure 9b shows an example of the XML mark-up to map the original data to the equivalent repository attributes. The standard attributes age and sex from the **DEMOGRAPHICS** class are mapped to the original variables **age** and **gender**. RMDQ scores for baseline and 3-month follow-up are mapped to the **RDQ** attribute from the **RMDQ** class.

The XML element `attributeName` accepts values for the original variable name (`originalName`) and the follow-up time point (`fu`) as XML attributes. The value of the `attributeName` XML element is set to the name of the repository attribute. In the example for class **RMDQ** the attribute name is **RDQ**.

Unlike in the original tabular data, the repository does not store different attribute names for each time point. Instead each time point will trigger a new object to be created. The XML `fu` attribute is used to track to which time point an original variable belongs.

Transforming clinical data

The original demographics and RMDQ scores have to be transformed into the repository standard before the data can be loaded into the repository database. Table 11 shows that the standard value for male is represented numerically by 1 and female is 2 for attribute **SEX**. Based on the same example (see Figure 9a), the values for male and female in the original data were entered as **M** and **F**, respectively. Thus, the transformation for the **SEX** attribute uses two match rules to find values **M** and **F**. When the value **M** is matched, the rule has been set to update the attribute's value to 1. Likewise, when the value **F** is matched, the attribute's value is updated to 2. There is no transformation rule for **AGE** attribute, as the repository accepts any valid integer value.

In the example for class **RMDQ**, the transformation uses a range rule to allow values of only between 0 and 24 to be imported. If any **RDQ** value falls outside this range then the system will transform the value to **Null** (empty).

Mapping and transforming health-care resource-use data

Mapping health-care resource-use variables was more challenging because the different types of resources used across all RCTs do not conform to any standard and are completely variable. However, each question and answer in a typical health-care resource-use questionnaire can be broken down to the recall period, the type of resource, the reason for using the resource, the location of the resource, the unit of measurement, the quantity, the cost or expenses incurred and the payer.

TABLE 11 A sample of the repository standard attributes and values

Class	Attribute short name	Attribute long name	Data type	Value	Label
DEMOGRAPHICS	SEX	Participant's sex	Integer	1	Male
				2	Female
DEMOGRAPHICS	AGE	Participant's age	Integer	> 0	
RMDQ	RDQ	RMDQ score	Integer	Range 0–24	
HE	RP	Recall period	Integer	> 0	
HE	TYPE	Types of resource	String	1a	Primary care doctor
				3a	Physiotherapist
				4M01	NSAIDs
				6	Aids and adaptations
HE	REASON	Resource reason	Integer	2	LBP
				4	Any condition
HE	LOCATION	Resource location	Integer	1	Primary care clinic
				3	Private clinic
				4	Community clinic
HE	UNIT	Resource units	Integer	1	Visit
				3	Prescription
				4	Item
HE	QUANTITY		Integer	> 0	
HE	COST		Integer	> 0	
HE	PAYER	Resource payer	Integer	1	Public health service
				4	Individual

HE, health economics; NSAID, non-steroidal anti-inflammatory drug; RP, recall period.

Figure 10 shows a simplified version of a typical health-care resource-use questionnaire. In this example, participants were asked to record all of the health-care resources that they used at the 3-month follow-up time point (see Figure 10a). The answers provided by the participants were stored in a tabular format, which used 12 columns to capture all of the responses to the five questions (see Figure 10b). By using this format, the number of required columns to accommodate the data would grow in line with the maximum number of responses provided by any one individual. For example, if only one participant listed three items he/she bought over the counter to treat his/her LBP, the number of columns required would have to be increased from 12 to 13.

Figure 10c shows a view of the repository health-care resources data, generated from the EAV tables. This view displays the eight standard repository health-care resource-use attributes (table columns) and an additional attribute called 'Text', which is used to store all of the characters that are captured as comments in the CRF.

The process for creating the transformed health-care resource-use data involves splitting the original questions into a number of derived parts that will map to the standard attributes. For example, question 1 asked how many times the participant had consulted his/her doctor or any primary care doctor, for any reason, in the last 3 months. Using the information contained in the question, the recall period is set to '3', the type of resource is 'GP', the reason for using the resource is 'Any condition', the location of the resource

(a)

Public healthcare professionals

	No. of visits
1. In the last 3 months, how many times have you consulted your doctor or another doctor for any reason?	<i>Pri1</i>
2. In the last 3 months, how many times have you consulted the NHS physiotherapist for low back pain?	<i>Pri2</i>

Private healthcare professionals

	No. of visits	Cost (£)
3. In the last 3 months, how many times have you consulted any private physiotherapist for low back pain?	<i>nPriv1</i>	<i>cPriv1</i>

Medication

4. In the last 3 months, list the medicines prescribed by your doctor for your low back pain.

Medicine prescribed	No. of prescriptions
i. <i>pmed1</i>	<i>nmed1</i>
ii. <i>pmed2</i>	<i>nmed2</i>

5. In the last 3 months, list the medicines or treatments that you personally bought over-the-counter to treat your low back pain.

Medicine bought	Cost (£)
i. <i>bmed1</i>	<i>cmed1</i>
ii. <i>bmed2</i>	<i>cmed2</i>

(b)

Subject

ID	Pri1	Pri2	nPriv1	cPriv1	pmed1	nmed1	pmed2	nmed2	bmed1	cmed1	bmed2	cmed2
1000	1	1	1	40	Ibuprofen	8.10						
1001	0	5	0	0					Mattress	325		
1002			3	120					Mattress	299	Walking stick	5.65

(c)

Qn.	FU	RP	Type	Text	Reason	Location	Unit	Quantity	Cost	Payer
1	3	3	GP	N/A	Any	Pri	Visit	<i>Pri1</i>	N/A	PHS
2	3	3	Physio	N/A	LBP	CC	Visit	<i>Pri2</i>	N/A	PHS
3	3	3	Physio	N/A	LBP	Priv	Visit	<i>nPriv1</i>	<i>cPriv1</i>	IND
4	3	3	<i>pmed1</i>	<i>pmed1</i>	LBP	N/A	Px	<i>nmed1</i>	N/A	PHS
4	3	3	<i>pmed2</i>	<i>pmed2</i>	LBP	N/A	Px	<i>nmed2</i>	N/A	PHS
5	3	3	Aid	<i>bmed1</i>	LBP	N/A	Items	N/A	<i>cmed1</i>	IND
5	3	3	Aid	<i>bmed2</i>	LBP	N/A	Items	N/A	<i>cmed2</i>	IND

FIGURE 10 (a) A sample of questions in a CRF at 3-month follow-up; (b) a sample of original tabular format health-care resource-use data; and (c) a sample of how the health-care resource-use data populate the repository standard. CC, community clinic; GP, general practice; IND, individual; N/A, not applicable; PHS, public health service; Pri, primary care; Physio, physiotherapist; Priv, private clinic; Px, prescription; RP, recall period.

is 'Primary Care Setting', the unit of measurement is 'Visit' and the payer is 'Public Health Service'. All of these values are derived solely from the information contained in the original question as opposed to the value of the variable. Only the attribute 'Quantity' is directly mapped to the original variable's value.

The health-care resource-use data are stored in the EAV tables by creating relationships between objects. For each time point, one or many resource-use objects can be created. The **HE** class is used to define the time points for collecting only the health-care resource-use data. The actual resource-use data are defined in the **HE-DATA** class and the time point value is used to link an **HE-DATA** object to an **HE** object. The XML schema was modified to allow related classes to be described, which, in turn, gets interpreted by the system to create the relationships in the **Object** table.

Figure 11 shows the **HE-DATA** class being used as a child class, that is, it has the **HE** class as its parent. Creating child classes signifies to the system that a relationship exists between two classes. The **linkedValue** attribute is used to specify a shared value between the parent and child classes. In a relational database, this shared value would be created as a foreign key constraint. In the example shown in Figure 11, an **HE** class has been defined for the 3-month follow-up time point using the attribute **fu**: `<attributeName fu="3"></attributeName>`. A child **HE-DATA** class has been defined and linked to the parent **HE** class by specifying the value "3" for the **linkedValue**: `<childClass name="HE-DATA" linkedValue="3">`. This corresponds with the 3-month follow-up time point specified in the **HE** class.

Child classes in the XML use **groupName** elements to signify the number of objects that need to be created. In a relational database, this would result in adding a new **groupName** element for every table row to be inserted. The value for the **groupName** element has no significance except that it must be unique. In the example shown in Figure 11, six groups have been created for the 3-month resource-use data, namely **3moResource1**, **3moResource2**, **3moResource3**, **3moResource4**, **3moResource5** and **3moResource6**. These groups represent each question in the CRF shown in Figure 10a and the data shown in Figure 10b.

The original tabular data required 13 columns across three rows to store all of the data for the three participants. Instead of creating a new column for every resource, the repository creates a new object. The seven groups are used to create objects for GP visit (**Pri1**), NHS physiotherapist visit (**Pri2**), private physiotherapist visit (**nPriv1**), two instances of prescribed medicine (**pmed1**, **pmed2**) and two instances of aids or medications bought over the counter (**bmed1**, **bmed2**). Although seven groups have been defined in this example, the ETL system will create objects only where data exist. For example, subject #1000 will create only four objects for GP visit (**Pri1**), NHS physiotherapist visit (**Pri2**), private physiotherapist visit (**nPriv1**) and medicine prescribed by GP (**pmed1**).

Once all resources have been identified and a group has been defined, the mapping rules are used to populate the repository's standard resource-use attributes. Within the `<mapping/>` structure, the **groupName** is used to allow the system to locate the correct object to process and the **originalName** is used to store the name of the original variable. The **attributeName** element stores the name of the mapped repository attribute.

The original variable **Pri1** stores the quantity of doctor visits and hence **Pri1** is mapped to the repository attribute **Quantity** for the group **3moResource1**. The other information required to make sense of this value are hard coded to the repository standard within the `<staticValue/>` structure, which is within the `<transform/>` structure. For example, the recall period (**RP**), the type (**Type**), the reason (**Reason**), the location (**Location**), the unit (**Unit**) and the payer (**Payer**) of the resource allocated in **3moResource1** group is hard coded to 3, 1a, 4, 1, 1 and 1, respectively (see Table 11 for list of values and corresponding labels). These values can be hard coded in the XML because they are known to be based on the CRF and do not affect the original data. When the system processes this mapping instruction, subject #1000 would have a health-care resource-use object that shows that there was one GP visit made during the 3-month follow-up time point (see Figure 10b).

```

<class name="HE">
  <mapping>
    <attributeName fu="3"></attributeName>
  </mapping>

  <childClass name="HE-DATA" linkedValue="3">
    <grouping>
      <groupName>3moResource1</groupName>
      <groupName>3moResource2</groupName>
      <groupName>3moResource3</groupName>
      <groupName>3moResource4</groupName>
      <groupName>3moResource5</groupName>
      <groupName>3moResource6</groupName>
    </grouping>

    <mapping>
      <attributeName originalName="Pri1" groupName="3moResource1">Quantity</attributeName>
      <attributeName originalName="Pri2" groupName="3moResource2">Quantity</attributeName>
      <attributeName originalName="nPriv1" groupName="3moResource3">Quantity</attributeName>
      <attributeName originalName="cPriv1" groupName="3moResource3">Cost</attributeName>
      <attributeName originalName="pmed1" groupName="3moResource4">Type</attributeName>
      <attributeName originalName="pmed1" groupName="3moResource4">Text</attributeName>
      <attributeName originalName="nmed1" groupName="3moResource4">Quantity</attributeName>
      <attributeName originalName="bmed1" groupName="3moResource5">Text</attributeName>
      <attributeName originalName="cmed1" groupName="3moResource5">Cost</attributeName>
      <attributeName originalName="bmed2" groupName="3moResource6">Text</attributeName>
      <attributeName originalName="cmed2" groupName="3moResource6">Cost</attributeName>
    </mapping>

    <transform>
      <staticValue>
        <!-- Public healthcare professionals: GP -->
        <newValue attributeName="RP" groupName="3moResource1">3</newValue>
        <newValue attributeName="Type" groupName="3moResource1">1a</newValue>
        <newValue attributeName="Reason" groupName="3moResource1">4</newValue>
        <newValue attributeName="Location" groupName="3moResource1">1</newValue>
        <newValue attributeName="Unit" groupName="3moResource1">1</newValue>
        <newValue attributeName="Payer" groupName="3moResource1">1</newValue>

        <!-- Public healthcare professionals: physiotherapist -->
        <newValue attributeName="RP" groupName="3moResource2">3</newValue>
        <newValue attributeName="Type" groupName="3moResource2">3a</newValue>
        <newValue attributeName="Reason" groupName="3moResource2">2</newValue>
        <newValue attributeName="Location" groupName="3moResource2">4</newValue>
        <newValue attributeName="Unit" groupName="3moResource2">1</newValue>
        <newValue attributeName="Payer" groupName="3moResource2">1</newValue>

        <!-- Private healthcare professionals -->
        <newValue attributeName="RP" groupName="3moResource3">3</newValue>
        <newValue attributeName="Type" groupName="3moResource3">3a</newValue>
        <newValue attributeName="Reason" groupName="3moResource3">2</newValue>
        <newValue attributeName="Location" groupName="3moResource3">3</newValue>
        <newValue attributeName="Unit" groupName="3moResource3">1</newValue>
        <newValue attributeName="Payer" groupName="3moResource3">4</newValue>

        <!-- Medicine prescribed -->
        <newValue attributeName="RP" groupName="3moResource4">3</newValue>
        <newValue attributeName="Reason" groupName="3moResource4">2</newValue>
        <newValue attributeName="Unit" groupName="3moResource4">3</newValue>
        <newValue attributeName="Payer" groupName="3moResource4">1</newValue>

        <!-- Medicine bought-->
        <newValue attributeName="RP" groupName="3moResource5">3</newValue>
        <newValue attributeName="Type" groupName="3moResource5">6</newValue>
        <newValue attributeName="Reason" groupName="3moResource5">2</newValue>
        <newValue attributeName="Unit" groupName="3moResource5">4</newValue>
        <newValue attributeName="Payer" groupName="3moResource5">4</newValue>

        <newValue attributeName="RP" groupName="3moResource6">3</newValue>
        <newValue attributeName="Type" groupName="3moResource6">6</newValue>
        <newValue attributeName="Reason" groupName="3moResource6">2</newValue>
        <newValue attributeName="Unit" groupName="3moResource6">4</newValue>
        <newValue attributeName="Payer" groupName="3moResource6">4</newValue>
      </staticValue>

      <match operator="equal" value="Ibuprofen">
        <newValue attributeName="Type" groupName="3moResource4">4M01</newValue>
      </match>
    </transform>
  </childClass>
</class>

```

FIGURE 11 The XML mapping and transformation instructions for the sample data in Figure 10.

Other `<transform/>` rules can be applied to manipulate the original health-care resource-use data. For example, the original medicines prescribed have to be transformed to the repository standard to the standardised drug coding. *Figure 11* shows a transformation for the `Type` attribute that uses a match rule to check for the value `Ibuprofen`. If matched, the rule has been set to update the attribute's value to `4M01`.

The XML mapping and transformation instructions shown in *Figure 11* were based on only one follow-up time point. For mapping data from more than one follow-up time point, simply create more `HE` objects, and map and transform health-care resource data within the child class `HE-DATA` that is linked to that follow-up time point, for example:

```
<class name="HE">
  <mapping>
    <attributeName fu="3"></attributeName>
    <attributeName fu="6"></attributeName>
    ...
    <attributeName fu="n"></attributeName>
  </mapping>

  <childClass name="HE-DATA" linkedValue="3">
    <grouping>
      <groupName>3moResource1</groupName>
      ...
      <groupName>3moResourceN</groupName>
    </grouping>
    <mapping>
      <attributeName originalName="Pri1" groupName="3moResource1">Quantity</attributeName>
      ...
      <attributeName originalName="cmed2" groupName="3moResource6">Cost</attributeName>
    </mapping>
    <transform>
      ...
    </transform>
  </childClass>

  <childClass name="HE-DATA" linkedValue="6">
    <grouping>
      <groupName>6moResource1</groupName>
      ...
      <groupName>6moResourceN</groupName>
    </grouping>
    <mapping>
      <attributeName originalName="Pri1" groupName="6moResource1">Quantity</attributeName>
      ...
      <attributeName originalName="cmedN" groupName="6moResourceN">Cost</attributeName>
    </mapping>
    <transform>
      ...
    </transform>
  </childClass>

  ...

  <childClass name="HE-DATA" linkedValue="n">
    <grouping>
      <groupName>nmoResource1</groupName>
      ...
      <groupName>nmoResourceN</groupName>
    </grouping>
    <mapping>
      <attributeName originalName="Pri1" groupName="nmoResource1">Quantity</attributeName>
      ...
      <attributeName originalName="cmedN" groupName="nmoResourceN">Cost</attributeName>
    </mapping>
    <transform>
      ...
    </transform>
  </childClass>

</class>
```

Using entity-attribute-value data

Using the EAV with classes and relationships data in its raw state for any kind of analysis work would be extremely difficult because of the fragmented nature of the EAV schema. For analysis purposes, it is therefore necessary to piece together the data to form complete data sets that are comparable to the data sets outputted from relational or tabular data sources. This task is achieved by processing the EAV table to derive a table for each class, a column for each attribute and a row for every object. An excerpt of the SQL statement to join the various data to extract the required data items for class **RMDQ** (whose identifier is 1 in this example) is shown below:

```
SELECT
    eav_objectid,
    prms_TrialName,
    subj_ID,
    subj_OriginalID,
    attr_ShortName,
    eav_Value
FROM
    attribute
    inner join eavobject
        on eav_AttributeID = attr_ID
    inner join
        object on obj_Id = eav_ObjectID
    inner join
        subject on obj_SubjectID = subj_ID
    inner join
        primarysource on prms_ID = subj_PrimarySourceID
WHERE
    obj_ClassID = 1
```

The statement produces a table in a long format, which was subsequently pivoted to produce a row for each object and a column for every attribute. The outcome of this query is a data set that resembles a tabular structure that can easily be processed for further analysis.

Although this solution provides a means for generating a usable tabular format, the scalability is severely limited. The server performance was found to decrease as the volume of data increases, and multiple pivot operations were used for transforming object relationships. Querying the derived data sets directly was also impractical because of the huge numbers of data that can be generated in the server's temporary database, causing the server to be unstable.

An initial solution used to overcome these issues was to disconnect from the actual query by using the in-built functionality of the statistical analysis software to create a copy of the query results. A more permanent solution, which is the current practice, is to periodically create a copy of the query results into actual tables within the database.

Extract, transform and load

The bespoke ETL application was required to read the original source data, automatically apply mapping and transformation rules from an XML document, and load the processed data into the repository. In addition to these basic functions, the ETL application was also required to permit end users to set up new RCTs for import, create new classes and attributes and make changes to existing ones, and to switch between a testing and live environment.

The bespoke ETL application was distributed as a Microsoft Windows® desktop application. It works by first uploading the original data set and the XML mapping and transformation rules. The instructions defined in the XML file are applied to the original data set and the transformed data are loaded into the repository database. The ETL application allows the statistician and health economist to execute these steps from their desktop computers. The ability to switch between a test and live environment gives the users the flexibility and convenience of checking whether or not the instructions that they have delineated in the XML file are correct before loading the data sets into the live database.

Data validation

Data integrity is vital throughout the repository ETL process. To check that the mapping and transformation procedures were carried out correctly, the repository data were routinely checked against the original data sets. To achieve this, at each time point (baseline and all follow-ups), a random sample of data was extracted and manually cross-checked against the source data. Any inconsistencies were flagged and, if required, the XML instructions were amended. This process was repeated until the data were deemed to have been transformed correctly.

Storage

In condition of our data sharing agreements to hold the RCT data sets and to meet local governance and standard operating procedures, the repository database server is held in a secure data centre, with robust disaster recovery policies in place.

The appeal of having this hybrid system architecture is that the structure takes up very little space in the server, and the time needed to query and retrieve data is very little, too. Naturally, the disk space needed to store the data in this repository will grow in proportion in accordance to the number of data points.

Future data sharing

At the end of this programme of work we would like to make the pooled data available for future analyses. We will go back to all of the principal investigators (PIs)/data custodians with a new data sharing agreement to enable us to share their pooled data. Once these agreements have been signed, we will set up a website with details of how to apply for the data. All requests will be:

1. forwarded to the study statistician who will carry out internal checks to ensure that the data being requested can be provided; the response from the study statistician will be supplied with the original request for the independent committee consideration
2. sent via e-mail to an independent committee, who will review the application and make a final decision on data sharing; for the data requested, if a PI/data custodian has:
 - i. agreed to sharing the data but has asked to see a copy of the request, a copy will be sent to them via e-mail for information purposes only
 - ii. not agreed to sharing their data, this data set will be removed from the pooled data before providing the requested data to the applicant.

Chapter 5 Crosswalking between disability questionnaire scores

This chapter presents our methodological development, exploring how to most accurately map multiple participant-reported outcome measures (PROMs), which measure the same domain, to a common scale (crosswalking). This work has now been published in *Spine*.¹⁴⁹ We sought to develop a 'crosswalk' of values from multiple measures of the same domain to a common single outcome score. This would allow us to pool measures more accurately than normalising to a single scale (e.g. 0–100) or expressing values as a proportion of their SD. The first step in this work is to ensure that changes in outcomes from two measures in the same individuals are both correlated and similarly responsive to change. The results from this work would inform us how, and if, we could pool various back pain-related disability outcomes into a single outcome for the main analyses (see *Chapter 6*).

Background

There are six PROMs that have been used in one or more studies within the repository that aim to measure back pain-related disability, namely the Chronic Pain Grade Scale (CPG) disability score (CPG-DS), which is one of the two domains in the CPG that aims to grade chronic pain status,¹⁵⁰ Hannover Functional Ability Questionnaire for measuring back pain-related functional limitations (Funktionsbeeinträchtigung durch Rückenschmerzen) (FFbHR),¹⁵¹ Oswestry Disability Index (ODI),¹⁵² Pain Disability Index (PDI),¹⁵³ Patient-Specific Functional Scale (PSFS)¹⁵⁴ and RMDQ.²⁸ Some trials also included generic health-related quality-of-life instruments, such as the Short Form questionnaire-12 items (SF-12)¹⁵⁵ or the Short Form questionnaire-36 items (SF-36),¹⁵⁶ for which the physical component score (PCS) measures the physical functioning. As mentioned later in *Chapter 6* (see *Outcome variables*), no common instrument was used by the trials that were included in the repository. We sought to assess the agreement of these instruments by determining their correlation and responsiveness at a trial level, in order to decide whether or not data pooling was feasible. After we had completed this work, a National Institutes of Health task force identified developing crosswalking values for 'legacy' measures of back pain outcome as a key priority for back pain research.¹⁵⁷

Data

We used data from 11 trials that had used at least two of the following measurements: CPG, FFbHR, PCS, PSFS, PDI, ODI and RMDQ. For all of these analyses we used the short-term change score, as this is where any treatment effects are likely to be greatest. For the purposes of this report we have defined a short-term follow-up as a measurement taken at between 2 and 3 months post randomisation or entry to the trial. The short-term change score is the difference between the baseline and the short-term follow-up (see *Chapter 6, Follow-up time point*). In each case we have standardised the reporting so that a positive change score is interpreted as an improvement. Where appropriate, we used the standardised response; change score divided by the SD of the change. We used this in preference to the standardised effect size (change score divided by the SD of the measure at baseline), so that all of the standardised scores had a SD of one. This enables visual comparisons to be made between all of the scatterplots.

Outcome conversion

All comparisons between instruments were carried out at an individual trial level. Each pair of outcome measures was fitted with simple linear regression models. Denoting the change scores for the two outcome measures by x and y , the simple linear model was:

$$y = \alpha + \beta X + \epsilon, \quad (1)$$

where the intercept, α , and the coefficient, β , are parameters to be estimated and ϵ is the error term. For the conversion to be meaningful, the standardised change scores have to be correlated and have similar responsiveness; the latter is explained below.¹⁵⁸

Correlation

Correlation was assessed by scatterplots and Pearson's correlation coefficient, with a correlation coefficient considered to be at least moderately high if it was > 0.5 .

Responsiveness

Responsiveness is the ability to detect a change in condition; if a participant's condition improves or worsens over time then this should be reflected by a change in the participant's score. If two 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 two outcome measures was examined by categorising the change scores as negative change (change score of < 0), no change (change score $= 0$) or positive change (change score of > 0), and applying Cohen's kappa (κ) to these categorisations.¹⁵⁹ We considered $\kappa > 0.4$ to indicate sufficiently similar responsiveness.¹⁶⁰ 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). We also planned to examine narrower categorisations in the event that the agreements within these three categories were good ($\kappa > 0.4$). However, as there was no standard on the levels of categorisations, a few would be examined.

For it to be acceptable to pool two measures, they needed to meet two criteria; to be at least moderately correlated (correlation > 0.5) and to have at least moderately similar responsiveness ($\kappa > 0.4$).

Results

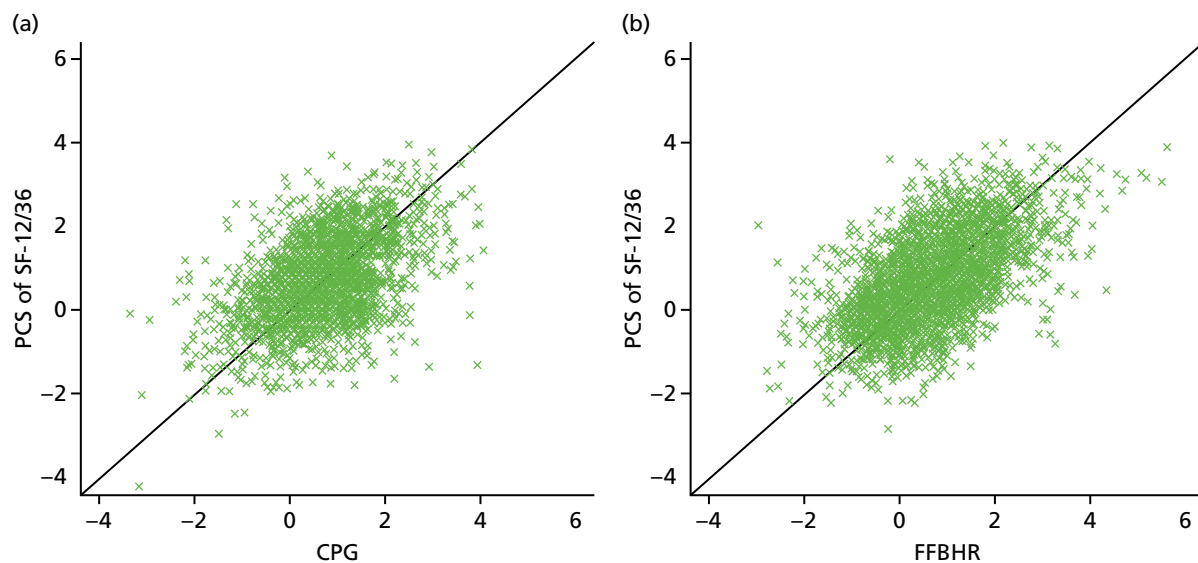
Eleven trials^{31,33,50,76,101,103,104,107,131,132,134} ($n = 6089$) and seven instruments were included in these analyses (Table 12). There was a total of 21 within-trial pairwise comparisons between two outcomes. Figures 12–16 show scatterplots of standardised change scores for each such pair of outcome measures. See Appendix 8 for scatterplots between raw change. It is clear from these plots that the outcomes were positively correlated. Note also that the standardised change scores were widely scattered around the reference line, suggesting that there was a lack of agreement between the outcomes.

TABLE 12 Instruments used and number of participants by trial

Trial	<i>n</i>	Outcome measures		
BeST ³³	426	RMDQ	CPG	PCS ^a
Brinkhaus ¹⁰¹	281	PCS	FFbHR	PDI
Haake ¹³²	1110	CPG	FFbHR	PCS
Hancock ¹³¹	235	RMDQ	PSFS	
HullExPro ⁷⁶	203	RMDQ	PCS	
Macedo ¹³⁴	158	RMDQ	PCS	PSFS
Pengel ¹⁰³	232	RMDQ	PSFS	
UK BEAM ³¹	885	RMDQ	CPG	PCS
VKBIA ¹⁰⁴	227	RMDQ	CPG	
Witt ⁵⁰	2229	PCS	FFbHR	
YACBAC ¹⁰⁷	206	PCS	ODI	

YACBAC, York Acupuncture Back Pain Trial.

a PCS of SF-12 or SF-36.

**FIGURE 12** Scatterplots of standardised change scores for (a) PCS vs. CPG (*n* = 2451); and (b) PCS vs. FFbHR (*n* = 3620) outcome measures. PCS of SF-12/36.

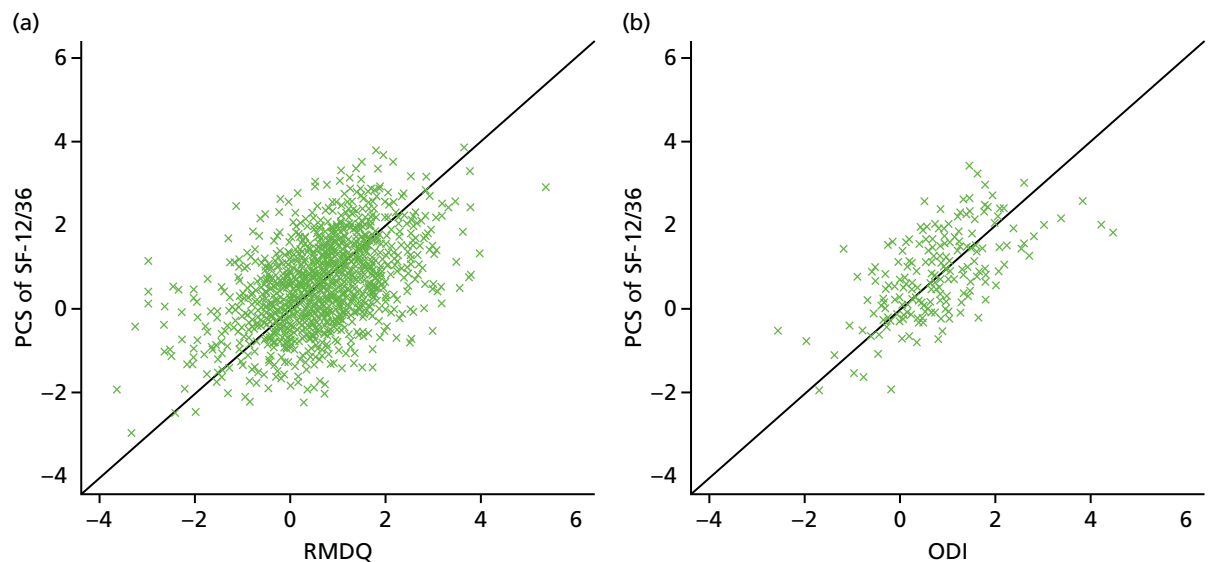


FIGURE 13 Scatterplots of standardised change scores for (a) PCS vs. RMDQ ($n = 1694$); and (b) PCS vs. ODI ($n = 206$) outcome measures. PCS of SF-12/36.

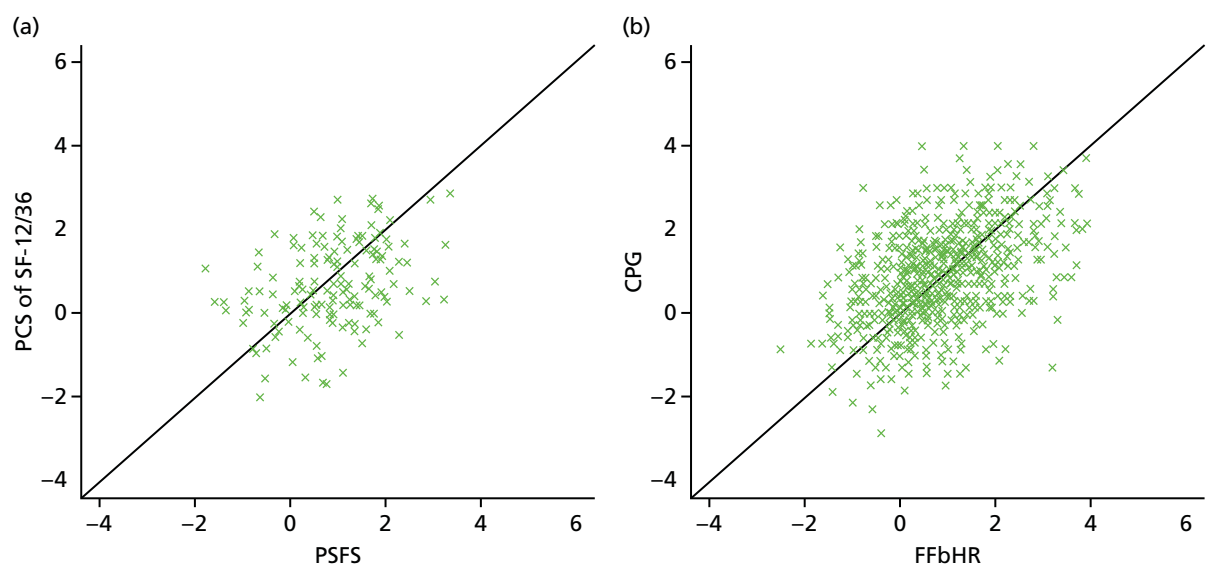


FIGURE 14 Scatterplots of standardised change scores for (a) PCS vs. PSFS ($n = 158$); and (b) CPG vs. FFbHR ($n = 1110$) outcome measures. PCS of SF-12/36.

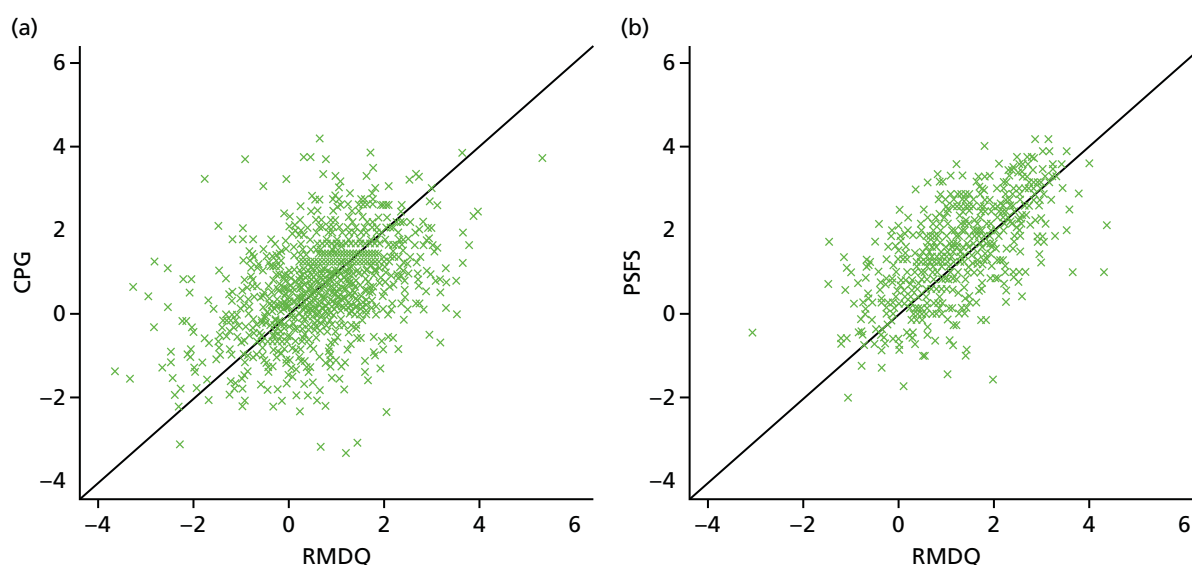


FIGURE 15 Scatterplots of standardised change scores for (a) CPG vs. RMDQ ($n = 1661$); and (b) PSFS vs. RMDQ ($n = 625$) outcome measures.

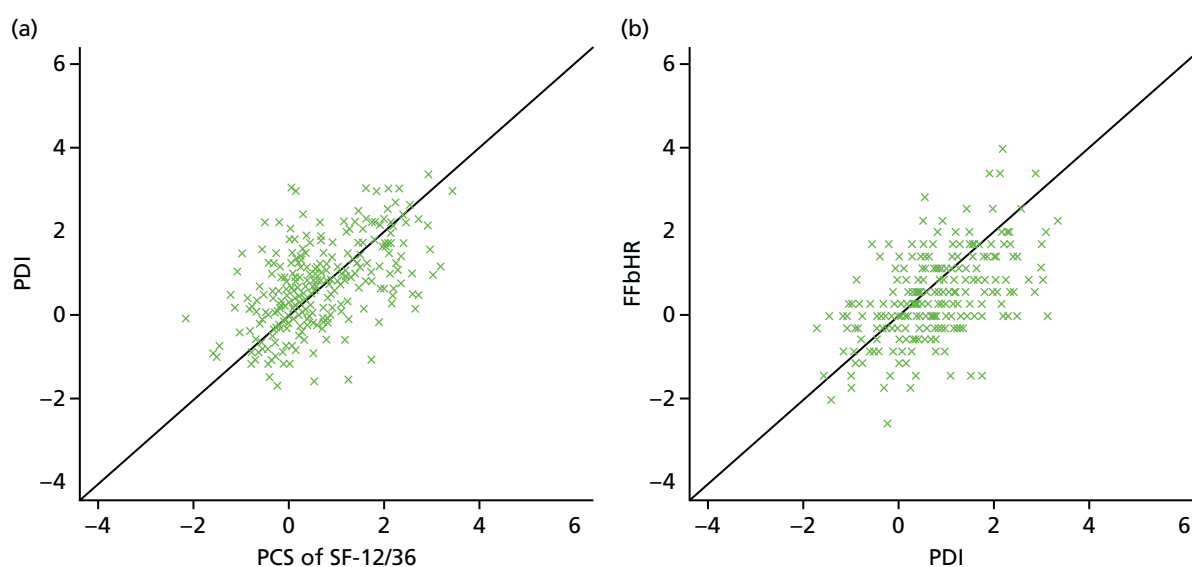


FIGURE 16 Scatterplots of standardised change scores for (a) PDI vs. PCS ($n = 281$), and (b) FFbHR vs. and PDI ($n = 284$) outcome measures. PCS, physical component scale of SF-12/36.

The correlations between outcomes ranged from 0.21 to 0.70; implying that the linear associations between them range from weak to moderately strong (*Table 13*). Three trials^{50,101,132} had both SF-12/36 PCS and FFbHR data, and their correlations were very similar, about 0.58. Another three trials^{33,101,132} had both SF-12/36 PCS and CPG, and the correlations were reasonably similar, ranging from 0.41 to 0.56, and four trials^{31,33,76,134} had both a SF-12/36 PCS and a RMDQ score with range 0.38–0.52, again similar. However, correlations between other outcomes were quite wide ranging: between CPG and RMDQ scores ($m = 3$ trials;^{31,33,104} range 0.21–0.47) and between PSFS and RMDQ scores ($m = 3$;^{103,131,134} range 0.40–0.70).

Cohen's kappa was < 0.4 for all 21 comparisons. Some were similar between trials, namely for PCS and FFbHR (range 0.27–0.30) and for PCS and CPG (range 0.27–0.31). However, the level of agreement was never more than fair.¹⁶⁰ As the Cohen's kappa agreement was not > 0.4 narrower categorisations were not investigated.

There were no pairs of outcomes that satisfied both criteria of at least moderately correlated (correlation > 0.5) and at least moderately similar responsive (Cohen's kappa > 0.4). Therefore, it was not meaningful to convert any outcome to another one.

TABLE 13 Pearson's correlation coefficient and Cohen's kappa agreement for responsiveness of each pairwise comparison of outcome measures by trial

Outcome measure		Trial	Pearson's correlation coefficient	Cohen's kappa
1	2			
CPG	RMDQ	BeST ³³	0.44	0.22
		UK BEAM ³¹	0.47	0.27
		VKBIA ¹⁰⁴	0.21	0.12
CPG	FFbHR	Haake ¹³²	0.48	0.25
PCS ^a	RMDQ	BeST ³³	0.38	0.17
		HullExPro ⁷⁶	0.45	0.29
		Macedo ¹³⁴	0.52	0.27
		UK BEAM ³¹	0.51	0.33
PCS	CPG	BeST ³³	0.41	0.27
		Haake ¹³²	0.49	0.27
		UK BEAM ³¹	0.56	0.31
PCS	FFbHR	Brinkhaus ¹⁰¹	0.59	0.30
		Haake ¹³²	0.58	0.29
		Witt ⁵⁰	0.59	0.27
PCS	PSFS	Macedo ¹³⁴	0.36	0.17
PCS	ODI	YACBAC ¹⁰⁷	0.60	0.28
RMDQ	PSFS	Hancock ¹³¹	0.70	0.38
		Macedo ¹³⁴	0.40	0.26
		Pengel ¹⁰³	0.53	0.18
PDI	FFbHR	Brinkhaus ¹⁰¹	0.55	0.32
PDI	PCS	Brinkhaus ¹⁰¹	0.54	0.31

YACBAC, York Acupuncture Back Pain Trial.

^a PCS of SF-12 or SF-36.

Conclusion

In view of the lack of correlation and responsiveness, it is not recommended to map any physical disability outcome measures to another considered in this investigation.

For each of our subsequent analyses we have pooled data only where the same participant-reported outcomes are available from multiple trials. The one exception is that the SF-12 and SF-36 are explicitly designed to have similar measurement properties when converted into their physical and mental component scores. We have therefore pooled the mental component score (MCS) and PCS from studies using SF-12 or SF-36.

Chapter 6 Preliminary statistical analyses and results

Background

In this chapter we present the results of preliminary statistical analyses performed on the individual participant data, specifically the analysis of covariance (ANCOVA) comparing all treatments with all controls (usual care plus sham) to identify individual potential moderators to take forward into our main analyses. The methodological development work to identify multiple covariates' baseline characteristics that moderate treatment effect is presented in later chapters (see *Chapters 7–10*). We do not, in this preliminary analysis, seek to define subgroups using multiple parameters.

Statistical analysis plan

In accordance with the standard operating procedure in the Warwick Clinical Trials Unit, a detailed statistical analysis plan was written by the study's statistician (SWH) and health economist (JJ). The plan was subsequently reviewed and approved by the study team and members of the repository oversight committee (see *Appendix 9*), whereas the overview of the plan is described in the following sections.

Definitions

Treatment arms

Treatments are broadly classified into intervention, sham control and control. The intervention grouping may be further classified into three broad categories, namely active physical, passive physical and psychological. Exercise and graded activity are considered as *active physical*; acupuncture, manual therapy and individual physiotherapy are considered as *passive physical*; and advice or education, and a cognitive-behavioural approach or cognitive-behavioural therapy, are considered as *psychological* interventions. Sham control may be sham acupuncture, sham electrotherapy, mock TENS or sham advice or education. The control arm is the non-active usual care, namely GP treatment or a waiting list control. Sham acupuncture may be a special case of a sham intervention. If it is the sensation of needling that is the active ingredient of acupuncture then the location of any needling, whether or not skin penetration takes place, or depth of any needling might have little effect on outcomes seen. Thus, sham acupuncture might be considered to be a 'true' intervention and is included in our analyses of passive physical treatments.

Follow-up time point

The follow-up times are classified into short term, mid term and long term. A short-term follow-up is a measurement taken between 2 and 3 months post randomisation or entry to the trial. A mid-term follow-up is a measurement taken at 6 months post randomisation or entry to the trial. A long-term follow-up is a measurement taken at 12 months post randomisation or entry to the trial. Data collected at immediate follow-up (< 2 months post randomisation or entry to the trial) and beyond the long-term follow-up (> 12 months post randomisation or entry to the trial) were also entered into the repository but were not considered for analysis.

Selection of follow-up time points

Some RCTs collected weekly data. For the short-term follow-up, data from the 3-month follow-up were considered for analysis. If data were missing (non-response), data from the nearest week to the 3-month follow-up were used as long as the time point was within the 2- and 3-month follow-up time point.

Outcome variables

Clinical outcomes

The response for each of the outcome variables of interest is presented as *change score* and *standardised change score*. The change score is the change from baseline to the follow-up time point. A positive change score is interpreted as an improvement.

Health-economic outcomes

For the initial economic analysis presented here, the outcome of QALYs was used. Estimated QALY gains from treatment were compared with the mean estimated costs of treatment to assess cost-effectiveness. Individual participant data on resource use or costs were available for some trials but, after allowing for availability of the European Quality of Life-5 Dimensions (EQ-5D) or SF-12/36 scores (required to calculate QALYs) and of a common set of moderator variables, no two studies provided both individual-level cost and QALY data for a common comparison. We were, therefore, unable to generate pooled cost/QALY data.

The heterogeneous nature of the trials posed some challenges for the economic analysis. To pool the data across trials, a consistent health outcome measure over time was required. The QALY is a standardised measure of health outcomes used for economic analysis, which summarises patients' profiles of health-related quality of life ('utility') over time. The QALY score for each patient was estimated using the EQ-5D, which is a generic measure of quality of life, suitable for calculation of QALYs. The EQ-5D index score, calculated using the UK Tariff, measures an individual's health state at a single time point.¹⁶¹ EQ-5D index scores can be integrated over time to estimate QALYs. QALYs were calculated for trial participants over 1 year of follow-up, using the area under the curve (AUC) method. For each participant the AUC was calculated from the EQ-5D index scores that were captured at each follow-up point for that participant, from baseline to 52 weeks (with linear interpolation between observations). Trials with more follow-up points arguably have greater resolution and therefore the QALY estimated will be more precise. However, in all regression analyses the differences between trials were controlled for, so this potential issue was mitigated.

For one trial,¹³² EQ-5D data were not available, but full data on patient responses to the SF-12 instrument were recorded. The SF-12 is a generic measure of health, similar to the EQ-5D, and a number of methods to estimate a utility index score from the SF-12 instrument have been published. To ensure that the index scores provided by the SF-12 are comparable to those obtained for the other trials using the EQ-5D, a mapping approach was applied. This mapped the SF-12 item responses on to the EQ-5D index scores. The specific mapping approach applied was based on the work of Gray *et al.*;¹⁶² in this study, a multinomial logit model was used to estimate the probability that a particular EQ-5D level would be chosen, based on the participants' SF-12 responses. The authors have made available an algorithm applying this method as an add-on programme in Stata 12 (StataCorp LP, College Station, TX, USA). This mapping approach was compared with other published methods by Rowen *et al.*¹⁶³ They found similar levels of performance across the alternative approaches. In our analysis, the mapped SF-12 index scores were integrated over time in the same manner as the EQ-5D scores to estimate an individual-level QALY. Use of SF-6D (Short Form questionnaire-6 Dimensions) to EQ-5D mapping might have introduced additional errors or bias, although the method was well developed and has been subject to validation. The potential for bias should also have been mitigated by the method of analysis, with a mixed model accounting for differences between trials. Furthermore, the outcomes of interest were the treatment subgroup coefficients rather than the magnitude of main effects per se.

One trial¹³² had data only up to 26 weeks. For this trial,¹³² it was assumed that the quality-of-life score measured at 26 weeks persisted up to 52 weeks, which allowed QALYs over 1 year to be estimated in the same way as for the other trials. This assumption might be seen as a limitation but, again, the potential for bias from this source should have been reduced through the inclusion of the trial as a random effect and the focus on treatment-subgroup interactions.

It is important to adjust for any baseline differences in EQ-5D scores when comparing QALY estimates between treatment groups. There are two ways of making this adjustment: by calculating a 'change from baseline' QALY at the individual level or adding the baseline EQ-5D score as a covariate in regression analysis. The latter approach has been used in the analyses presented here, as it is recommended as more efficient.¹⁶⁴

Selection of instrument

Clinical outcomes are classified broadly into physical disability, pain, psychological distress and non-utility quality of life. Nine instruments in the repository have been identified as measurements for physical disability and four instruments for pain (see *Appendix 9*). No single instrument was used by all RCTs to measure physical disability and hence we explored how to map some of these instruments to one single outcome. The mapping methodology is described in *Chapter 5*. We concluded that it was not possible to map to one single outcome. Therefore, analyses were undertaken on common outcomes only.

Most of the RCTs in the repository had asked the participant to rate or mark on a numerical rating scale or a visual analogue scale (VAS) that described either their average or worst pain at the present time or over defined weeks or months. This item was presented either as a single stand-alone instrument or as an item that was part of a collective pain measurement; for example, in the McGill Pain Questionnaire a VAS was presented as a line that anchors with 'no pain' at one end and 'worst possible pain' at the other end.¹⁶⁵ For the analyses of average pain, one of the following instruments from each trial, where available, was chosen (in descending order):

1. individual VAS on average pain today
2. average pain over the past 1 week
3. average pain over the past 2 weeks, average pain over the past 1 month
4. average pain over the past 3 months
5. the individual item of the CPG pain intensity score (CPG-PS) that is equivalent to the VAS if it is available¹⁵⁰
6. the summary score of the CPG-PS otherwise or
7. the bodily pain domain of SF-12/36.^{155,156}

Where a numerical rating scale (range 0–10) was used, it was scaled to an analogue scale so that it gives a range from 0 to 100.

There are two dimensions of psychological distress that are of interest: depression and anxiety. Six and four instruments have been identified to measure depression and anxiety, respectively (see *Appendix 9*). Within each instrument there is usually a classification system that is widely used to classify participants into ordinal categories, for example with a minimal, a moderate or a severe level of depression. Thus, all instruments were mapped into a single ordinal categorical variable. Instruments with no threshold guideline to discriminate level of risk or severity were categorised into tertiles to discriminate the low- and high-risk or low- and high-severity groups from the moderate-risk or moderate-severity group. Other psychosocial measures – catastrophising, coping and fear avoidance – were handled in the same manner. In each case, the reference standard for comparison was the tertile with the least favourable score.

Data sets

Individual participant data without treatment assignment were excluded from the repository. This exclusion criterion applies to individual participants whose data were included in the data set but the treatment allocation was not available in the data set. We were not able to allocate these participants to a treatment group and they were thus excluded.

Clinical analysis

The main analysis, which is to confirm proof of concept, was based on complete case analysis. Missing data due to non-responders or withdrawals were not imputed. Missing items were imputed and the method for imputation is as described in the statistical analysis plan (see *Appendix 9*). When available, individual items were used to obtain the composite score for each measurement, otherwise the composite score provided to the repository was used for all analyses.

For the overall exploration of moderation by single variables, the sham control was grouped with non-active usual care. All direct analyses were based on pairwise comparisons, that is, only two treatment arms were compared each time. For the overall analysis, intervention was compared against control/placebo arm, where intervention was any therapist-delivered intervention given either singly or in combination with another intervention, and the control/placebo arm was either the non-active usual care control or sham treatment. Other pairwise comparisons considered were active physical against non-active usual care control, passive physical against non-active usual care control, psychological against non-active usual care control, and sham against non-active usual care control. In all cases for the pairwise comparisons we separated sham and usual care controls, as this reflects more accurately the clinical choice than adding of an intervention on to a sham control intervention.

Direct analyses were performed if the individual participant data were from at least two trials, that is, no direct analysis was performed if the individual participant data were from one single trial.

Health-economic analysis

The health-economic analysis focused on the QALY score as the outcome measure. QALYs were calculated for individuals, using the estimated EQ-5D index scores or a mapped SF-12 outcome at multiple follow-up points. This means that missing data can be more of a problem than for outcomes measured at a single time point. If data are missing at any follow-up point, the QALY cannot be estimated and the entire observation is lost. An observation was also lost if data on the moderator at baseline were missing. All analyses were based on complete cases only; therefore, caution must be taken in interpretation of the results, as the missing data may be a source of bias.

In order to simplify the analysis it was split into four overarching comparisons; all interventions collectively against non-active usual care, active physical interventions against non-active usual care, passive physical interventions against non-active usual care and active physical against passive physical. For each analysis, the treatment arms for the included trials were pooled appropriately by the type of treatment and used collectively as the intervention group for each of the respective analyses. Seven trials in total were included in the analysis. The first three analyses described limited the sample to a maximum of six trials, which included a non-active usual care as the control arm and reporting EQ-5D outcomes or a mapped SF-12 outcome. The comparison between active physical and passive physical allowed the inclusion of one additional trial. Data for comparisons against a sham treatment arm were excluded from this analysis, as these are not plausible choices for a health-economic analysis.

Methods

Descriptive summary

The baseline data were summarised by treatment arm (non-active usual care, active physical, passive physical, psychological, combination or sham control). The continuous data were summarised as mean and SD, and the categorical data were summarised as the number of participants and percentage.

One-step meta-analysis

In a one-step meta-analysis, individual participant data from all studies were modelled simultaneously in a single model adjusting for the study effect.¹⁶⁶ It can be viewed analogously as an analysis of a multicentre study, for which, instead of multicentres in a study, we have multitrials in a study.

The one-step meta-analysis was performed to explore the efficacy between treatment arms. A mixed-effects model was used as analysis, for which the intercept and the interaction between treatment arm and trial were modelled as random effects, and treatment arm as the fixed effect.

Moderator identification

Systematic review

We identified potential moderators from the literature via a systematic review. Details of this review and the outcomes are presented in *Chapter 2*.

Analysis of covariance

Analysis of covariance was performed to identify any covariate that moderates outcomes. Similarly, the one-step meta-analysis approach was used, that is, all of the available individual participant data were pooled into a single mixed-effects model for which the intercept and the interaction between treatment and trial were modelled as random effects. The treatment arm (intervention against control), covariate and the interaction between treatment and covariate were modelled as fixed effects. For analysis with QALYs as the outcome measure, the baseline EQ-5D score was also included as a fixed effects model in the mixed-effects model described above.

As stated in the statistical analysis plan, covariates were declared weakly statistically significant at the two-sided 20% level and statistically significant at the two-sided 5% level. This ensured that covariates that approach the conventional statistical significance at 5% level would not be missed for the final clinical and health-economic prediction rule analyses. All moderators identified from the systematic review and ANCOVA were considered for the clinical and health-economic prediction rule analyses. The prediction rule analyses were to determine which participant characteristics at baseline were optimal to different treatments and associated with the end points of interest, namely disability or pain, or cost-effective treatments for LBP. The methodology of identifying a combination of characteristics is presented in detail in *Chapters 7–10*.

As seen in the results from the one-step meta-analysis, the estimated efficacy between intervention and control/placebo arm for most of the outcomes at mid-term and long term were not statistically significant. Therefore, the ANCOVA was not performed for the mid- and long-term outcomes. In addition, the short-term outcomes were those in which the maximum clinical effects were observed between intervention and control/placebo. This is where the largest differential subgroup effects are likely to be seen. In the absence of substantial short-term effect moderation there is little point in exploring mid- and long-term effect moderation.

The list of moderators assessed for each of the short-term clinical outcomes and QALYs were presented. As not all trials have the same moderators, the sample size varied depending on which moderator was being assessed and for which outcome.

Results

Descriptive

Table 14 shows the response rates for each of the outcomes of interest per treatment groups in different time points. Most trials collected data 3 months post randomisation or entry to the trial, and this is recorded as 13 weeks, whereas one RCT had specifically mentioned in its protocol to collect data at 12 weeks and thus this was recorded as per protocol.

Most of the RCTs collected short- and mid-term outcomes and some collected more immediate outcomes (typically measured within 6 weeks post randomisation or entry to the trial) (see *Table 14*). Two RCTs collected longer-term effects (outcomes measured at or after 12 months post randomisation or entry to the trial). Each of the RCTs was designed with a unique protocol and this was apparent from the choice of different instruments used to measure the physical disability, pain and psychological distress outcomes, and at different time points.

TABLE 14 Number of trials (*m*) and participants (*n*) for each outcome by follow-up time points and treatment arms

Outcomes	Follow-up (weeks)	Active physical (<i>m</i> = 7; <i>n</i> = 914)	Passive physical (<i>m</i> = 12; <i>n</i> = 3270)	Psychological (<i>m</i> = 7; <i>n</i> = 1120)	Combination (<i>m</i> = 3; <i>n</i> = 451)	Sham (<i>m</i> = 6; <i>n</i> = 688)	Control (<i>m</i> = 10; <i>n</i> = 2885)	All (<i>m</i> = 19; <i>n</i> = 9326)
Physical disability								
CPG-DS ^a	0	<i>m</i> = 1; <i>n</i> = 284	<i>m</i> = 2; <i>n</i> = 721	<i>m</i> = 2; <i>n</i> = 572	<i>m</i> = 1; <i>n</i> = 312	<i>m</i> = 1; <i>n</i> = 387	<i>m</i> = 4; <i>n</i> = 1052	<i>m</i> = 5; <i>n</i> = 3328
	4	<i>m</i> = 1; <i>n</i> = 228	<i>m</i> = 1; <i>n</i> = 315	–	<i>m</i> = 1; <i>n</i> = 280	–	<i>m</i> = 1; <i>n</i> = 262	<i>m</i> = 4; <i>n</i> = 1085
	8	–	–	<i>m</i> = 1; <i>n</i> = 109	–	–	<i>m</i> = 1; <i>n</i> = 120	<i>m</i> = 2; <i>n</i> = 229
	13	<i>m</i> = 1; <i>n</i> = 214	<i>m</i> = 2; <i>n</i> = 653	<i>m</i> = 1; <i>n</i> = 345	<i>m</i> = 1; <i>n</i> = 252	<i>m</i> = 1; <i>n</i> = 376	<i>m</i> = 3; <i>n</i> = 797	<i>m</i> = 5; <i>n</i> = 2637
	26	–	<i>m</i> = 1; <i>n</i> = 377	<i>m</i> = 2; <i>n</i> = 491	–	<i>m</i> = 1; <i>n</i> = 376	<i>m</i> = 3; <i>n</i> = 656	<i>m</i> = 2; <i>n</i> = 1900
	52	<i>m</i> = 1; <i>n</i> = 212	<i>m</i> = 1; <i>n</i> = 267	<i>m</i> = 2; <i>n</i> = 473	<i>m</i> = 1; <i>n</i> = 254	–	<i>m</i> = 3; <i>n</i> = 530	<i>m</i> = 5; <i>n</i> = 1736
	104	–	–	<i>m</i> = 1; <i>n</i> = 94	–	–	<i>m</i> = 1; <i>n</i> = 92	<i>m</i> = 2; <i>n</i> = 186
FFbHR	0	–	<i>m</i> = 3; <i>n</i> = 1927	–	–	<i>m</i> = 2; <i>n</i> = 460	<i>m</i> = 3; <i>n</i> = 1789	<i>m</i> = 3; <i>n</i> = 4176
	6	–	<i>m</i> = 1; <i>n</i> = 370	–	–	<i>m</i> = 1; <i>n</i> = 375	<i>m</i> = 1; <i>n</i> = 362	<i>m</i> = 1; <i>n</i> = 1107
	8	–	<i>m</i> = 1; <i>n</i> = 140	–	–	<i>m</i> = 1; <i>n</i> = 70	<i>m</i> = 1; <i>n</i> = 74	<i>m</i> = 1; <i>n</i> = 284
	13	–	<i>m</i> = 2; <i>n</i> = 1723	–	–	<i>m</i> = 1; <i>n</i> = 376	<i>m</i> = 2; <i>n</i> = 1605	<i>m</i> = 2; <i>n</i> = 3704
	26	–	<i>m</i> = 3; <i>n</i> = 1825	–	–	<i>m</i> = 2; <i>n</i> = 446	<i>m</i> = 3; <i>n</i> = 1620	<i>m</i> = 3; <i>n</i> = 3891
	52	–	<i>m</i> = 1; <i>n</i> = 137	–	–	<i>m</i> = 1; <i>n</i> = 68	<i>m</i> = 1; <i>n</i> = 70	<i>m</i> = 1; <i>n</i> = 275
ODI	0	–	<i>m</i> = 1; <i>n</i> = 159	–	–	–	<i>m</i> = 1; <i>n</i> = 80	<i>m</i> = 1; <i>n</i> = 239
	13	–	<i>m</i> = 1; <i>n</i> = 146	–	–	–	<i>m</i> = 1; <i>n</i> = 71	<i>m</i> = 1; <i>n</i> = 217
	52	–	<i>m</i> = 1; <i>n</i> = 136	–	–	–	<i>m</i> = 1; <i>n</i> = 57	<i>m</i> = 1; <i>n</i> = 193
	104	–	<i>m</i> = 1; <i>n</i> = 114	–	–	–	<i>m</i> = 1; <i>n</i> = 50	<i>m</i> = 1; <i>n</i> = 164
PDI	0	–	<i>m</i> = 1; <i>n</i> = 146	–	–	<i>m</i> = 1; <i>n</i> = 73	<i>m</i> = 1; <i>n</i> = 79	<i>m</i> = 1; <i>n</i> = 298
	8	–	<i>m</i> = 1; <i>n</i> = 140	–	–	<i>m</i> = 1; <i>n</i> = 70	<i>m</i> = 1; <i>n</i> = 74	<i>m</i> = 1; <i>n</i> = 284
	26	–	<i>m</i> = 1; <i>n</i> = 138	–	–	<i>m</i> = 1; <i>n</i> = 70	<i>m</i> = 1; <i>n</i> = 73	<i>m</i> = 1; <i>n</i> = 281
	52	–	<i>m</i> = 1; <i>n</i> = 137	–	–	<i>m</i> = 1; <i>n</i> = 66	<i>m</i> = 1; <i>n</i> = 69	<i>m</i> = 1; <i>n</i> = 272

TABLE 14 Number of trials (*m*) and participants (*n*) for each outcome by follow-up time points and treatment arms (*continued*)

Outcomes	Follow-up (weeks)	Active physical (<i>m</i> = 7; <i>n</i> = 914)	Passive physical (<i>m</i> = 12; <i>n</i> = 3270)	Psychological (<i>m</i> = 7; <i>n</i> = 1120)	Combination (<i>m</i> = 3; <i>n</i> = 451)	Sham (<i>m</i> = 6; <i>n</i> = 688)	Control (<i>m</i> = 10; <i>n</i> = 2885)	All (<i>m</i> = 19; <i>n</i> = 9326)
PSFS	0	<i>m</i> = 2; <i>n</i> = 150	<i>m</i> = 1; <i>n</i> = 119	<i>m</i> = 2; <i>n</i> = 148	<i>m</i> = 1; <i>n</i> = 62	<i>m</i> = 2; <i>n</i> = 188	–	<i>m</i> = 3; <i>n</i> = 667
	1	–	<i>m</i> = 1; <i>n</i> = 119	–	–	<i>m</i> = 1; <i>n</i> = 118	–	<i>m</i> = 2; <i>n</i> = 237
	2	–	<i>m</i> = 1; <i>n</i> = 119	–	–	<i>m</i> = 1; <i>n</i> = 119	–	<i>m</i> = 1; <i>n</i> = 238
	4	–	<i>m</i> = 1; <i>n</i> = 118	–	–	<i>m</i> = 1; <i>n</i> = 117	–	<i>m</i> = 1; <i>n</i> = 235
	6	<i>m</i> = 1; <i>n</i> = 58	–	<i>m</i> = 1; <i>n</i> = 54	<i>m</i> = 1; <i>n</i> = 57	<i>m</i> = 1; <i>n</i> = 59	–	<i>m</i> = 1; <i>n</i> = 228
	8	<i>m</i> = 1; <i>n</i> = 82	–	<i>m</i> = 1; <i>n</i> = 76	–	–	–	<i>m</i> = 1; <i>n</i> = 158
	12	<i>m</i> = 1; <i>n</i> = 57	–	<i>m</i> = 1; <i>n</i> = 56	<i>m</i> = 1; <i>n</i> = 58	<i>m</i> = 1; <i>n</i> = 61	–	<i>m</i> = 1; <i>n</i> = 232
	13	–	<i>m</i> = 1; <i>n</i> = 118	–	–	<i>m</i> = 1; <i>n</i> = 117	–	<i>m</i> = 1; <i>n</i> = 235
	26	<i>m</i> = 1; <i>n</i> = 81	–	<i>m</i> = 1; <i>n</i> = 74	–	–	–	<i>m</i> = 1; <i>n</i> = 155
	52	<i>m</i> = 2; <i>n</i> = 136	–	<i>m</i> = 2; <i>n</i> = 132	<i>m</i> = 1; <i>n</i> = 56	<i>m</i> = 1; <i>n</i> = 56	–	<i>m</i> = 2; <i>n</i> = 380
RMDQ	0	<i>m</i> = 7; <i>n</i> = 907	<i>m</i> = 7; <i>n</i> = 1087	<i>m</i> = 7; <i>n</i> = 1120	<i>m</i> = 3; <i>n</i> = 446	<i>m</i> = 3; <i>n</i> = 212	<i>m</i> = 6; <i>n</i> = 938	<i>m</i> = 14; <i>n</i> = 4710
	1	–	<i>m</i> = 1; <i>n</i> = 119	–	–	<i>m</i> = 1; <i>n</i> = 118	–	<i>m</i> = 1; <i>n</i> = 237
	2	–	<i>m</i> = 2; <i>n</i> = 119	–	–	<i>m</i> = 1; <i>n</i> = 118	–	<i>m</i> = 1; <i>n</i> = 237
	4	<i>m</i> = 1; <i>n</i> = 234	<i>m</i> = 2; <i>n</i> = 436	–	<i>m</i> = 1; <i>n</i> = 283	<i>m</i> = 1; <i>n</i> = 117	<i>m</i> = 1; <i>n</i> = 264	<i>m</i> = 2; <i>n</i> = 1334
	6	<i>m</i> = 2; <i>n</i> = 144	<i>m</i> = 1; <i>n</i> = 23	<i>m</i> = 1; <i>n</i> = 55	<i>m</i> = 1; <i>n</i> = 58	<i>m</i> = 2; <i>n</i> = 81	<i>m</i> = 1; <i>n</i> = 94	<i>m</i> = 3; <i>n</i> = 455
	8	<i>m</i> = 1; <i>n</i> = 82	–	<i>m</i> = 2; <i>n</i> = 186	–	–	<i>m</i> = 1; <i>n</i> = 120	<i>m</i> = 2; <i>n</i> = 388
	10	<i>m</i> = 1; <i>n</i> = 107	–	–	<i>m</i> = 1; <i>n</i> = 55	–	<i>m</i> = 1; <i>n</i> = 50	<i>m</i> = 1; <i>n</i> = 212
	12	<i>m</i> = 1; <i>n</i> = 58	–	<i>m</i> = 1; <i>n</i> = 58	<i>m</i> = 1; <i>n</i> = 59	<i>m</i> = 1; <i>n</i> = 61	–	<i>m</i> = 1; <i>n</i> = 236
	13	<i>m</i> = 3; <i>n</i> = 433	<i>m</i> = 7; <i>n</i> = 963	<i>m</i> = 4; <i>n</i> = 670	<i>m</i> = 1; <i>n</i> = 255	<i>m</i> = 2; <i>n</i> = 135	<i>m</i> = 3; <i>n</i> = 537	<i>m</i> = 9; <i>n</i> = 2993
	26	<i>m</i> = 4; <i>n</i> = 371	<i>m</i> = 2; <i>n</i> = 262	<i>m</i> = 5; <i>n</i> = 706	<i>m</i> = 1; <i>n</i> = 53	–	<i>m</i> = 5; <i>n</i> = 474	<i>m</i> = 8; <i>n</i> = 1866
	52	<i>m</i> = 7; <i>n</i> = 722	<i>m</i> = 5; <i>n</i> = 771	<i>m</i> = 7; <i>n</i> = 903	<i>m</i> = 3; <i>n</i> = 365	<i>m</i> = 1; <i>n</i> = 56	<i>m</i> = 6; <i>n</i> = 690	<i>m</i> = 12; <i>n</i> = 3507
	104	<i>m</i> = 1; <i>n</i> = 83	<i>m</i> = 1; <i>n</i> = 95	<i>m</i> = 1; <i>n</i> = 94	–	–	<i>m</i> = 1; <i>n</i> = 92	<i>m</i> = 2; <i>n</i> = 364

continued

TABLE 14 Number of trials (*m*) and participants (*n*) for each outcome by follow-up time points and treatment arms (*continued*)

Outcomes	Follow-up (weeks)	Active physical (<i>m</i> = 7; <i>n</i> = 914)	Passive physical (<i>m</i> = 12; <i>n</i> = 3270)	Psychological (<i>m</i> = 7; <i>n</i> = 1120)	Combination (<i>m</i> = 3; <i>n</i> = 451)	Sham (<i>m</i> = 6; <i>n</i> = 688)	Control (<i>m</i> = 10; <i>n</i> = 2885)	All (<i>m</i> = 19; <i>n</i> = 9326)
Troublesomeness	0	<i>m</i> = 2; <i>n</i> = 344	<i>m</i> = 3; <i>n</i> = 556	<i>m</i> = 1; <i>n</i> = 426	<i>m</i> = 1; <i>n</i> = 312	–	<i>m</i> = 3; <i>n</i> = 604	<i>m</i> = 4; <i>n</i> = 2242
	4	<i>m</i> = 1; <i>n</i> = 225	<i>m</i> = 1; <i>n</i> = 313	–	<i>m</i> = 1; <i>n</i> = 279	–	<i>m</i> = 1; <i>n</i> = 262	<i>m</i> = 1; <i>n</i> = 1079
	13	<i>m</i> = 2; <i>n</i> = 280	<i>m</i> = 3; <i>n</i> = 494	–	<i>m</i> = 1; <i>n</i> = 253	–	<i>m</i> = 2; <i>n</i> = 318	<i>m</i> = 3; <i>n</i> = 1345
	52	<i>m</i> = 2; <i>n</i> = 302	<i>m</i> = 3; <i>n</i> = 493	–	<i>m</i> = 1; <i>n</i> = 252	–	<i>m</i> = 2; <i>n</i> = 297	<i>m</i> = 8; <i>n</i> = 1344
	104	–	<i>m</i> = 1; <i>n</i> = 113	–	–	–	<i>m</i> = 1; <i>n</i> = 50	<i>m</i> = 3; <i>n</i> = 162
Pain								
CPG-PS ^b	0	<i>m</i> = 1; <i>n</i> = 283	<i>m</i> = 2; <i>n</i> = 721	<i>m</i> = 2; <i>n</i> = 582	<i>m</i> = 1; <i>n</i> = 312	<i>m</i> = 1; <i>n</i> = 387	<i>m</i> = 4; <i>n</i> = 1054	<i>m</i> = 4; <i>n</i> = 3339
	4	<i>m</i> = 1; <i>n</i> = 228	<i>m</i> = 1; <i>n</i> = 316	–	<i>m</i> = 1; <i>n</i> = 281	–	<i>m</i> = 1; <i>n</i> = 261	<i>m</i> = 1; <i>n</i> = 1086
	6	–	<i>m</i> = 1; <i>n</i> = 370	–	–	<i>m</i> = 1; <i>n</i> = 375	<i>m</i> = 1; <i>n</i> = 362	<i>m</i> = 1; <i>n</i> = 1107
	8	–	–	<i>m</i> = 1; <i>n</i> = 110	–	–	<i>m</i> = 1; <i>n</i> = 120	<i>m</i> = 1; <i>n</i> = 230
	13	<i>m</i> = 1; <i>n</i> = 214	<i>m</i> = 2; <i>n</i> = 653	<i>m</i> = 1; <i>n</i> = 354	<i>m</i> = 1; <i>n</i> = 252	<i>m</i> = 1; <i>n</i> = 376	<i>m</i> = 3; <i>n</i> = 799	<i>m</i> = 3; <i>n</i> = 2648
	26	–	<i>m</i> = 1; <i>n</i> = 377	<i>m</i> = 2; <i>n</i> = 497	–	<i>m</i> = 1; <i>n</i> = 376	<i>m</i> = 3; <i>n</i> = 661	<i>m</i> = 3; <i>n</i> = 1911
	52	<i>m</i> = 1; <i>n</i> = 211	<i>m</i> = 1; <i>n</i> = 269	<i>m</i> = 2; <i>n</i> = 491	<i>m</i> = 1; <i>n</i> = 253	–	<i>m</i> = 4; <i>n</i> = 536	<i>m</i> = 3; <i>n</i> = 1760
	104	–	–	<i>m</i> = 1; <i>n</i> = 94	–	–	<i>m</i> = 1; <i>n</i> = 92	<i>m</i> = 1; <i>n</i> = 186
VAS								
Average pain today	0	<i>m</i> = 2; <i>n</i> = 253	<i>m</i> = 3; <i>n</i> = 461	<i>m</i> = 1; <i>n</i> = 196	<i>m</i> = 1; <i>n</i> = 61	<i>m</i> = 1; <i>n</i> = 120	<i>m</i> = 1; <i>n</i> = 51	<i>m</i> = 3; <i>n</i> = 1142
	1	–	<i>m</i> = 1; <i>n</i> = 119	–	–	<i>m</i> = 1; <i>n</i> = 119	–	<i>m</i> = 1; <i>n</i> = 238
	2	–	<i>m</i> = 1; <i>n</i> = 119	–	–	<i>m</i> = 1; <i>n</i> = 119	–	<i>m</i> = 1; <i>n</i> = 238
	3	–	<i>m</i> = 1; <i>n</i> = 118	–	–	<i>m</i> = 1; <i>n</i> = 118	–	<i>m</i> = 1; <i>n</i> = 236
	4	<i>m</i> = 1; <i>n</i> = 83	<i>m</i> = 1; <i>n</i> = 118	<i>m</i> = 1; <i>n</i> = 80	–	<i>m</i> = 1; <i>n</i> = 118	–	<i>m</i> = 2; <i>n</i> = 399
	6	–	<i>m</i> = 1; <i>n</i> = 36	–	–	<i>m</i> = 1; <i>n</i> = 38	–	<i>m</i> = 1; <i>n</i> = 74
	8	<i>m</i> = 1; <i>n</i> = 81	<i>m</i> = 1; <i>n</i> = 24	<i>m</i> = 1; <i>n</i> = 79	–	<i>m</i> = 1; <i>n</i> = 23	–	<i>m</i> = 2; <i>n</i> = 207
	10	<i>m</i> = 1; <i>n</i> = 107	<i>m</i> = 1; <i>n</i> = 16	–	<i>m</i> = 1; <i>n</i> = 55	<i>m</i> = 1; <i>n</i> = 18	<i>m</i> = 1; <i>n</i> = 49	<i>m</i> = 2; <i>n</i> = 245

TABLE 14 Number of trials (*m*) and participants (*n*) for each outcome by follow-up time points and treatment arms (*continued*)

Outcomes	Follow-up (weeks)	Active physical (<i>m</i> = 7; <i>n</i> = 914)	Passive physical (<i>m</i> = 12; <i>n</i> = 3270)	Psychological (<i>m</i> = 7; <i>n</i> = 1120)	Combination (<i>m</i> = 3; <i>n</i> = 451)	Sham (<i>m</i> = 6; <i>n</i> = 688)	Control (<i>m</i> = 10; <i>n</i> = 2885)	All (<i>m</i> = 19; <i>n</i> = 9326)
	11	–	<i>m</i> = 1; <i>n</i> = 15	–	–	<i>m</i> = 1; <i>n</i> = 17	–	<i>m</i> = 1; <i>n</i> = 32
	12	–	<i>m</i> = 1; <i>n</i> = 15	–	–	<i>m</i> = 1; <i>n</i> = 17	–	<i>m</i> = 1; <i>n</i> = 32
	13	<i>m</i> = 1; <i>n</i> = 81	<i>m</i> = 1; <i>n</i> = 153	<i>m</i> = 2; <i>n</i> = 231	–	–	–	<i>m</i> = 1; <i>n</i> = 465
	17	<i>m</i> = 1; <i>n</i> = 79	–	<i>m</i> = 1; <i>n</i> = 75	–	–	–	<i>m</i> = 1; <i>n</i> = 154
	21	<i>m</i> = 1; <i>n</i> = 81	–	<i>m</i> = 1; <i>n</i> = 76	–	–	–	<i>m</i> = 1; <i>n</i> = 157
	26	<i>m</i> = 2; <i>n</i> = 186	–	<i>m</i> = 1; <i>n</i> = 75	<i>m</i> = 1; <i>n</i> = 53	–	–	<i>m</i> = 2; <i>n</i> = 314
	30	<i>m</i> = 1; <i>n</i> = 79	–	<i>m</i> = 1; <i>n</i> = 72	–	–	–	<i>m</i> = 1; <i>n</i> = 151
	34	<i>m</i> = 1; <i>n</i> = 81	–	<i>m</i> = 1; <i>n</i> = 73	–	–	–	<i>m</i> = 1; <i>n</i> = 154
	39	<i>m</i> = 1; <i>n</i> = 80	–	<i>m</i> = 1; <i>n</i> = 74	–	–	–	<i>m</i> = 1; <i>n</i> = 154
	43	<i>m</i> = 1; <i>n</i> = 78	–	<i>m</i> = 1; <i>n</i> = 74	–	–	–	<i>m</i> = 1; <i>n</i> = 152
	47	<i>m</i> = 1; <i>n</i> = 76	–	<i>m</i> = 1; <i>n</i> = 71	–	–	–	<i>m</i> = 1; <i>n</i> = 147
	52	<i>m</i> = 2; <i>n</i> = 183	<i>m</i> = 1; <i>n</i> = 164	<i>m</i> = 2; <i>n</i> = 238	<i>m</i> = 1; <i>n</i> = 53	–	–	<i>m</i> = 6; <i>n</i> = 638
Average pain over past 1 week	0	<i>m</i> = 2; <i>n</i> = 150	<i>m</i> = 2; <i>n</i> = 235	<i>m</i> = 3; <i>n</i> = 349	<i>m</i> = 1; <i>n</i> = 63	<i>m</i> = 2; <i>n</i> = 84	–	<i>m</i> = 4; <i>n</i> = 881
	1	–	<i>m</i> = 1; <i>n</i> = 235	–	–	<i>m</i> = 1; <i>n</i> = 119	–	<i>m</i> = 1; <i>n</i> = 238
	2	–	<i>m</i> = 1; <i>n</i> = 235	–	–	<i>m</i> = 1; <i>n</i> = 119	–	<i>m</i> = 1; <i>n</i> = 238
	3	–	<i>m</i> = 1; <i>n</i> = 235	–	–	<i>m</i> = 1; <i>n</i> = 118	–	<i>m</i> = 1; <i>n</i> = 237
	4	<i>m</i> = 1; <i>n</i> = 82	<i>m</i> = 2; <i>n</i> = 152	<i>m</i> = 1; <i>n</i> = 80	–	<i>m</i> = 2; <i>n</i> = 134	–	<i>m</i> = 3; <i>n</i> = 448
	6	<i>m</i> = 1; <i>n</i> = 59	<i>m</i> = 1; <i>n</i> = 49	<i>m</i> = 1; <i>n</i> = 55	<i>m</i> = 1; <i>n</i> = 58	<i>m</i> = 2; <i>n</i> = 97	–	<i>m</i> = 2; <i>n</i> = 306
	8	<i>m</i> = 1; <i>n</i> = 81	<i>m</i> = 1; <i>n</i> = 24	<i>m</i> = 1; <i>n</i> = 79	–	<i>m</i> = 1; <i>n</i> = 24	–	<i>m</i> = 2; <i>n</i> = 208
	10	–	<i>m</i> = 1; <i>n</i> = 16	–	–	<i>m</i> = 1; <i>n</i> = 19	–	<i>m</i> = 1; <i>n</i> = 35
	11	–	<i>m</i> = 1; <i>n</i> = 11	–	–	<i>m</i> = 1; <i>n</i> = 17	–	<i>m</i> = 1; <i>n</i> = 33

continued

TABLE 14 Number of trials (*m*) and participants (*n*) for each outcome by follow-up time points and treatment arms (*continued*)

Outcomes	Follow-up (weeks)	Active physical (<i>m</i> = 7; <i>n</i> = 914)	Passive physical (<i>m</i> = 12; <i>n</i> = 3270)	Psychological (<i>m</i> = 7; <i>n</i> = 1120)	Combination (<i>m</i> = 3; <i>n</i> = 451)	Sham (<i>m</i> = 6; <i>n</i> = 688)	Control (<i>m</i> = 10; <i>n</i> = 2885)	All (<i>m</i> = 19; <i>n</i> = 9326)
	12	<i>m</i> = 1; <i>n</i> = 58	<i>m</i> = 1; <i>n</i> = 15	<i>m</i> = 1; <i>n</i> = 58	<i>m</i> = 1; <i>n</i> = 59	<i>m</i> = 2; <i>n</i> = 78	–	<i>m</i> = 2; <i>n</i> = 268
	13	<i>m</i> = 1; <i>n</i> = 81	<i>m</i> = 2; <i>n</i> = 180	<i>m</i> = 2; <i>n</i> = 231	–	<i>m</i> = 1; <i>n</i> = 9	–	<i>m</i> = 3; <i>n</i> = 501
	17	<i>m</i> = 1; <i>n</i> = 79	–	<i>m</i> = 1; <i>n</i> = 75	–	–	–	<i>m</i> = 1; <i>n</i> = 154
	21	<i>m</i> = 1; <i>n</i> = 81	–	<i>m</i> = 1; <i>n</i> = 76	–	–	–	<i>m</i> = 1; <i>n</i> = 157
	26	<i>m</i> = 1; <i>n</i> = 81	<i>m</i> = 1; <i>n</i> = 21	<i>m</i> = 1; <i>n</i> = 75	–	<i>m</i> = 1; <i>n</i> = 6	–	<i>m</i> = 2; <i>n</i> = 183
	30	<i>m</i> = 1; <i>n</i> = 79	–	<i>m</i> = 1; <i>n</i> = 72	–	–	–	<i>m</i> = 1; <i>n</i> = 151
	34	<i>m</i> = 1; <i>n</i> = 81	–	<i>m</i> = 1; <i>n</i> = 73	–	–	–	<i>m</i> = 1; <i>n</i> = 154
	39	<i>m</i> = 1; <i>n</i> = 80	–	<i>m</i> = 1; <i>n</i> = 74	–	–	–	<i>m</i> = 1; <i>n</i> = 154
	43	<i>m</i> = 1; <i>n</i> = 78	–	<i>m</i> = 1; <i>n</i> = 74	–	–	–	<i>m</i> = 1; <i>n</i> = 152
	47	<i>m</i> = 1; <i>n</i> = 77	–	<i>m</i> = 1; <i>n</i> = 71	–	–	–	<i>m</i> = 1; <i>n</i> = 148
	52	<i>m</i> = 2; <i>n</i> = 140	<i>m</i> = 1; <i>n</i> = 163	<i>m</i> = 3; <i>n</i> = 297	<i>m</i> = 1; <i>n</i> = 57	<i>m</i> = 1; <i>n</i> = 56	–	<i>m</i> = 3; <i>n</i> = 713
Average pain over past 1 month	0	–	<i>m</i> = 1; <i>n</i> = 24	–	–	<i>m</i> = 1; <i>n</i> = 24	–	<i>m</i> = 1; <i>n</i> = 48
	6	–	<i>m</i> = 1; <i>n</i> = 23	–	–	<i>m</i> = 1; <i>n</i> = 22	–	<i>m</i> = 1; <i>n</i> = 45
	13	–	<i>m</i> = 1; <i>n</i> = 22	–	–	<i>m</i> = 1; <i>n</i> = 18	–	<i>m</i> = 1; <i>n</i> = 40
Worst pain today	0	<i>m</i> = 1; <i>n</i> = 111	–	–	<i>m</i> = 1; <i>n</i> = 61	–	<i>m</i> = 1; <i>n</i> = 51	<i>m</i> = 1; <i>n</i> = 223
	10	<i>m</i> = 1; <i>n</i> = 107	–	–	<i>m</i> = 1; <i>n</i> = 53	–	<i>m</i> = 1; <i>n</i> = 49	<i>m</i> = 1; <i>n</i> = 209
	26	<i>m</i> = 1; <i>n</i> = 103	–	–	<i>m</i> = 1; <i>n</i> = 53	–	–	<i>m</i> = 1; <i>n</i> = 156
	52	<i>m</i> = 1; <i>n</i> = 103	–	–	<i>m</i> = 1; <i>n</i> = 52	–	–	<i>m</i> = 1; <i>n</i> = 155
Worst pain over past 1 month	0	–	<i>m</i> = 2; <i>n</i> = 24	–	–	<i>m</i> = 1; <i>n</i> = 24	–	<i>m</i> = 2; <i>n</i> = 48
	6	–	<i>m</i> = 1; <i>n</i> = 23	–	–	<i>m</i> = 1; <i>n</i> = 22	–	<i>m</i> = 2; <i>n</i> = 45
	13	–	<i>m</i> = 1; <i>n</i> = 22	–	–	<i>m</i> = 1; <i>n</i> = 18	–	<i>m</i> = 2; <i>n</i> = 40

TABLE 14 Number of trials (*m*) and participants (*n*) for each outcome by follow-up time points and treatment arms (*continued*)

Outcomes	Follow-up (weeks)	Active physical (<i>m</i> = 7; <i>n</i> = 914)	Passive physical (<i>m</i> = 12; <i>n</i> = 3270)	Psychological (<i>m</i> = 7; <i>n</i> = 1120)	Combination (<i>m</i> = 3; <i>n</i> = 451)	Sham (<i>m</i> = 6; <i>n</i> = 688)	Control (<i>m</i> = 10; <i>n</i> = 2885)	All (<i>m</i> = 19; <i>n</i> = 9326)
Quality of life								
SF-12/SF-36 PCS ^a	0	<i>m</i> = 4; <i>n</i> = 617	<i>m</i> = 7; <i>n</i> = 2544	<i>m</i> = 2; <i>n</i> = 507	<i>m</i> = 1; <i>n</i> = 305	<i>m</i> = 2; <i>n</i> = 460	<i>m</i> = 6; <i>n</i> = 2262	<i>m</i> = 9; <i>n</i> = 6695
	4	<i>m</i> = 1; <i>n</i> = 214	<i>m</i> = 1; <i>n</i> = 300	–	<i>m</i> = 1; <i>n</i> = 264	–	<i>m</i> = 1; <i>n</i> = 249	<i>m</i> = 1; <i>n</i> = 1027
	8	<i>m</i> = 1; <i>n</i> = 82	<i>m</i> = 1; <i>n</i> = 139	<i>m</i> = 1; <i>n</i> = 76	–	<i>m</i> = 1; <i>n</i> = 69	<i>m</i> = 1; <i>n</i> = 73	<i>m</i> = 2; <i>n</i> = 439
	13	<i>m</i> = 3; <i>n</i> = 415	<i>m</i> = 6; <i>n</i> = 2276	<i>m</i> = 1; <i>n</i> = 332	<i>m</i> = 1; <i>n</i> = 243	<i>m</i> = 1; <i>n</i> = 376	<i>m</i> = 5; <i>n</i> = 2006	<i>m</i> = 7; <i>n</i> = 5648
	26	<i>m</i> = 2; <i>n</i> = 185	<i>m</i> = 4; <i>n</i> = 1850	<i>m</i> = 2; <i>n</i> = 436	–	<i>m</i> = 2; <i>n</i> = 444	<i>m</i> = 4; <i>n</i> = 1711	<i>m</i> = 6; <i>n</i> = 4626
	52	<i>m</i> = 4; <i>n</i> = 469	<i>m</i> = 5; <i>n</i> = 719	<i>m</i> = 2; <i>n</i> = 449	<i>m</i> = 1; <i>n</i> = 235	<i>m</i> = 1; <i>n</i> = 68	<i>m</i> = 4; <i>n</i> = 545	<i>m</i> = 7; <i>n</i> = 2485
	104	<i>m</i> = 1; <i>n</i> = 83	<i>m</i> = 2; <i>n</i> = 206	–	–	–	<i>m</i> = 1; <i>n</i> = 49	<i>m</i> = 2; <i>n</i> = 338
SF-12/SF-36 MCS ^b	0	<i>m</i> = 4; <i>n</i> = 617	<i>m</i> = 7; <i>n</i> = 2544	<i>m</i> = 2; <i>n</i> = 507	<i>m</i> = 1; <i>n</i> = 305	<i>m</i> = 2; <i>n</i> = 460	<i>m</i> = 6; <i>n</i> = 2262	<i>m</i> = 9; <i>n</i> = 6695
	4	<i>m</i> = 1; <i>n</i> = 214	<i>m</i> = 1; <i>n</i> = 300	–	<i>m</i> = 1; <i>n</i> = 264	–	<i>m</i> = 1; <i>n</i> = 249	<i>m</i> = 1; <i>n</i> = 1027
	8	<i>m</i> = 1; <i>n</i> = 82	<i>m</i> = 1; <i>n</i> = 139	<i>m</i> = 1; <i>n</i> = 76	–	<i>m</i> = 1; <i>n</i> = 69	<i>m</i> = 1; <i>n</i> = 73	<i>m</i> = 2; <i>n</i> = 439
	13	<i>m</i> = 3; <i>n</i> = 415	<i>m</i> = 6; <i>n</i> = 2276	<i>m</i> = 1; <i>n</i> = 332	<i>m</i> = 1; <i>n</i> = 243	<i>m</i> = 1; <i>n</i> = 376	<i>m</i> = 5; <i>n</i> = 2006	<i>m</i> = 7; <i>n</i> = 5648
	26	<i>m</i> = 2; <i>n</i> = 185	<i>m</i> = 4; <i>n</i> = 1850	<i>m</i> = 2; <i>n</i> = 436	–	<i>m</i> = 2; <i>n</i> = 444	<i>m</i> = 4; <i>n</i> = 1711	<i>m</i> = 6; <i>n</i> = 4626
	52	<i>m</i> = 4; <i>n</i> = 469	<i>m</i> = 5; <i>n</i> = 719	<i>m</i> = 2; <i>n</i> = 449	<i>m</i> = 1; <i>n</i> = 235	<i>m</i> = 1; <i>n</i> = 68	<i>m</i> = 4; <i>n</i> = 545	<i>m</i> = 7; <i>n</i> = 2485
	104	<i>m</i> = 1; <i>n</i> = 83	<i>m</i> = 2; <i>n</i> = 206	–	–	–	<i>m</i> = 1; <i>n</i> = 49	<i>m</i> = 2; <i>n</i> = 338
Health utility								
EQ-5D-3L	0	<i>m</i> = 1; <i>n</i> = 85	–	–	–	–	<i>m</i> = 1; <i>n</i> = 94	<i>m</i> = 1; <i>n</i> = 179
	6	<i>m</i> = 1; <i>n</i> = 85	–	–	–	–	<i>m</i> = 1; <i>n</i> = 94	<i>m</i> = 1; <i>n</i> = 179
	26	<i>m</i> = 1; <i>n</i> = 77	–	–	–	–	<i>m</i> = 1; <i>n</i> = 86	<i>m</i> = 1; <i>n</i> = 163
	52	<i>m</i> = 1; <i>n</i> = 82	–	–	–	–	<i>m</i> = 1; <i>n</i> = 88	<i>m</i> = 1; <i>n</i> = 170

EQ-5D-3L, EQ-5D three-level version.

a PCS of SF-12/SF-36.

b MCS of SF-12/SF-36.

There were 9328 participants in the trials included in the repository. *Table 15* shows the demographics and clinical characteristics at baseline by treatment arms. All of the trials were able to provide information on sex and age. Of the 9326 participants (missing data from two participants), 5316 (57%) were females. The proportion of males and females was similar across all treatment arms. The average age of the participants in the repository was 49 years (SD 14 years). The average age of participants from trials that had active physical therapies (APTs) was slightly lower [44 years ($n = 914$; SD 12 years)] compared with the average age from trials that had passive and psychological treatments [49 years ($n = 3270$; SD 14 years) and 50 years ($n = 1118$; SD 14 years)], respectively. This difference is mainly due to the inclusion criteria of the trials.

TABLE 15 Demographics and clinical characteristics at baseline by treatment arms

Characteristics	Active physical ($m = 7$; $n = 914$)	Passive physical ($m = 12$; $n = 3270$)	Psychological ($m = 7$; $n = 1120$)	Combination ($m = 3$; $n = 451$)	Sham ($m = 6$; $n = 688$)	Control ($m = 10$; $n = 2885$)	All ($m = 19$; $n = 9328$)
Demographics							
<i>Age, years</i>							
Number of trials, m	7	12	7	3	6	10	19
n	914	3270	1118	451	688	2885	9326
Mean	43.67	49.39	50.08	43.77	48.54	50.51	48.92
SD	11.74	14.13	14.22	12.51	15.22	13.37	13.88
<i>Sex</i>							
Number of trials, m	7	12	7	3	6	10	19
Female (%)	497 (54.4)	1907 (58.3)	655 (58.5)	237 (52.6)	412 (59.9)	1641 (56.9)	5349 (57.4)
Male (%)	417 (45.6)	1363 (41.7)	464 (41.5)	214 (47.5)	276 (40.1)	1243 (43.1)	3977 (42.6)
<i>Ethnicity</i>							
Number of trials, m	1	1	4	–	–	4	5
White (%)	65 (75.6)	159 (100.0)	667 (87.8)	–	–	478 (89.4)	1369 (88.9)
Mixed (%)	–	–	4 (0.5)	–	–	3 (0.6)	7 (0.5)
Black (%)	–	–	26 (3.4)	–	–	21 (3.9)	47 (3.1)
Asian (Indian, Pakistani, Bangladeshi, others) (%)	7 (8.1)	–	37 (4.9)	–	–	17 (3.2)	61 (4.0)
Chinese (%)	1 (1.2)	–	1 (0.1)	–	–	1 (0.2)	3 (0.2)
Others (%)	13 (15.1)	–	25 (3.3)	–	–	15 (2.8)	53 (3.4)
<i>Smoking status</i>							
Number of trials, m	5	3	3	1	1	1	6
No (%)	333 (66.7)	211 (52.4)	167 (76.3)	52 (82.5)	54 (79.4)	69 (70.4)	886 (65.6)
Yes (%)	167 (33.3)	192 (47.6)	52 (23.7)	11 (17.5)	14 (20.6)	29 (29.6)	465 (34.4)

TABLE 15 Demographics and clinical characteristics at baseline by treatment arms (*continued*)

Characteristics	Active physical (<i>m</i> = 7; <i>n</i> = 914)	Passive physical (<i>m</i> = 12; <i>n</i> = 3270)	Psychological (<i>m</i> = 7; <i>n</i> = 1120)	Combination (<i>m</i> = 3; <i>n</i> = 451)	Sham (<i>m</i> = 6; <i>n</i> = 688)	Control (<i>m</i> = 10; <i>n</i> = 2885)	All (<i>m</i> = 19; <i>n</i> = 9328)
<i>Employment status</i>							
Number of trials, <i>m</i>	5	6	5	1	1	6	11
Full-time employment (%)	307 (51.3)	424 (51.7)	360 (42.2)	165 (64.7)	4 (25.0)	485 (54.3)	1745 (50.8)
Part-time employment (%)	120 (20.0)	130 (15.9)	132 (15.5)	60 (23.5)	–	190 (21.3)	632 (18.4)
No employment (%)	172 (28.7)	266 (32.4)	362 (42.4)	30 (11.8)	12 (75.0)	218 (24.4)	1060 (30.8)
<i>BMI</i>							
Number of trials, <i>m</i>	2	4	2	–	2	2	5
<i>n</i>	222	811	156	–	453	462	2,104
Mean	27.03	26.60	26.52	–	26.45	26.42	26.57
SD	5.31	4.60	5.22	–	4.73	4.48	4.73
<i>Physical disability</i> <i>CPG-DS (0–100; 100 = worst)^a</i>							
Number of trials, <i>m</i>	1	2	2	1	1	5	4
<i>n</i>	284	721	572	312	387	1052	3328
Mean	47.44	51.82	49.38	44.76	55.36	49.87	50.16
SD	22.66	20.9	23.77	21.86	18.92	22.14	21.99
<i>FFbHR (0–100; 100 = best)</i>							
Number of trials, <i>m</i>	–	3	–	–	2	3	3
<i>n</i>	–	1927	–	–	460	1789	4176
Mean	–	58.33	–	–	48.01	59.38	57.64
SD	–	20.63	–	–	16.14	20.69	20.5
<i>ODI (0–100; 100 = worst)</i>							
Number of trials, <i>m</i>	–	1	–	–	–	1	1
<i>n</i>	–	159	–	–	–	80	239
Mean	–	33.72	–	–	–	31.36	32.93
SD	–	15.40	–	–	–	14.24	15.03
<i>PDI (0–70; 70 = worst)</i>							
Number of trials, <i>m</i>	–	1	–	–	1	1	1
<i>n</i>	–	146	–	–	73	79	298
Mean	–	28.92	–	–	31.53	30.95	30.10
SD	–	11.12	–	–	11.14	13.27	11.75

continued

TABLE 15 Demographics and clinical characteristics at baseline by treatment arms (*continued*)

Characteristics	Active physical (<i>m</i> = 7; <i>n</i> = 914)	Passive physical (<i>m</i> = 12; <i>n</i> = 3270)	Psychological (<i>m</i> = 7; <i>n</i> = 1120)	Combination (<i>m</i> = 3; <i>n</i> = 451)	Sham (<i>m</i> = 6; <i>n</i> = 688)	Control (<i>m</i> = 10; <i>n</i> = 2885)	All (<i>m</i> = 19; <i>n</i> = 9328)
<i>PSFS (0–10; 10 = best)</i>							
Number of trials, <i>m</i>	2	1	2	1	2	–	3
<i>n</i>	150	119	148	62	188	–	667
Mean	3.57	3.78	3.76	3.83	3.97	–	3.79
SD	1.79	1.60	1.67	1.94	1.84	–	1.76
<i>RMDQ (0–24; 24 = worst)</i>							
Number of trials, <i>m</i>	7	7	7	3	3	6	14
<i>n</i>	907	1087	1,120	446	212	938	4710
Mean	10.07	10.89	9.85	9.59	11.09	8.57	9.91
SD	5.08	5.03	5.33	4.33	5.95	4.69	5.09
Troublesomeness							
Number of trials, <i>m</i>	2	3	1	1	–	3	4
Not at all troublesome (%)	3	4	–	–	–	4	11
Slightly troublesome (%)	41	62	26	29	–	51	209
Moderately troublesome (%)	146	213	211	154	–	284	1008
Very troublesome (%)	115	205	151	107	–	211	789
Extremely troublesome (%)	39	72	38	22	–	54	225
Pain							
<i>CPG-PS (0–100; 100 = worst)^a</i>							
Number of trials, <i>m</i>	1	2	3	1	1	5	5
<i>n</i>	283	721	582	312	387	1054	3,339
Mean	60.82	64.93	58.93	59.91	67.60	62.65	62.66
SD	17.62	16.79	18.53	17.91	13.16	17.41	17.31
<i>Average pain (0–100; 100 = worst)^b</i>							
Number of trials, <i>m</i>	4	6	6	3	5	6	12
<i>n</i>	472	922	969	380	493	1118	4354
Mean	52.42	59.79	48.20	50.63	65.54	52.53	54.40
SD	22.49	20.96	24.74	21.50	15.20	24.64	23.18
Quality of life							
<i>SF-12/SF-36 PCS^a (0–100; 100 = best)</i>							
Number of trials, <i>m</i>	4	7	2	1	2	6	9
<i>n</i>	617	2544	507	305	460	2262	6695
Mean	37.14	36.03	37.15	38.14	32.87	36.30	36.19
SD	7.42	8.05	9.06	7.46	7.09	8.74	8.29

TABLE 15 Demographics and clinical characteristics at baseline by treatment arms (*continued*)

Characteristics	Active physical (<i>m</i> = 7; <i>n</i> = 914)	Passive physical (<i>m</i> = 12; <i>n</i> = 3270)	Psychological (<i>m</i> = 7; <i>n</i> = 1120)	Combination (<i>m</i> = 3; <i>n</i> = 451)	Sham (<i>m</i> = 6; <i>n</i> = 688)	Control (<i>m</i> = 10; <i>n</i> = 2885)	All (<i>m</i> = 19; <i>n</i> = 9328)
SF-12/SF-36 MCS^b (0–100; 100 = best)							
Number of trials, <i>m</i>	4	7	2	1	2	6	9
<i>n</i>	617	2544	507	305	460	2262	6695
Mean	43.94	44.89	44.38	44.84	46.61	45.89	45.22
SD	11.66	12.23	11.28	10.84	11.42	11.90	11.90
Health utility							
EQ-5D-3L (–0.11 to 1; 1 = best)							
Number of trials, <i>m</i>	4	4	2	2	–	5	7
<i>n</i>	593	740	652	371	–	724	3080
Mean	0.57	0.61	0.6	0.58	–	0.59	0.59
SD	0.27	0.27	0.29	0.25	–	0.26	0.27
Depression (DE)							
DASS–DE (0–42; 42 = worst)							
Number of trials, <i>m</i>	1	–	1	1	1	–	1
<i>n</i>	65	–	62	63	68	–	258
Mean	7.11	–	7.55	7.08	7.06	–	7.19
SD	7.84	–	7.67	8.79	7.61	–	7.94
DRAM							
Number of trials, <i>m</i>	2	1	–	1	–	2	2
Type N (%)	135 (36.49)	122 (36.75)	–	116 (37.54)	–	184 (44.88)	557 (39.20)
Type R (%)	147 (39.73)	147 (44.28)	–	120 (38.83)	–	158 (38.54)	572 (40.25)
Type DD (%)	55 (14.86)	41 (12.35)	–	46 (14.89)	–	49 (11.95)	191 (13.44)
Type DS (%)	33 (8.92)	22 (6.63)	–	27 (8.74)	–	19 (4.63)	101 (7.11)
HADS–DE (0–21; 21 = worst)							
Number of trials, <i>m</i>	–	–	1	–	–	1	1
<i>n</i>	–	–	464	–	–	231	695
Mean	–	–	6.04	–	–	5.54	5.87
SD	–	–	3.81	–	–	3.6	3.75
MZDI (0–69; 69 = worst)							
Number of trials, <i>m</i>	2	2	1	1	–	2	3
<i>n</i>	411	485	148	309	–	411	1724
Mean	19.77	21.44	22.41	21.24	–	19.77	21.06
SD	10.75	10.55	9.37	10.93	–	10.75	10.70

continued

TABLE 15 Demographics and clinical characteristics at baseline by treatment arms (*continued*)

Characteristics	Active physical (<i>m</i> = 7; <i>n</i> = 914)	Passive physical (<i>m</i> = 12; <i>n</i> = 3270)	Psychological (<i>m</i> = 7; <i>n</i> = 1120)	Combination (<i>m</i> = 3; <i>n</i> = 451)	Sham (<i>m</i> = 6; <i>n</i> = 688)	Control (<i>m</i> = 10; <i>n</i> = 2885)	All (<i>m</i> = 19; <i>n</i> = 9328)
Anxiety (AN)							
<i>DASS-AN (0–42; 42 = worst)</i>							
Number of trials, <i>m</i>	1	–	1	1	1	–	1
<i>n</i>	65	–	62	63	68	–	258
Mean	6.22	–	5.23	4.76	5.35	–	5.40
SD	7.57	–	7.44	6.68	6.92	–	7.14
<i>HADS-AN (0–21; 21 = worst)</i>							
Number of trials, <i>m</i>	–	–	1	–	–	1	1
<i>n</i>	–	–	458	–	–	230	688
Mean	–	–	8.22	–	–	7.49	7.98
SD	–	–	4.3	–	–	4.43	4.35
Fear avoidance							
<i>ALBPSQ-FA (0–30; 30 = worst)</i>							
Number of trials, <i>m</i>	2	–	2	1	1	–	2
<i>n</i>	121	–	117	36	33	–	307
Mean	18.14	–	18.58	17.14	18.42	–	18.22
SD	6.91	–	6.16	5.97	5.90	–	6.40
<i>FABQ-PC (0–24; 24 = worst)</i>							
Number of trials, <i>m</i>	2	3	1	1	2	4	5
<i>n</i>	366	840	443	311	506	1016	3482
Mean	14.70	16.65	13.59	14.96	17.79	15.85	15.84
SD	5.27	5.24	6.34	5.30	4.87	5.65	5.61
<i>TSK (16–68; 68 = worst)</i>							
Number of trials, <i>m</i>	2	1	4	2	1	3	5
<i>n</i>	176	177	472	124	68	285	1302
Mean	39.08	44.05	41.64	39.33	38.07	39.71	40.79
SD	7.44	7.09	8.14	7.51	8.16	8.58	8.12
Catastrophising (CAT)							
<i>CSQ-CAT (0–36; 36 = worst)</i>							
Number of trials, <i>m</i>	1	1	2	–	–	–	2
<i>n</i>	86	193	282	–	–	–	561
Mean	10.84	7.83	9.62	–	–	–	9.19
SD	7.61	6.65	7.22	–	–	–	7.16

TABLE 15 Demographics and clinical characteristics at baseline by treatment arms (*continued*)

Characteristics	Active physical (<i>m</i> = 7; <i>n</i> = 914)	Passive physical (<i>m</i> = 12; <i>n</i> = 3270)	Psychological (<i>m</i> = 7; <i>n</i> = 1120)	Combination (<i>m</i> = 3; <i>n</i> = 451)	Sham (<i>m</i> = 6; <i>n</i> = 688)	Control (<i>m</i> = 10; <i>n</i> = 2885)	All (<i>m</i> = 19; <i>n</i> = 9328)
PRSS-CAT (0–45; 45 = worst)							
Number of trials, <i>m</i>	1	1	1	1	2	–	2
<i>n</i>	65	119	62	63	188	–	497
Mean	17.92	16.43	17.9	17.29	17.23	–	17.22
SD	8.61	8.12	10.55	9.05	8.53	–	8.77
Coping (CSS)							
CSQ-CSS (0–36; 36 = best)							
Number of trials, <i>m</i>	–	1	1	–	–	–	1
<i>n</i>	–	198	196	–	–	–	394
Mean	–	25.13	25.33	–	–	–	25.23
SD	–	6.23	6.64	–	–	–	6.43
PRSS-CSS (0–45; 45 = best)							
Number of trials, <i>m</i>	1	2	1	1	2	–	2
<i>n</i>	65	119	62	63	188	–	497
Mean	30.18	31.26	30.06	30.37	31.97	–	31.13
SD	7.34	6.95	8.36	6.81	6.85	–	7.15
PSEQ (0–60; 60 = best)							
Number of trials, <i>m</i>	3	1	3	1	1	1	4
<i>n</i>	268	117	601	63	67	223	1,339
Mean	40.49	36.85	40.12	44.38	43.70	41.15	40.46
SD	12.93	10.94	13.17	12.77	13.38	12.54	12.90
Somatic perception							
MSPQ (0–39; 39 = worst)							
Number of trials, <i>m</i>	2	2	1	1	–	2	3
<i>n</i>	372	526	195	310	–	411	1814
Mean	6.78	6.43	5.58	7.07	–	6.14	6.45
SD	5.52	5.38	4.29	5.43	–	5.34	5.32
Sensory index (SE)							
McGill-SE (0–33; 33 = worst)							
Number of trials, <i>m</i>	–	1	1	–	–	–	1
<i>n</i>	–	185	170	–	–	–	355
Mean	–	14.21	14.26	–	–	–	14.24
SD	–	6.10	6.36	–	–	–	6.22

continued

TABLE 15 Demographics and clinical characteristics at baseline by treatment arms (*continued*)

Characteristics	Active physical (<i>m</i> = 7; <i>n</i> = 914)	Passive physical (<i>m</i> = 12; <i>n</i> = 3270)	Psychological (<i>m</i> = 7; <i>n</i> = 1120)	Combination (<i>m</i> = 3; <i>n</i> = 451)	Sham (<i>m</i> = 6; <i>n</i> = 688)	Control (<i>m</i> = 10; <i>n</i> = 2885)	All (<i>m</i> = 19; <i>n</i> = 9328)
<i>SES–SE (10–40; 40 = worst)</i>							
Number of trials, <i>m</i>	–	1	–	–	1	1	1
<i>n</i>	–	146	–	–	73	79	298
Mean	–	49.7	–	–	49.11	49.77	49.57
SD	–	9.05	–	–	8.39	11.06	9.45
Affective index (AF)							
<i>McGill–AF (0–12; 12 = worst)</i>							
Number of trials, <i>m</i>	–	1	1	–	–	–	1
<i>n</i>	–	192	187	–	–	–	379
Mean	–	4.21	4.25	–	–	–	4.23
SD	–	3.31	3.36	–	–	–	3.33
<i>SES–AF (14–56; 56 = worst)</i>							
Number of trials, <i>m</i>	–	1	–	–	1	1	1
<i>n</i>	–	146	–	–	73	79	298
Mean	–	50.19	–	–	50.88	50.01	50.31
SD	–	8.38	–	–	8.17	9.34	8.57
AF, Affective Index; ALBPSQ, Acute Low Back Pain Screening Questionnaire; BMI, body mass index; CAT, Catastrophising; CSQ, Coping Strategy Questionnaire; CSS, Coping; DASS, Depression Anxiety Stress Scales; DD, distressed–depressive; DRAM, Distress and Risk Assessment Method; DS, distressed–somatic; FABQ, Fear-Avoidance Beliefs Questionnaire; HADS, Hospital Anxiety and Depression Scale; McGill, McGill Pain Questionnaire; MSPQ, Modified Somatic Perception Questionnaire; MZDI, Modified Zung Depression Index; N, normal; PSEQ, Pain Self-Efficacy Questionnaire; PRSS, Pain-Related Self-Statement; R, at risk; SES, Pain Experience Scale (Schmerzempfindungsskala); TSK, Tampa Scale for Kinesiophobia. a PCS of SF-12/SF-36. b MCS of SF-12/SF-36.							

Most of the participants with data in the repository had similar physical disability or functional limitation at baseline. One trial⁸⁸ (*n* = 239) used the ODI as its outcome measure and the average baseline score was 33 (SD 15), which was somewhere between no disability and moderate disability. Three trials^{50,101,132} (*n* = 4176) used the FFbHR and the average baseline score was 58 (SD 21), which was slightly above moderate functional limitation. Fourteen trials^{31,33,65,70,76,102–106,131,133,134,136} (*n* = 4710) used the RMDQ as their outcome measure and the average baseline score was 10 (SD 5), which was slightly below moderate disability.

Nine trials^{31,33,50,76,101,102,107,132,134} (*n* = 6695) collected quality-of-life information with either the SF-12 or SF-36 instrument. The mean PCS at baseline was 36 (SD 8) and the mean MCS at baseline was 45 (SD 12). The mean values were similar across treatment arms.

Only a minority of the RCTs provided information on psychological distress at baseline and were insufficient to provide any qualitative comparison across treatment arms.

One-step meta-analysis

Box plots of change of outcome measures from baseline to short-, mid- and long-term follow-up by treatment arms show that participants in all groups are behaving as expected, with all groups improving over time (data not shown). This observation was examined further in the one-step meta-analysis (adjusting for study effects) and the results are shown in *Figures 17–19* and *Table 16*. There was a statistically significant difference between control and intervention for all outcomes at the short-term follow-up.

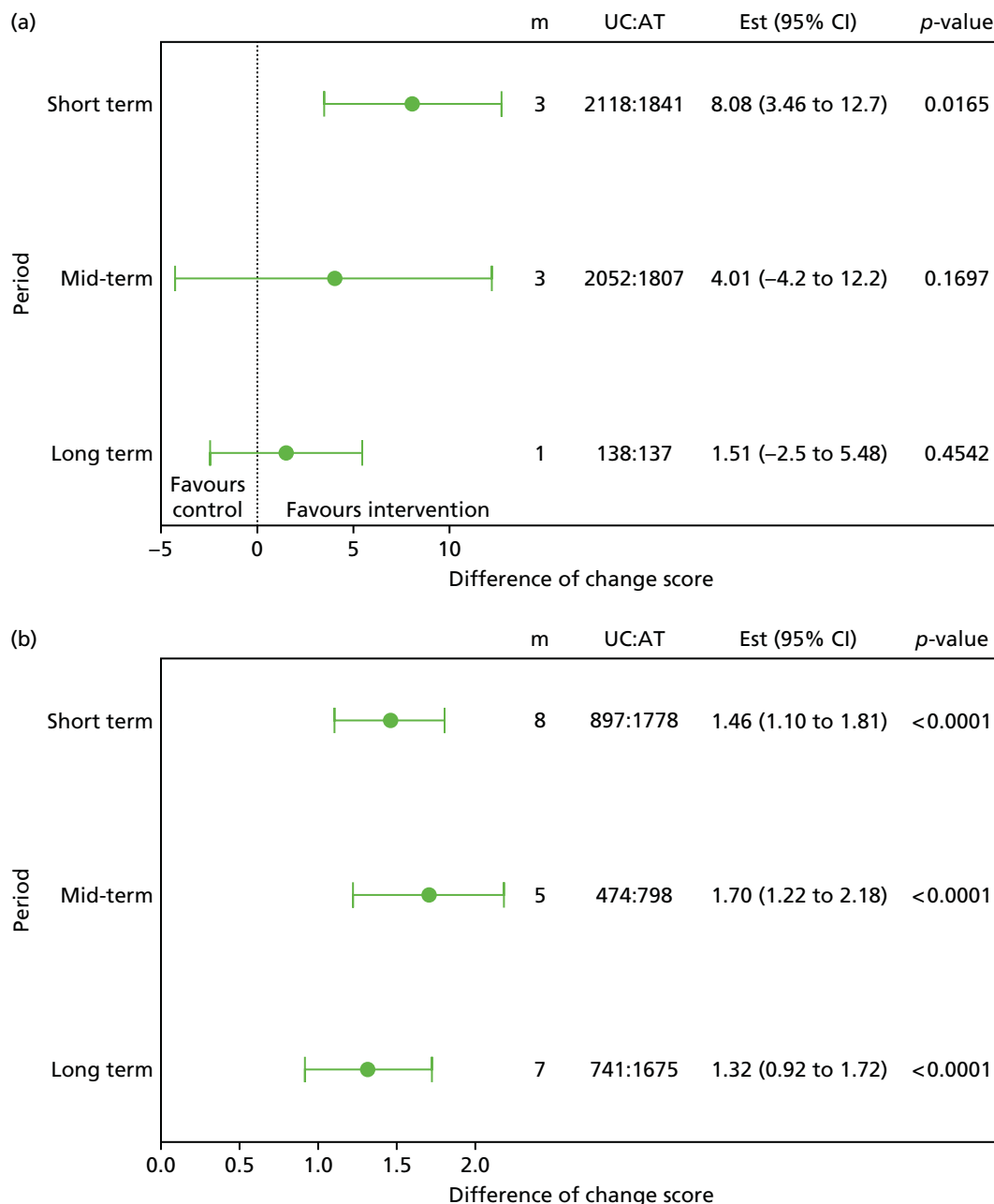


FIGURE 17 The estimated efficacy between control (non-active usual care and sham) and intervention treatments from one-step meta-analysis for (a) FFbHR; and (b) RMDQ score. AT, number of participants in the intervention arm; Est (95% CI), estimated treatment efficacy and 95% CI; *m*, number of trials; UC, number of participants in the control arm.

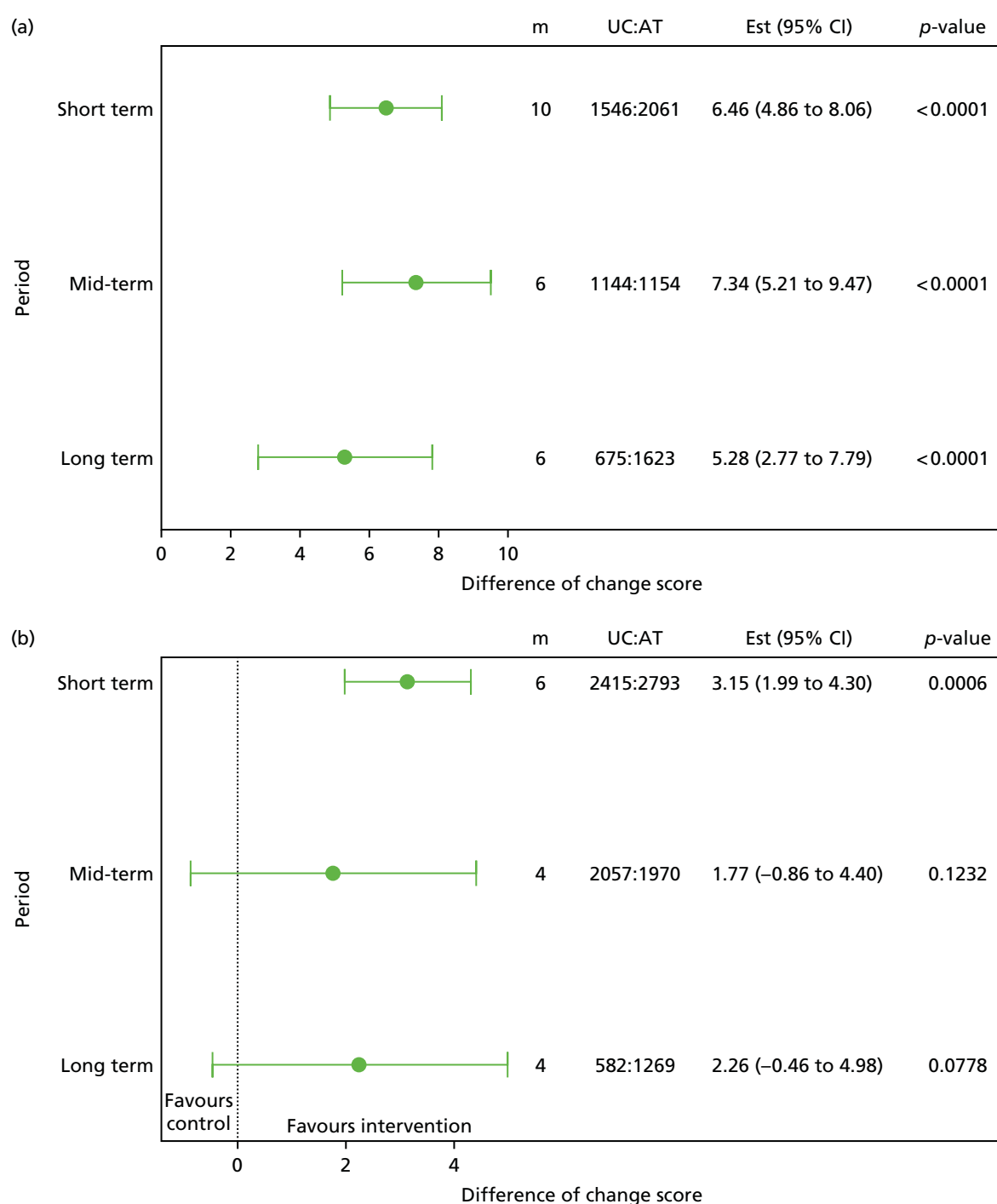


FIGURE 18 The estimated efficacy between control (non-active usual care and sham) and intervention treatments from one-step meta-analysis for (a) average pain (based on VAS); and (b) PCS of SF-12/36. AT, number of participants in the intervention arm; Est (95% CI), estimated treatment efficacy and 95% CI; *m*, number of trials; UC, number of participants in the control arm.

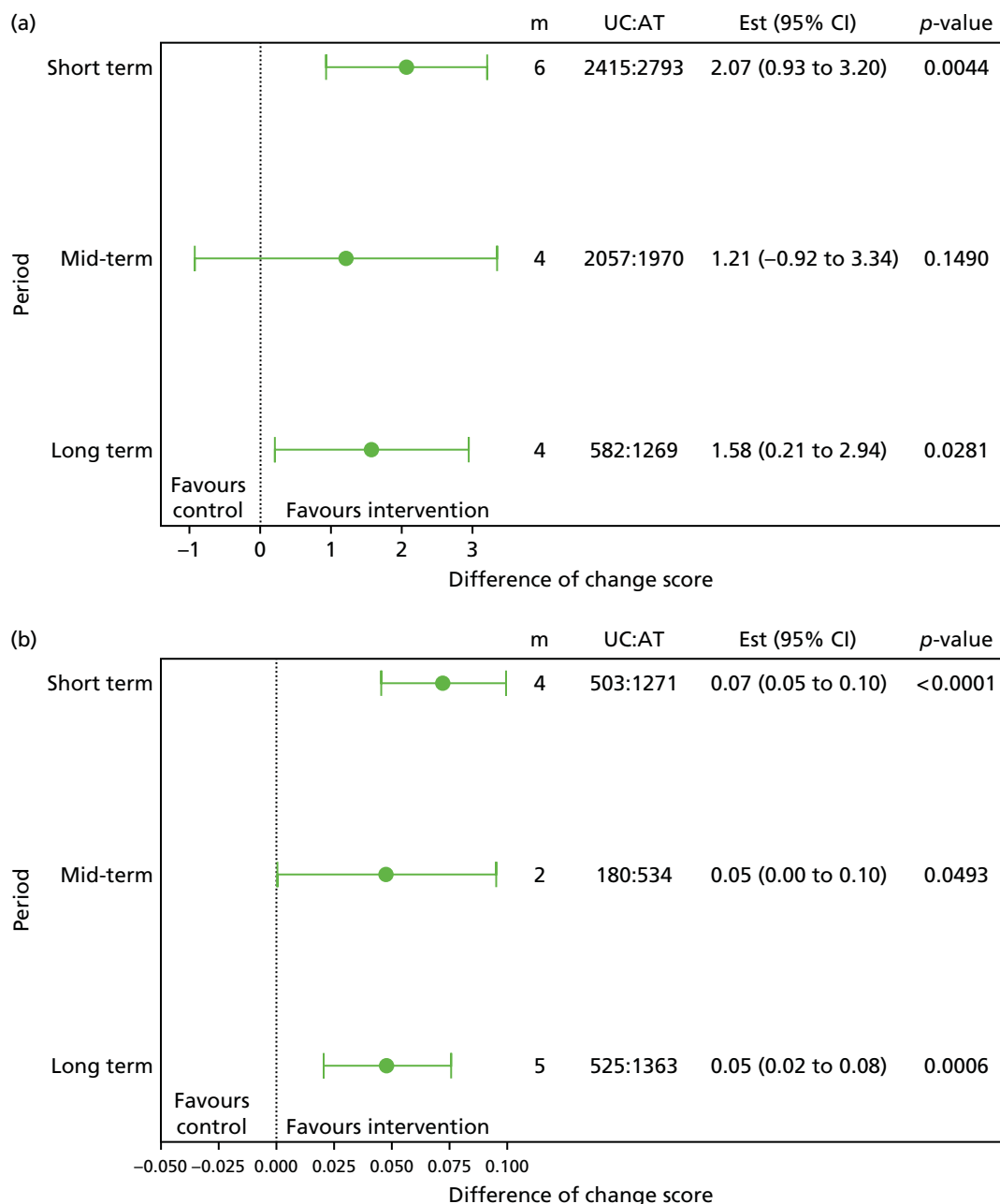


FIGURE 19 The estimated efficacy between control (non-active usual care and sham) and intervention treatments from one-step meta-analysis for (a) MCS of SF-12/36; and (b) EQ-5D. AT, number of participants in the intervention arm; Est (95% CI), estimated treatment efficacy and 95% CI; *m*, number of trials; UC, number of participants in the control arm.

TABLE 16 One-step meta-analysis: estimated mean change from baseline to short-term follow-up by treatment arms and the estimated difference between treatment arms (95% CI)^a

Outcomes	Number of trials, <i>m</i>	Intervention	Control ^b	Difference ^c	<i>p</i> -value
FFbHR	3	<i>n</i> = 1841	<i>n</i> = 2118		0.0165
		13.88	5.80	8.08	
		1.24 to 26.51	−6.93 to 18.53	3.46 to 12.69	
RMDQ	8	<i>n</i> = 1778	<i>n</i> = 897		< 0.0001
		4.43	2.97	1.46	
		1.56 to 7.29	0.10 to 5.84	1.10 to 1.81	
Average pain ^d	10	<i>n</i> = 2061	<i>n</i> = 1546		< 0.0001
		18.03	11.57	6.46	
		8.65 to 27.41	2.18 to 20.97	4.86 to 8.06	
PCS ^e	6	<i>n</i> = 2793	<i>n</i> = 2415		0.0006
		6.86	3.72	3.15	
		4.90 to 8.83	1.75 to 5.68	1.99 to 4.30	
MCS ^f	6	<i>n</i> = 2793	<i>n</i> = 2415		0.0044
		2.69	0.62	2.07	
		1.54 to 3.84	−0.55 to 1.79	0.93 to 3.20	
EQ-5D	4	<i>n</i> = 1271	<i>n</i> = 503		< 0.0001
		0.1065	0.03422	0.072	
		0.008 to 0.205	−0.059 to 0.127	0.04538 to 0.099	

a Adjusted by random intercept, trial and interaction between treatment and trial effects.

b Control, usual care/GP and sham control.

c Difference, intervention–control (thus, positive = favours intervention arm).

d Obtained from either VAS or CPG-PS (see *Selection of instrument*, above).

e PCS of SF-12/36.

f MCS of SF-12/36.

Analyses of covariance

Table 17 shows the list of moderators for each of the outcomes of interest at short-term follow-up, namely FFbHR, RMDQ, average pain, PCS and MCS. There were three trials^{50,101,132} with FFbHR short-term outcomes and the explanatory variables that may potentially be treatment moderators provided by these trials were age, sex, SF-12/36 PCS and SF-12/36 MCS. For the change of FFbHR from baseline to short-term follow-up, the treatment effect for younger participants was weakly statistically significant ($p = 0.2018$). Participants with lower value of FFbHR at baseline (more physical disability) had a larger treatment effect and this was statistically significant ($p < 0.0001$). Similarly, participants with lower value of PCS at baseline (substantial physical limitations) had larger treatment effect ($p < 0.0001$). Therefore, age, and the baseline values of FFbHR and PCS were considered for inclusion in further analyses.

TABLE 17 Analysis of covariance: analysis for short-term outcomes (change from baseline to short-term follow-up)^a

Outcome	Covariates	Number of trials, <i>m</i>	Number of participants, AT: UC	Estimate (interaction term)	LCI	UCI	<i>p</i> -value
FFbHR	Age	3	1841:2118	-0.051	-0.131	0.028	0.2018
	Sex (male vs. female) ^b	3	1841:2118	-0.684	-2.851	1.483	0.5361
	FFbHR	3	1841:2118	-0.177	-0.229	-0.125	<0.0001
	PCS (< 50 vs. ≥ 50) ^c	3	1718:2000	2.521	-2.361	7.403	0.3114
	PCS (continuous)	3	1718:2000	-0.318	-0.451	-0.186	<0.0001
	MCS (< 50 vs. ≥ 50)	3	1718:2000	0.612	-1.618	2.842	0.5903
	MCS (continuous)	3	1718:2000	-0.039	-0.130	0.051	0.3949
RMDR	Age	8	1778:897	-0.009	-0.036	0.018	0.514
	Sex (male vs. female)	8	1778:896	0.136	-0.591	0.863	0.7133
	RMDQ	8	1778:897	-0.017	-0.085	0.050	0.6176
	Average pain	8	1649:790	-0.003	-0.018	0.011	0.6548
	PCS (continuous)	2	1009:401	-0.016	-0.076	0.044	0.594
	PCS (< 50 vs. ≥ 50)	2	1009:401	0.546	-1.463	2.556	0.5939
	MCS (continuous)	2	1009:401	-0.002	-0.046	0.042	0.9177
	MCS (< 50 vs. ≥ 50)	2	1009:401	-0.423	-1.435	0.589	0.4123
	EQ-5D	3	1201:460	-0.366	-2.162	1.429	0.6892
	Anxiety	4	1388:523				0.3332
	Low risk ^d			-0.295	-1.713	1.123	0.6832
	Moderate risk ^e			0.452	-1.089	1.994	0.5649
	Depression	4	1387:525				0.5684
	Low risk			0.078	-1.337	1.492	0.9143
	Moderate risk			0.559	-0.933	2.051	0.4622
	Catastrophising	2	293:178				0.2360
	Positive ^f			0.387	-2.271	3.046	0.7747
	Moderate ^g			2.030	-0.461	4.521	0.1099
	Coping	3	620:348				0.6797
	Positive ^h			0.428	-1.127	1.982	0.5895
	Moderate ⁱ			0.729	-0.904	2.362	0.3813
	Fear avoidance	7	1706:858				0.1933
	Positive ^j			0.786	-0.125	1.697	0.0907
	Moderate ^k			0.714	-0.225	1.653	0.1361
Average pain ^l	Age	10	2061:1546	-0.047	-0.162	0.068	0.4216
	Sex (male vs. female)	10	2061:1545	0.784	-2.381	3.950	0.6272
	RMDQ	8	1657:794	0.156	-0.293	0.604	0.497
	Average pain	10	2061:1546	0.047	-0.017	0.111	0.1451

continued

TABLE 17 Analysis of covariance: analysis for short-term outcomes (change from baseline to short-term follow-up)^a (*continued*)

Outcome	Covariates	Number of trials, <i>m</i>	Number of participants, AT:UC	Estimate (interaction term)	LCI	UCI	<i>p</i> -value
SF-12/36 PCS ^m	PCS (continuous)	3	1390 : 1144	-0.167	-0.400	0.066	0.1587
	PCS (< 50 vs. ≥ 50)	3	1390 : 1144	1.569	-8.473	11.610	0.7594
	MCS (continuous)	3	1390 : 1144	0.111	-0.047	0.268	0.1677
	MCS (< 50 vs. ≥ 50)	3	1390 : 1144	-1.270	-4.942	2.403	0.498
	EQ-5D	3	1208 : 464	-3.192	-13.603	7.219	0.5477
	Anxiety	4	1394 : 528				0.2488
	Low risk			-6.939	-15.111	1.233	0.096
	Moderate risk			-5.509	-14.423	3.405	0.2256
	Depression	4	1394 : 530				0.9355
	Low risk			-1.519	-9.809	6.772	0.7195
	Moderate risk			-1.076	-9.843	7.692	0.8099
	Catastrophising	2	198 : 85				0.9797
	Positive			-0.400	-19.050	18.250	0.9664
	Moderate			-1.573	-17.280	14.133	0.8438
	Coping	3	544 : 264				0.4009
	Positive			-6.107	-14.999	2.786	0.178
	Moderate			-2.864	-11.995	6.266	0.5382
	Fear avoidance	8	1991 : 1505				0.3577
	Positive			1.396	-2.525	5.317	0.4851
	Moderate			2.808	-1.031	6.646	0.1516
	Age	6	2793 : 2415	-0.034	-0.068	0.001	0.0538
	Sex (male vs. female)	6	2793 : 2414	-0.176	-1.106	0.755	0.7111
	FFbHR	3	1675 : 1955	-0.016	-0.045	0.013	0.2766
	RMDQ	2	966 : 383	0.012	-0.210	0.234	0.9187
	Average pain	3	1346 : 1125	-0.011	-0.044	0.023	0.5313
	PCS (continuous)	6	2793 : 2415	-0.057	-0.109	-0.005	0.0313
	PCS (< 50 vs. ≥ 50)	6	2793 : 2415	1.995	0.018	3.973	0.048
	MCS (continuous)	6	2793 : 2415	0.023	-0.015	0.060	0.2395
	MCS (< 50 vs. ≥ 50)	6	2793 : 2415	-0.913	-1.827	0.002	0.0504
	EQ-5D	3	1046 : 425	1.216	-2.364	4.795	0.5054
	Anxiety	3	1051 : 428				0.6537
	Low risk			1.315	-1.638	4.267	0.3826
	Moderate risk			1.398	-1.750	4.545	0.3839
	Depression	3	1053 : 430				0.6277
	Low risk			1.261	-1.640	4.163	0.3939
	Moderate risk			1.462	-1.559	4.483	0.3427

TABLE 17 Analysis of covariance: analysis for short-term outcomes (change from baseline to short-term follow-up)^a (continued)

Outcome	Covariates	Number of trials, <i>m</i>	Number of participants, AT: UC	Estimate (interaction term)	LCI	UCI	<i>p</i> -value
SF-12/36 MCS ^b	Fear avoidance	3	1332 : 1114				0.8438
	Positive			−0.311	−2.029	1.408	0.7229
	Moderate			0.211	−1.435	1.857	0.8019
	Somatic symptoms	2	805 : 365				0.9147
	Positive ⁿ			0.542	−1.989	3.072	0.6746
	Moderate ^o			0.249	−1.907	2.405	0.8206
	Age	6	2793 : 2415	0.008	−0.035	0.050	0.7273
	Sex (male vs. female)	6	2793 : 2414	−0.324	−1.470	0.822	0.579
	FFbHR	3	1675 : 1955	−0.046	−0.081	−0.011	0.0093
	RMDQ	2	966 : 383	−0.011	−0.298	0.276	0.9395
	Average pain	3	1346 : 1125	−0.007	−0.048	0.034	0.7423
	PCS (continuous)	6	2793 : 2415	−0.035	−0.102	0.033	0.3133
	PCS (< 50 vs. ≥ 50)	6	2793 : 2415	0.649	−1.821	3.118	0.6067
	MCS (continuous)	6	2793 : 2415	−0.052	−0.093	−0.011	0.0128
	MCS (< 50 vs. ≥ 50)	6	2793 : 2415	1.490	0.442	2.539	0.0054
	EQ-5D	3	1046 : 425	−0.059	−4.576	4.458	0.9795
	Anxiety	3	1051 : 428				0.4267
	Low risk			−1.201	−4.918	2.517	0.5265
	Moderate risk			0.406	−3.558	4.369	0.8409
	Depression	3	1053 : 430				0.863
	Low risk			−0.334	−3.983	3.314	0.8573
	Moderate risk			0.343	−3.456	4.142	0.8594
	Fear avoidance	3	1332 : 1114				0.7926
	Positive			0.732	−1.378	2.843	0.4964
	Moderate			0.278	−1.744	2.299	0.7877
EQ-5D	Somatic symptoms	2	805 : 365				0.575
	Least			−0.978	−4.351	2.395	0.5695
	Moderate			0.789	−2.087	3.665	0.5906
	Age	4	1271 : 503	0.001	−0.001	0.003	0.503
	Sex (male vs. female)	4	1271 : 502	−0.040	−0.094	0.015	0.1543
	RMDQ	3	1177 : 455	0.007	0.001	0.013	0.0219
	Average pain	3	1183 : 459	0.002	0.000	0.003	0.0094
	PCS (continuous)	3	1068 : 439	−0.004	−0.008	−0.001	0.0128
	PCS (< 50 vs. ≥ 50)	3	1068 : 439	0.045	−0.072	0.162	0.4494
	MCS (continuous)	3	1068 : 439	−0.002	−0.004	0.001	0.1834

continued

TABLE 17 Analysis of covariance: analysis for short-term outcomes (change from baseline to short-term follow-up)^a (*continued*)

Outcome	Covariates	Number of trials, <i>m</i>	Number of participants, AT : UC	Estimate (interaction term)	LCI	UCI	<i>p</i> -value
QALY	MCS (< 50 vs. ≥ 50)	3	1068 : 439	0.024	-0.034	0.082	0.4102
	EQ-5D	4	1271 : 503	-0.054	-0.144	0.035	0.2358
	Anxiety	4	1269 : 500				0.0032
	Low risk			-0.143	-0.232	-0.055	0.0015
	Moderate risk			-0.086	-0.180	0.009	0.0753
	Depression	4	1265 : 500				0.5331
	Low risk			-0.033	-0.120	0.054	0.4573
	Moderate risk			-0.003	-0.094	0.088	0.9511
	Fear avoidance	3	1163 : 450				0.0533
	Positive			-0.001	-0.072	0.071	0.9856
	Moderate			0.073	-0.002	0.147	0.0565
	Age	6	1539 : 814	0.001	-0.0003	0.002	0.1850
	RMDQ	4	1092 : 422	0.003	-0.001	0.008	0.1270
	PCS (continuous)	4	1273 : 715	-0.001	-0.003	0.0004	0.1160
	MCS (continuous)	4	1273 : 715	-0.0001	-0.002	0.001	0.8340
	EQ-5D	4	1273 : 715	-0.018	-0.082	0.045	0.5730

AT, number of patients in the intervention arm (active physical, passive physical, psychological or combination; LCI, lower limit of the 95% CI; UC, number of patients in the control arm (usual care/GP or sham); UCI upper limit of the 95% CI.

a Mixed-effects models with intercept, trials and interaction between treatments and trials as random effects, and covariate and interaction between covariates.

b Estimate of the treatment effect for male was less than for female.

c Estimate of the treatment effect for participants with SF-12/36 PCS lower than general norm (< 50) was greater than for those with a score at or above the general norm (≥ 50).

d Estimate of the treatment effect for participants with low risk of anxiety was less than for those with high risk of anxiety.

e Estimate of the treatment effect for participants with moderate risk of anxiety was greater than for those with high risk of anxiety.

f Estimate of the treatment effect for participants with positive attitude of catastrophising (low catastrophising score) was greater than for those with negative attitude (high catastrophising score).

g Estimate of the treatment effect for participants with moderate attitude of catastrophising was greater than for those with negative attitude.

h Estimate of the treatment effect for participants with positive attitude of coping strategy (high coping score) was greater than for those with negative attitude (low coping score).

i Estimate of the treatment effect for participants with moderate attitude of coping strategy was greater than for those with negative attitude.

j Estimate of the treatment effect for participants with positive belief (low fear avoidance) of fear avoidance belief was greater than for those with negative attitude.

k Estimate of the treatment effect for participants with moderate belief of fear avoidance was greater than for those with negative attitude.

l Obtained from either VAS or CPG-PS (see *Selection of instrument*, above).

m PCS of SF-12/36.

n Estimate of the treatment effect for participants with least general somatic symptoms was greater than for those with more general somatic symptoms.

o Estimate of the treatment effect for participants with moderate general somatic symptoms was greater than for those with more general somatic symptoms.

p MCS of SF-12/36.

Roland–Morris Disability Questionnaire

There were eight trials^{31,33,70,103–105,131,136} with RMDQ score as a short-term outcome, and the explanatory covariates provided by them were age, sex, RMDQ score, average pain, PCS, MCS, EQ-5D, anxiety level, depression level, catastrophising, coping strategy and fear avoidance at baseline. Seven trials^{31,33,70,103–105,131} provided information on fear avoidance at baseline and the original values were mapped to a single ordinal categorical variable. The covariate was weakly statistically significant, at our lower threshold for inclusion in further analyses ($p < 0.20$), in moderating the change of RMDQ score over the short term, for which those with either positive or moderate attitude (lower fear avoidance score) had greater treatment effect than those with negative attitude (higher fear avoidance score). Although the covariate catastrophising was not statistically significant ($p = 0.236$) in predicting the change of RMDQ score in the short term, there was a weakly statistically significant difference between the moderate and negative statement (mean difference = 2.03; $p = 0.1099$), that is, those with a moderate attitude towards catastrophising had greater treatment effect than those with a negative attitude. Therefore, both fear avoidance and catastrophising were considered for the prediction rule analyses.

Pain

Ten trials^{31,33,70,103–105,131,132,135,136} provided an average pain short-term outcome. The list of covariates that were considered in the ANCOVA were age, sex, RMDQ score, average pain, PCS, MCS, EQ-5D, anxiety level, depression level, catastrophising, coping strategy and fear avoidance at baseline. Similar to the results seen for the change of RMDQ score in the short term, anxiety level, coping strategy and fear avoidance were not statistically significant but there was weakly significant difference between the low and high risk of anxiety level ($p = 0.0960$), between the positive and negative statement of coping strategy ($p = 0.1780$), and between the moderate and negative statement of fear avoidance ($p = 0.1516$). Similar to the results seen above, those with moderate fear avoidance belief had greater treatment effect than those with a negative attitude. However, those with low risk of anxiety had less treatment effect than those with high risk of anxiety. Similarly, those with a positive attitude towards coping had less treatment effect than those with a negative attitude. As the average pain increased, the estimated treatment effect was greater, that is, as participants had worse average pain, they gained greater treatment effect and this was weakly significant ($p = 0.1451$). The estimated treatment effect decreased as PCS increased, that is, as a participant's physical functioning score got worse, he/she had greater treatment effect ($p = 0.1587$). The interaction term between treatment and MCS was also weakly statistically significant ($p = 0.1677$), for which participants with higher (better) MCS had larger treatment effect. Therefore, average pain, PCS, MCS, anxiety level, coping strategy and fear avoidance at baseline were considered for the prediction rule analyses.

Mental component score and physical component score

There were six trials^{31,33,50,88,101,132} with PCS and MCS short-term outcomes and the covariates considered were age, sex, FFbHR, RMDQ, average pain, PCS, MCS, EQ-5D, anxiety level, depression level, fear avoidance and somatic symptoms. Psychological distress at baseline measured by the MCS instrument was not significant in predicting the change of PCS at short term but when the score was dichotomised to < 50 against ≥ 50 , that is, 'below the norm' against 'at or above the general population norm', participants with more psychological distress (score of < 50) had worse treatment effect and this was possibly statistically significant ($p = 0.0504$). In addition, age and PCS at baseline were significant when those who were younger and those with substantial physical limitations had a larger treatment effect. Therefore, age, PCSs and MCSs at baseline were included for the prediction rule analyses for the change of SF-12/36 PCS at short term.

For the short-term MCS outcome, only FFbHR and MCS at baseline were found to be statistically significant in predicting the change of SF-12/36 MCS. Those with higher physical disability and more psychological distress had a greater treatment effect. Therefore, both FFbHR score and MCS at baseline were included for the prediction rule analyses for the change of SF-12/36 MCS in the short term.

European Quality of Life-5 Dimensions (EQ-5D)

Four trials^{31,33,70,105} provided health utility measured by EQ-5D over the short term. The covariates examined in the ANCOVA were age, sex, RMDQ score, average pain, PCS, MCS, EQ-5D, anxiety level, depression level and fear avoidance. Seven of these were statistically or weakly significant in predicting the change of EQ-5D at short term and these were sex, RMDQ score, average pain, PCS, MCS, anxiety level and fear avoidance at baseline. Females had greater treatment effect ($p = 0.1543$) and so had those with worse physical disability (RMDQ score, $p = 0.0219$; average pain, $p = 0.0094$; PCS, $p = 0.0128$). Participants with more psychological distress at baseline, high risk of anxiety, high risk of depression, negative beliefs about physical activity affecting their LBP (fear avoidance) or frequent psychological distress (MCS) had a larger treatment effect. Therefore, these were considered for the prediction rule analyses.

Quality-adjusted life-years

There were six trials^{31,33,70,105,132,133} with QALY data, age and baseline RMDQ score and PCS, which were possibly statistically significant in moderating QALYs. The age-by-treatment interaction was possibly significant, with a coefficient of 0.001 and a p -value of 0.19. The coefficient was positive, suggesting that older participants within this sample achieved a higher treatment effect. The RMDQ score by treatment interaction was significant ($p = 0.13$) at our prespecified level of 0.2. The coefficient of 0.003 was positive. The scale on the RMDQ score is such that lower scores denote better health states; therefore, participants with better (lower) RMDQ scores should be peeled off first for the health-economic prediction rule analyses (see *Chapter 9*). The coefficient of PCS-by-treatment interaction was -0.001 ($p = 0.12$). The negative coefficient indicates that participants with a worse physical functioning score at baseline achieved a greater treatment effect than those with better physical functioning scores at baseline. The baseline scores of EQ-5D and MCS were not significant. The EQ-5D-by-treatment interaction was not significant, with a coefficient of -0.018 ($p = 0.57$). The coefficient was negative, suggesting that participants with worse baseline EQ-5D scores achieved better treatment outcomes. However, this result should not be considered reliable, given the low level of significance. The coefficient of MCS by treatment interaction was -0.0001 ($p = 0.83$).

Summary

This analysis has provided the largest analysis of possible treatment moderation in LBP. Overall, these analyses do not provide strong evidence for substantial effect moderation. Using conventional criteria for statistical significance we can conclude, overall, only that back pain disability moderates effect size on back pain disability outcomes (FFbHR moderates FFbHR); physical state and back pain moderate effect size on physical outcomes (PCS and FFbHR moderate PCS); psychological state moderates effect size on psychological outcomes (MCS moderates MCS); overall psychological state and anxiety moderate effect size on quality of life (PCS and anxiety moderate EQ-5D); and back pain severity moderates effect size on psychological outcomes (FFbHR moderates MCS).

Age, gender, back pain disability, pain severity, MCS, PCS, anxiety, catastrophising and coping were all at least weakly statistically significant ($p < 0.2$) in one, or more, ANCOVA and were considered further for our main analyses.

Chapter 7 Methodology and statistical developments 1: subgroup identification with recursive partitioning

In *Chapter 2* we concluded that current approaches using tests for interactions on single potential moderators may not be the best approach to identifying subgroups; specifically in the case of LBP but this may be generalisable to other disorders. We argued that new statistical methods may be needed to improve subgroup identification. In the succeeding chapters we describe our exploration of the different methods that we have applied to addressing this problem. In particular, we were interested in how subgroups might be defined using multiple parameters. We first describe two recursive partitioning approaches then an adaptive peeling approach and, finally, an indirect meta-analytical approach.

This chapter presents the two methodological developments, using recursive partitioning, to identify subgroup characteristics that moderate response to treatment. Both methods were the works of a PhD project that was part of this programme grant.¹⁶⁷ The other methods are described in later chapters (see *Chapters 8–10*).

Background

Two methods were considered as suitable and appropriate to perform subgroup analyses using a recursive partitioning approach. They are the interaction tree (IT) and subgroup identification based on a differential effect search (SIDES).^{97,99} These methods were initially developed and implemented in a single-trial setting. Therefore, they have to be extended so that they can be applied in an individual patient data (IPD) meta-analysis framework. The extended IT and SIDES methods are known as IPD-IT and IPD-SIDES, respectively. Details of each of these methods are given below.

Both IT and SIDES are tree-based methods that rely on technique referred to as recursive partitioning. This technique recursively forms binary splits of the covariate space in order to grow a tree-like structure. An example of a tree structure is displayed in *Figure 20*. In this example, we start off with the root node of the tree, which consists of the entire data set. The method then searches all possible binary splits for every covariate to find the best split that maximises some splitting criterion. Suppose that 'sex' is identified as the first best split. The method, therefore, splits the root node using the sex covariate to form two child nodes;

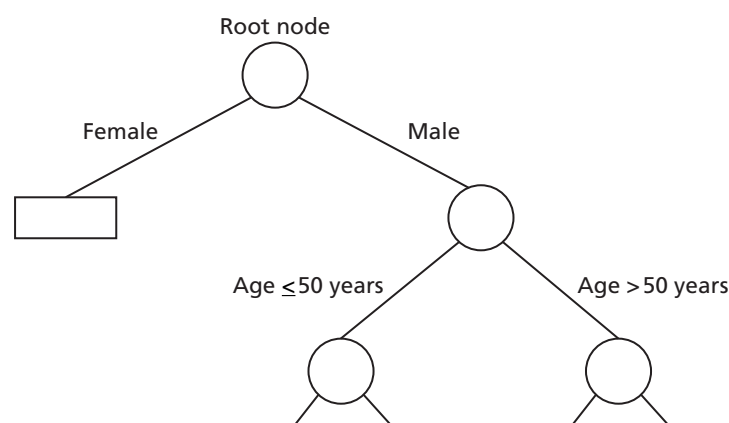


FIGURE 20 Example of a tree structure.

females (left child node) and males (right child node). The newly formed child nodes are also referred to as internal nodes. The same search process is then conducted on all of the internal nodes of the tree, that is, the two child nodes, to try and identify the next best split. No additional splits are identified for the left child node and hence the node is not split any further. This node is thus referred to as a terminal node, as it cannot be split any further and is represented by a square box in *Figure 20*. For the right child node, the method identifies age of ≤ 50 years as the next best split and thus forms two new child nodes accordingly. In the same manner, this search process is repeated until a full tree is grown.

The objective of both the IPD-IT and IPD-SIDES methods are somewhat different. The aim of the IPD-IT method is to identify moderators of treatment effect whereas the aim of the IPD-SIDES method is to identify candidate subgroups with enhanced treatment effect. In other words, the IPD-IT method is driven by identifying the split that results in the largest interaction effect whereas the IPD-SIDES method is driven by identifying the split that maximises the overall treatment benefit in one of the subgroups formed from the split.

Individual patient data interaction tree

The IPD-IT method primarily consists of three steps:

1. growing an initial tree
2. pruning the initial tree
3. selecting the best tree.

The third and final step in the process will result in either tree structure with just the root node (i.e. no moderators identified) or a larger tree structure that stems from the root node (i.e. some moderators identified). In the latter case, the subgroups identified by the final selected tree are interpreted using its terminal nodes.

Growing an initial tree

The first iteration of the procedure starts at the root node and evaluates a splitting criterion that assesses the interaction effect for every possible binary split of each covariate in order to identify an optimal split. For a continuous or discrete ordered covariate, the total number of binary split points is just one fewer than the total number of distinct values. For example, a discrete ordered covariate with 10 distinct values will have $10 - 1 = 9$ possible split points. For a categorical covariate with k different categories, there are $2^{k-1} - 1$ different split points. For example, a categorical covariate such as ethnicity with four different categories (white, Asian/Asian British, black/African/Caribbean/black British, and other) will have seven possible ways of forming two groups using a binary split.

The splitting criterion is used to evaluate the interaction effect for any particular split. The original IT method used a splitting criterion that was equivalent to the square of the t -test statistic of the interaction term in a linear regression model consisting of a treatment indicator variable T , a covariate indicator representing a particular split X and the interaction between T and X . As we are now applying this method to IPD from different trials, we extended the original method so that the splitting criterion adjusts for the between-trial variability when evaluating the interaction. This was done by fitting the same linear regression model but also including dummy variables for each trial, that is, fitting a fixed-effects model.¹⁶⁷ A split with a larger splitting criterion value indicates a larger interaction effect. Therefore, an optimal split

is defined as the split that maximises the splitting criterion having searched every possible split point of each covariate. Having defined the splitting criterion, the algorithm for growing a full tree can be applied as follows:

- Start at the root node consisting of the entire data set.
- Iteration:
 - *Step 1* Evaluate the splitting criterion for all possible splits for every single covariate.
 - *Step 2* Select the optimal split from step 1 and form a split to create two new child nodes.
 - *Step 3* Repeat steps 1 and 2 for each of the newly formed child nodes.
 - *Step 4* Repeat steps 1–3 until either a full tree is grown or some stopping criterion is satisfied, for example minimum number of observations in a node is 30.

Pruning the initial tree

The fully grown tree is well fitted to the available data; however, it would be quite poorly fitted and unstable if applied to new data. For this reason, a pruning procedure is applied to the full tree to sequentially remove any branches of the tree that least contribute to the overall predictive accuracy of the tree. The procedure continues until we are just left with the root node and thus have a sequence of subtrees from which the optimal final subtree will be chosen. A more detailed description of the pruning procedure can be found elsewhere.^{97,167,168}

Selecting the best tree

Once the sequence of subtrees has been determined, an interaction complexity measure is used to evaluate the quality of each tree. The interaction complexity is basically the total amount of interaction of the internal nodes for a tree. Although the interaction–complexity measure is computed for each of the subtrees, these estimates are known to be over-optimistic and thus need to be validated to obtain more reliable estimates. To validate the tree selection, the method applies a bootstrapping procedure, used by LeBlanc and Crowley,¹⁶⁹ for validating the trees. As a guideline, LeBlanc and Crowley¹⁶⁹ suggested that around 25–100 bootstrap samples is sufficient. The subtree with the largest interaction–complexity measure estimated from the bootstrapping procedure is chosen as the best tree. Conclusions can then be drawn from the best tree by simply computing the treatment effect in each of the terminal nodes of the tree.

Individual patient data subgroup identification based on a differential effect search

The IPD-SIDES method consists of two key steps:

1. growing an initial tree
2. selecting the final candidate subgroups.

The tree growing procedure for the IPD-SIDES method (step 1) relies on two different criteria; a splitting criterion to help search the covariate space for the best splits and a continuation criterion to control the complexity of the tree. Details are given below. Unlike the IPD-IT procedure, the IPD-SIDES method does not require a pruning step, as the tree complexity is controlled using the continuation criterion. Ultimately, after step 2, the method outputs a list of candidate subgroups that have enhanced treatment effect.

Growing an initial tree

We first describe the algorithm for the IPD-SIDES procedure followed by a more detailed description of the splitting criterion and the continuation criterion. The algorithm for growing the tree is as follows:

- Start at the root node consisting of the entire data set.
- Iteration:
 - *Step 1* Evaluate the splitting criterion for all splits of every covariate, excluding any covariates that have already been used to define the parent node, retaining only the best split for each covariate. Order the covariates from smallest adjusted p -value to largest adjusted p -value, where the adjusted p -values are computed using the Sidak-based multiplicity adjustment.
 - *Step 2* Select the best M covariates from the ordered best splits. The value of M is specified by the user where the recommended value is 5. For each of the M splits, form the split creating two child nodes and retain the child node with the larger positive treatment effect, providing that it satisfies the continuation criterion. The retained nodes now become parent nodes for the next iteration.
 - *Step 3* Repeat steps 1 and 2 for the newly formed parent nodes.
 - *Step 4* Repeat steps 1–3 until either a prespecified maximum number of levels is reached or no more splits can be formed, that is, the continuation criterion is not satisfied. In both cases, the previously formed parent nodes become terminal nodes.

The IPD-SIDES procedure starts at the root node consisting of the entire data set. The method then evaluates the splitting criterion for all splits for every covariate, retaining only the single best split for each covariate. The original SIDES method used a splitting criterion in a single-trial setting, which tested the difference in the treatment effect precision between two child nodes with the aim of identifying the subgroup or child node with the most significant treatment effect. This objective is different from what we require the method to do; we require the method to test the differential treatment effect between the two groups in an IPD meta-analysis setting. For this reason, a new splitting criterion was proposed, which uses the same fixed-effects model described earlier for the IPD-IT method but instead uses the p -value of the interaction effect, for which a smaller p -value is indicative of a larger interaction effect. If a covariate has more than two distinct cut-off points, the p -value computed using the splitting criterion is adjusted to overcome variable selection bias – a well-known issue with recursive partitioning-based methods when covariates with a larger number of splits have a greater probability of being chosen as the splitting variable.^{170,171} The method adjusts the p -value by applying a Sidak-based multiplicity adjustment, as described in the original SIDES method paper.⁹⁹

Continuation criterion

In step 2 of the IPD-SIDES iteration algorithm, a child node with a large positive treatment effect is retained only if it satisfies the continuation criterion. The continuation criterion is given by *Equation 2*:

$$p_c \leq \gamma \cdot p_p, \quad (2)$$

where p_c is the treatment effect p -value of the child node, p_p is the treatment effect p -value of the parent node and γ is the relative improvement parameter that controls the complexity of the tree. Prior to running the method, the user must specify the maximum number of covariates, L , that defines a subgroup, for which the recommended value is '3'. This means that any identified subgroups will at most be defined by L covariates and hence the tree will have at most L levels. Each level of the tree has a relative improvement parameter value that ranges from 0 to 1, for which a smaller value makes the procedure more selective. The values for each level can be either user specified or optimally selected using a cross-validation procedure as described by the authors.⁹⁹ Hence once the relative improvement parameter values are in place, a child node is retained only if its treatment effect p -value is less than or equal to the right-hand side of the continuation criterion.

Selecting the final candidate subgroups

The first step of the IPD-SIDES procedure grows the tree and produces a list of candidate subgroups. Many of these subgroups may be spurious findings and thus need to be removed. To control for this, the authors of the original SIDES method proposed a resampling-based procedure that computes an adjusted treatment effect p -value for each of the identified candidate subgroups to control the overall type I error in the weak sense.⁹⁹ Comparing the unadjusted p -value to the adjusted p -value gives a good indication of whether or not the identified subgroups are spurious.

Analyses

Two sets of analyses were performed using the repository data. In the first analyses (analysis 1), we grouped all of the interventions together as being one arm, and grouped the non-active usual care and sham control together as being the comparator arm. We then sought to identify subgroups within these data by applying the IPD-IT and IPD-SIDES methods. These analyses were performed for all of the following absolute change from baseline to short-term follow-up outcome variables: average pain, EQ-5D, FFbHR, MCS of SF-12/36, PCS of SF-12/36 and RMDQ.

In addition to the above outcome measures, we also looked at the QALYs health-economics outcome. This analysis provides proof of principle that the analytical techniques are robust when used with real data rather than simply in the simulated data sets in which we originally developed our techniques.¹⁶⁷

In the second set of analyses (analysis 2), the following interventions against the non-active usual care comparisons were investigated for subgroups:

1. active physical against non-active usual care
2. passive physical against non-active usual care
3. psychological against non-active usual care
4. sham against non-active usual care.

Both the IPD-IT and IPD-SIDES methods were applied to the above for each of the short-term outcomes common to all trials. For example, active physical against non-active usual care may consist of three trials with RMDQ, MCS and PCS as common short-term outcome measures. Thus the analyses would be applied to only these three outcome measures.

Prior to performing each of the analyses, any observations with missing data were removed from the data set. A mixed-effects model was then applied to adjust for the clustering inherent within the data and thus obtain an estimate of the overall treatment effect. In both sets of analyses, the potential moderator variables identified from the univariate analyses as well as those moderators identified in systematic review 1 (see *Chapter 2*) were considered. From this set of moderator variables, only the variables that were most common across all trials were entered into each of the analyses in order to retain as much data as possible.

The IPD-IT and IPD-SIDES methods both require certain parameters to be prespecified to aid or control the methods when applied to the data. For both methods, the minimum number of participants in any given node of a tree was set to $r = 1/20$ of the population being analysed. The maximum number of splits for the fully grown IPD-IT tree was set as 15. For the IPD-SIDES methods, the maximum number of levels, that is, the maximum number of covariates defining any particular subgroup, was set as being the number of potential moderators being considered. Moreover, the maximum number of best splits to consider for each node during the IPD-SIDES procedure was set to '3', with a restriction of $p \leq 0.20$ placed on the splitting criterion. This is the same constraint that we set in the identification of a promising moderator.

Before applying the IPD-SIDES method, we performed a grid search to obtain an optimal sequence of complexity control parameters for the first three levels of the tree. The grid search considered all permutations from 0.2 to 1, in steps of 0.2 at the first level and then from 0 to 1 in steps of 0.2 at levels two and three. When validating or selecting the final subgroups, we used 500 bootstraps for the IPD-IT procedure and used 1000 repetitions of the resampling procedure for the IPD-SIDES procedure. Any identified subgroups from the analyses were then summarised using the treatment effect and 95% CI. All analyses were performed using R version 3.0.3 (The R Foundation for Statistical Computing, Vienna, Austria).

Results

Analysis 1

The intervention (active physical, passive physical or psychological given either singly or as combined regimen with the other interventions) against control/placebo data were searched for subgroups for the first set of analyses. *Table 18* provides a summary of the trials included and the variables used to search for subgroups for each short-term outcome measure. Number included from each trial is dependent on the number of complete cases available for each analysis.

Subgroups identified by the individual patient data interaction tree method

The IPD-IT method did not identify any subgroups that moderate treatment effect when comparing any intervention compared with usual care control/sham.

TABLE 18 Summary of the included trials and variables used for each short-term outcome measure in analysis 1

Outcome ^a	Trials	Variables
Average pain	$m = 2$; $n = 1377$ ^b UK BEAM ³¹ ($n = 910$) ^c BeST ³³ ($n = 467$)	Age, sex, anxiety, fear avoidance, MCS, PCS, average pain and RMDQ score at baseline
EQ-5D	$m = 2$; $n = 1339$ UK BEAM ³¹ ($n = 883$) BeST ³³ ($n = 456$)	Age, sex, anxiety, fear avoidance, MCS, PCS, RMDQ and average pain at baseline
FFbHR	$m = 3$; $n = 3718$ ^d Brinkhaus ¹⁰¹ ($n = 284$) ^e Haake ¹³² ($n = 1110$) ^f Witt ⁵⁰ ($n = 2324$)	Age, sex, PCS, FFbHR and MCS at baseline
MCS ^g	$m = 3$; $n = 3630$ Brinkhaus ¹⁰¹ ($n = 281$) Haake ¹³² ($n = 1110$) Witt ⁵⁰ ($n = 2239$)	Age, sex, FFbHR, MCS and PCS at baseline

TABLE 18 Summary of the included trials and variables used for each short-term outcome measure in analysis 1 (*continued*)

Outcome ^a	Trials	Variables
PCS ^h	<i>m</i> = 6; <i>n</i> = 5208 UK BEAM ³¹ (<i>n</i> = 893) BeST ³³ (<i>n</i> = 470) Brinkhaus ¹⁰¹ (<i>n</i> = 281) Haake ¹³² (<i>n</i> = 1110) Witt ⁵⁰ (<i>n</i> = 2248) ⁱ YACBAC ¹⁰⁷ (<i>n</i> = 206)	Age, sex, MCS and PCS at baseline
RMDQ	<i>m</i> = 7; <i>n</i> = 2564 UK BEAM ³¹ (<i>n</i> = 951) BeST ³³ (<i>n</i> = 488) ^j Hancock ¹³¹ (<i>n</i> = 235) ^k Pengel ¹⁰³ (<i>n</i> = 236) ^l Smeets ⁷⁰ (<i>n</i> = 212) ^m VKBIA ¹⁰⁴ (<i>n</i> = 229) ⁿ VKSC2 ¹⁰⁵ (<i>n</i> = 213)	Age, sex, fear avoidance and RMDQ score at baseline
QALY ^o	<i>m</i> = 4; <i>n</i> = 1514 UK BEAM ³¹ (<i>n</i> = 728) BeST ³³ (<i>n</i> = 468) Smeets ⁷⁰ (<i>n</i> = 151) ^p York BP ¹³³ (<i>n</i> = 167)	Age and RMDQ score at baseline

YACBAC, York Acupuncture Back Pain Trial.

a Change from baseline to short-term follow-up (between 2 and 3 months post randomisation or entry to the trial).

b UK BEAM (Exercise, spinal manipulation, combined, best care).

c BeST (cognitive behavioural approach, control).

d Brinkhaus (acupuncture, minimal acupuncture, waiting list).

e Haake (verum acupuncture, sham acupuncture, conventional therapy).

f Witt (acupuncture, control).

g MCS of SF-12/36.

h PCS of SF-12/36.

i YACBAC (traditional acupuncture, usual care).

j Hancock (spinal manipulation, placebo spinal manipulation, advice).

k Pengel (exercise, sham exercise, advice, sham advice).

l Smeets (APT, cognitive-behavioural treatment, combined treatment, waiting list).

m Von Korff BIA (Brief Individualised Programme, usual care).

n Von Korff SC2 (Self-Care, usual care).

o The QALY that was measured over 1 year of follow-up using the AUC method.

p York BP (exercise, control).

Subgroups identified by the individual patient data subgroup identification based on a differential effect search method

The application of the IPD-SIDES method for the first set of analyses found candidate subgroups for three of the short-term outcome measures when comparing intervention with control/placebo (*Table 19*); namely short-term FFbHR (*Figure 21*), SF-12/36 MCS (*Figure 22*) and SF-12/36 PCS (*Figure 23*). No candidate subgroups were identified for the average pain, EQ-5D and RMDQ short-term outcomes, as well as the QALY health outcome measure.

Short-term Hannover Functional Ability Questionnaire for measuring back pain-related functional limitations outcome

For the short-term FFbHR outcome, five variables were included in the IPD-SIDES analyses. The overall treatment effect for the FFbHR outcome was 8.93 (95% CI 7.81 to 10.05). Three candidate subgroups with enhanced treatment effect were identified by the IPD-SIDES procedure. Those with baseline FFbHR score of ≤ 54.2 had a treatment effect of 11.31 (95% CI 9.38 to 13.23), those with baseline FFbHR score of ≤ 54.2 and age ≤ 60 years had a treatment effect of 13.17 (95% CI 10.56 to 15.77) and those with FFbHR score of ≤ 54.2 and age ≤ 66 years had a treatment effect of 12.26 (95% CI 10.06 to 14.46).

- Those with more disability at baseline and who are younger are likely to gain a greater benefit on disability.

Short-term mental component scale of SF-12/36 outcome

For the short-term MCS outcome, five variables were included in the IPD-SIDES analyses. The overall treatment effect for the MCS outcome was 2.61 (95% CI 1.92 to 3.29). Only one candidate subgroup was identified for MCS outcome. Those with baseline MCS of ≤ 54.4 had a treatment effect of 3.46 (95% CI 2.62 to 4.30).

- Those with more psychological distress at baseline will get better outcomes on psychological distress.

Short-term physical component scale of SF-12/36 outcome

For the short-term PCS outcome, four variables were included in the analyses and four candidate subgroups were identified. The overall treatment effect for the PCS outcome was 3.48 (95% CI 3.01 to 3.96). Those with baseline MCS of > 50.9 had a treatment effect of 4.09 (95% CI 3.32 to 4.87), those with baseline MCS of > 50.9 and female had a treatment effect of 4.72 (95% CI 3.67 to 5.78), those with baseline MCS of > 50.9 and baseline PCS of ≤ 43.2 had a treatment effect of 4.62 (95% CI 3.75 to 5.49) and, finally, those with baseline MCS of > 50.9 and baseline PCS of ≤ 40.0 had a treatment effect of 4.89 (95% CI 3.96 to 5.82).

- Those with less psychological distress and worse physical status will get better outcomes on physical status.
- Women with low levels of psychological distress will get better outcomes on physical status.

These analyses do not consider any differences between different treatment approaches.

Analysis 2: pairwise comparisons

Each of the subgrouped interventions (active physical, passive physical or psychological) against non-active usual care data were searched for subgroups for the second set of analyses. *Table 20* provides a summary of the trials included and the variables used to search for subgroups for each short-term outcome measure analysed for the different comparisons.

Subgroups identified by the 'individual patient data interaction tree' method

The IPD-IT method did not identify any subgroups that moderate treatment effect when comparing any of the subgrouped interventions against non-active usual care.

TABLE 19 Candidate subgroups identified by the IPD-SIDES method for the intervention vs. control/ placebo comparison^a

Subgroups	n	Treatment effect (95% CI)	Interaction effect	Unadjusted p-value
Outcome: short-term FFbHR^b				
Overall treatment effect (95% CI): 8.93 (7.81 to 10.05)				
Candidate 1				
FFbHR ≤ 54.2	1709	11.31 (9.38 to 13.23)	4.69	<0.001
FFbHR > 54.2	2009	6.62 (5.46 to 7.78)		
Candidate 2				
FFbHR ≤ 54.2 and Age ≤ 60	1043	13.17 (10.56 to 15.77)	5.03	0.019
FFbHR ≤ 54.2 and Age > 60	666	8.14 (5.47 to 10.80)		
Candidate 3				
FFbHR ≤ 54.2 and Age ≤ 66	1367	12.26 (10.06 to 14.46)	5.14	0.043
FFbHR ≤ 54.2 and Age > 66	342	7.12 (3.42 to 10.82)		
Outcome: short-term MCS^c				
Overall treatment effect (95% CI): 2.61 (1.92 to 3.29)				
Candidate 1				
MCS ≤ 54.4	2541	3.46 (2.62 to 4.30)	2.62	0.002
MCS > 54.4	1089	0.84 (0.01 to 1.67)		
Outcome: short-term PCS^d				
Overall treatment effect (95% CI): 3.48 (3.01 to 3.96)				
Candidate 1				
MCS > 50.9	2082	4.09 (3.32 to 4.87)	0.97	0.033
MCS ≤ 50.9	3126	3.12 (2.54 to 3.71)		
Candidate 2				
MCS > 50.9 and Sex = Female	1125	4.72 (3.67 to 5.78)	1.38	0.097
MCS > 50.9 and Sex = Male	957	3.34 (2.20 to 4.48)		
Candidate 3				
MCS > 50.9 and PCS ≤ 43.2	1666	4.62 (3.75 to 5.49)	2.61	0.020
MCS > 50.9 and PCS > 43.2	416	2.01 (0.69 to 3.33)		
Candidate 4				
MCS > 50.9 and PCS ≤ 40.0	1457	4.89 (3.96 to 5.82)	2.61	0.007
MCS > 50.9 and PCS > 40.0	625	2.28 (1.12 to 3.44)		
^a The first row of each candidate subgroup is the selected subgroup with enhanced treatment effect. ^b FFbHR, ranging from 0 to 100 where a lower score represents greater disability. ^c MCS of SF-12/36 ranging from 0 to 100 where a lower score represents worse mental functioning. ^d PCS of SF-12/36 ranging from 0 to 100 where a lower score represents worse physical functioning.				

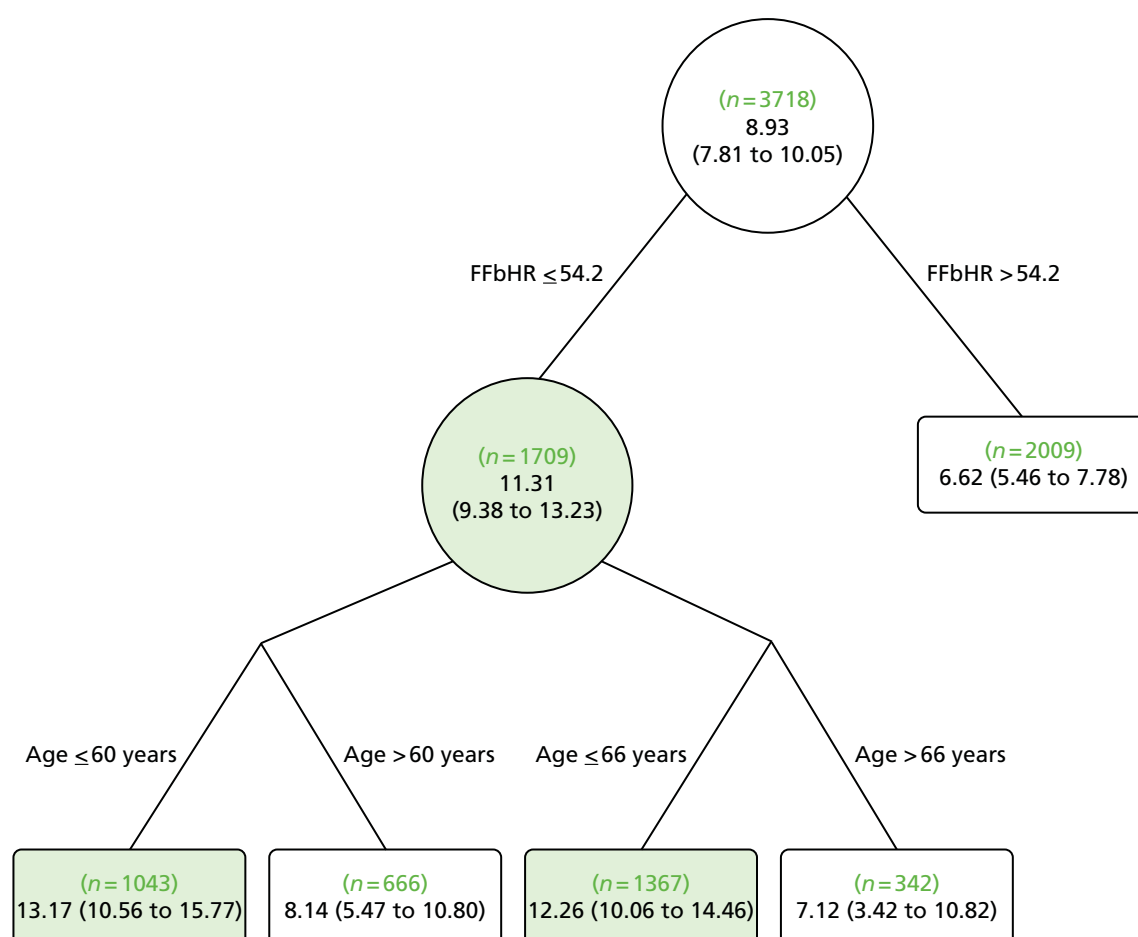


FIGURE 21 Candidate subgroups identified (shaded green) by the IPD-SIDES method when applied to change from baseline to short-term FFbHR (range 0–100; lower score implies greater disability) outcome for the intervention against control/placebo comparison. Results presented as treatment effect (95% CI) in each node.

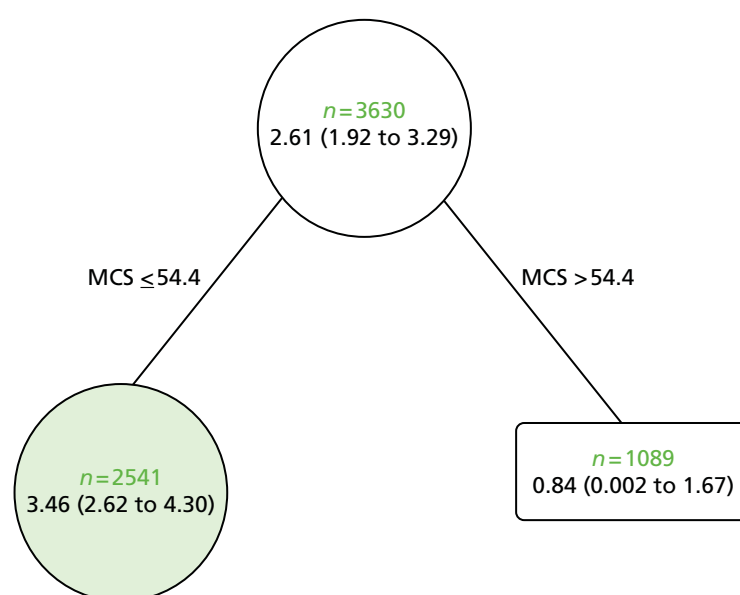


FIGURE 22 Candidate subgroup identified (shaded green) by the IPD-SIDES method when applied to change from baseline to short-term SF-12/36 MCS outcome (range 0–100; lower score implies worse mental functioning) for the intervention against control/placebo comparison. Results presented as treatment effect (95% CI) in each node.

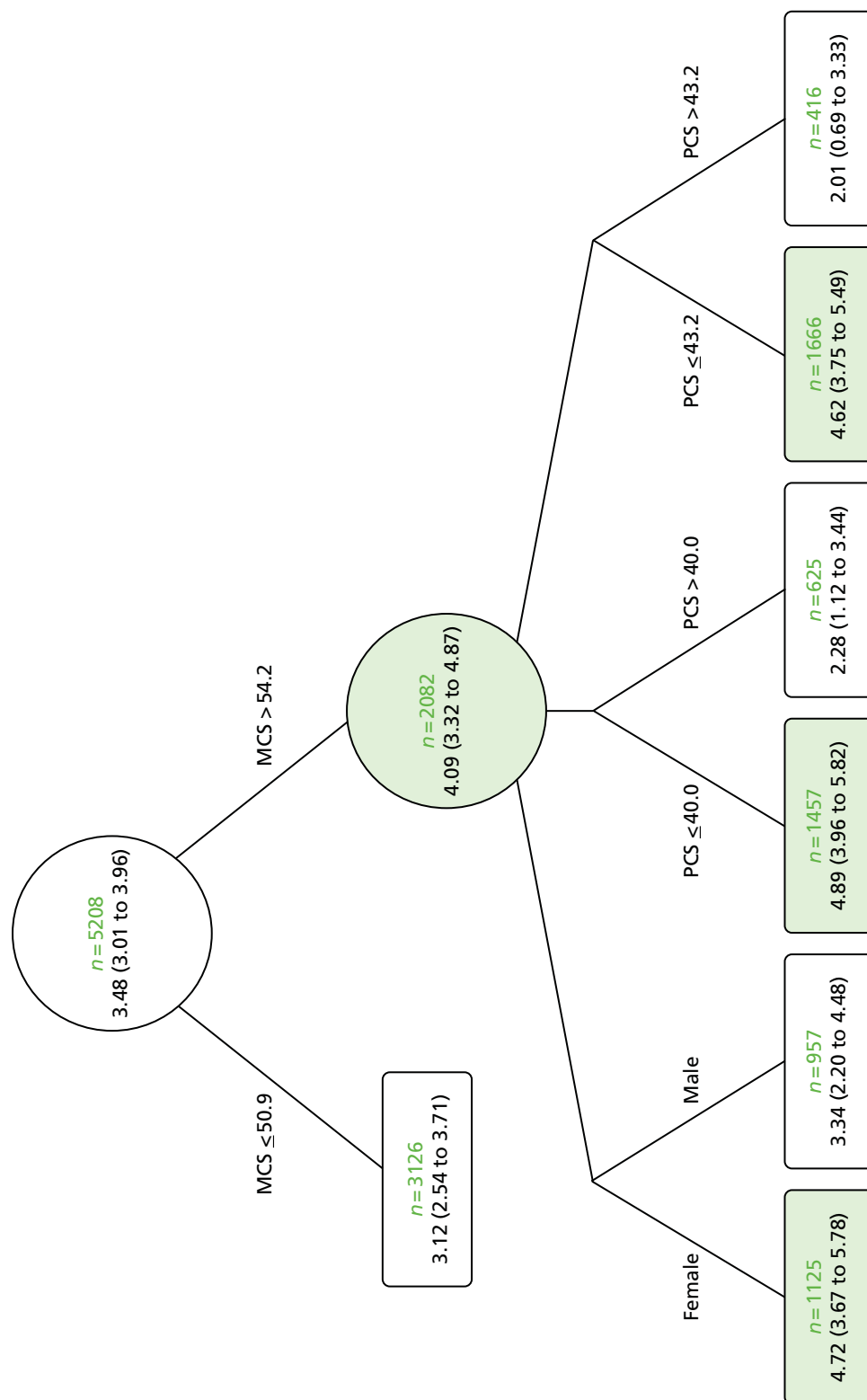


FIGURE 23 Candidate subgroups identified (shaded green) by the IPD-SIDES method when applied to change from baseline to short-term SF-12/36 PCS outcome (range 0–100; lower score implies worse physical functioning) for the intervention against control/placebo comparison. Results presented as treatment effect (95% CI) in each node.

TABLE 20 Summary of the trials included and variables used for each change from baseline to short-term outcome measure and the QALY health outcome measure analysed for the different comparisons

Comparison	Short-term outcome measures						QALY	
	FFbHR	RMDQ		MCS		PCS	Variables	Variables
	Trials ^a	Variables	Trials ^a	Variables	Trials ^a	Variables	Trials ^a	Variables
Active vs. non-active usual care	–	–	m = 2; n = 576 UK BEAM ³¹ (n = 421), Smeets ⁷⁰ (n = 155)	Fear avoidance, age, sex, RMDQ, average pain today, EQ-5D, HADS anxiety, HADS depression	–	–	m = 2; n = 496 UK BEAM ³¹ (n = 329), York Bp ¹³³ (n = 167)	Age, RMDQ
Passive vs. non-active usual care	m = 3; n = 3272 Brinkhaus ¹⁰¹ (n = 214), Haake ¹³² (n = 734), Witt ⁵⁰ (n = 2324)	Age, PCS, FFbHR, sex, MCS	–	–	m = 5; n = 3879 UK BEAM ³¹ (n = 479), Brinkhaus ¹⁰¹ (n = 212), Haake ¹³² (n = 734), Witt ⁵⁰ (n = 2248), YACBAC ¹⁰⁷ (n = 206)	MCS, age, sex, PCS	m = 5; n = 3879 UK BEAM ³¹ (n = 479), Brinkhaus ¹⁰¹ (n = 212), Haake ¹³² (n = 734), Witt ⁵⁰ (n = 2248), YACBAC ¹⁰⁷ (n = 206)	Age, MCS, PCS, sex
Psychological vs. non-active usual care	–	–	m = 3; n = 928 BeST ³³ (n = 487), VKBIA ¹⁰⁴ (n = 229), VKSC2 ¹⁰⁵ (n = 212)	Fear avoidance, age, sex, RMDQ, average pain today	–	–	–	–
Sham vs. non-active usual care	m = 2; n = 881 Brinkhaus ¹⁰¹ (n = 144), Haake ¹³² (n = 737)	Age, PCS, FFbHR, sex, MCS	–	–	m = 2; n = 879 Brinkhaus ¹⁰¹ (n = 142), Haake ¹³² (n = 737)	MCS, age, sex, PCS	m = 2; n = 879 Brinkhaus ¹⁰¹ (n = 142), Haake ¹³² (n = 737)	Age, MCS, PCS, sex

HADS, Hospital Anxiety and Depression Scale; YACBAC, York Acupuncture Back Pain Trial.

a UK BEAM³¹ (exercise, spinal manipulation, combined, best care); Smeets⁷⁰ (APT, cognitive-behavioural treatment, combined treatment, waiting list); York Bp¹³³ (exercise, control); Brinkhaus¹⁰¹ (acupuncture, minimal acupuncture, sham acupuncture, conventional therapy); Witt⁵⁰ (acupuncture, control); YACBAC¹⁰⁷ (traditional acupuncture, usual care); BeST³³ (cognitive-behavioural approach, control); KBIA¹⁰⁴ (brief individualised programme, usual care); VKSC2¹⁰⁵ (self-care, usual care).

Subgroups identified by the 'individual patient data subgroup identification based on a differential effect search' method

The application of the IPD-SIDES method for the second set of analyses found candidate subgroups for one or more short-term outcome measures for the passive physical against non-active usual care (*Table 21*), psychological against non-active usual care (*Table 22*) and sham against non-active usual care (*Table 23*). No candidate subgroups were identified for the active physical against non-active usual care comparison.

TABLE 21 Candidate subgroups identified by the IPD-SIDES method for the passive physical vs. non-active usual care comparison^{a,b}

Subgroups	<i>n</i>	Treatment effect (95% CI)	Interaction effect	Unadjusted <i>p</i> -value
Outcome: short-term FFbHR				
Overall treatment effect (95% CI): 9.95 (8.80 to 11.11)				
Candidate 1				
FFbHR score of ≤ 54.2	1424	12.86 (10.81 to 14.91)	5.45	<0.001
FFbHR score of > 54.2	1848	7.41 (6.23 to 8.59)		
Candidate 2				
FFbHR score of ≤ 54.2 and age ≤ 57 years	731	15.86 (12.80 to 18.92)	6.63	0.002
FFbHR score of ≤ 54.2 and age > 57 years	693	9.23 (6.64 to 11.82)		
Candidate 3				
FFbHR score of ≤ 54.2 and age ≤ 53 years	571	16.67 (13.16 to 20.18)	6.85	0.001
FFbHR score of ≤ 54.2 and age > 53 years	853	9.83 (7.43 to 12.22)		
Candidate 4				
FFbHR score of ≤ 41.7	792	15.03 (12.06 to 18.01)	6.71	<0.001
FFbHR score of > 41.7	2480	8.32 (7.19 to 9.45)		
Outcome: short-term MCS				
Overall treatment effect (95% CI): 2.96 (2.31 to 3.61)				
Candidate 1				
MCS of ≤ 54.3	2714	3.76 (2.97 to 4.55)	2.82	<0.001
MCS of > 54.3	1165	0.93 (0.10 to 1.76)		
Candidate 2				
MCS of ≤ 54.3 and PCS ≤ 43.9	2171	4.27 (3.39 to 5.15)	2.43	0.019
MCS of ≤ 54.3 and PCS > 43.9	543	1.85 (0.11 to 3.59)		
Candidate 3				
MCS of ≤ 51.3	2327	3.83 (2.96 to 4.70)	2.57	<0.001
MCS of > 51.3	1552	1.26 (0.52 to 1.99)		

continued

TABLE 21 Candidate subgroups identified by the IPD-SIDES method for the passive physical vs. non-active usual care comparison^{a,b} (*continued*)

Subgroups	<i>n</i>	Treatment effect (95% CI)	Interaction effect	Unadjusted <i>p</i> -value
Outcome: short-term PCS				
Overall treatment effect (95% CI): 4.10 (3.56 to 4.63)				
Candidate 1				
PCS of ≤ 43.6	3103	4.39 (3.78 to 4.99)	1.61	0.013
PCS of > 43.6	776	2.77 (1.87 to 3.67)		
Candidate 2				
PCS of ≤ 43.6 and age ≤ 44 years	942	5.35 (4.21 to 6.49)	1.45	0.040
PCS of ≤ 43.6 and age > 44 years	2161	3.90 (3.20 to 4.60)		
Candidate 3				
PCS of ≤ 37.8	2326	4.61 (3.90 to 5.32)	1.23	0.025
PCS of > 37.8	1553	3.37 (2.66 to 4.09)		
Candidate 4				
PCS of ≤ 37.8 and age ≤ 62 years	1682	5.08 (4.21 to 5.94)	1.97	0.016
PCS of ≤ 37.8 and age > 62 years	644	3.11 (1.94 to 4.28)		
Candidate 5				
PCS of ≤ 37.8 and MCS > 44.0	1396	5.48 (4.55 to 6.41)	1.80	0.011
PCS of ≤ 37.8 and MCS ≤ 44.0	930	3.68 (2.64 to 4.71)		
Candidate 6				
PCS of ≤ 37.8 and MCS > 51.8	932	5.77 (4.63 to 6.91)	1.78	0.012
PCS of ≤ 37.8 and MCS ≤ 51.8	1394	3.99 (3.11 to 4.87)		
Candidate 7				
PCS of ≤ 37.8 and MCS > 51.8 and sex = female	520	6.64 (5.12 to 8.16)	1.73	0.167
PCS of ≤ 37.8 and MCS > 51.8 and sex = male	412	4.91 (3.17 to 6.65)		
Candidate 8				
PCS of ≤ 40.3	2715	4.51 (3.85 to 5.16)	1.61	0.006
PCS of > 40.3	1164	2.90 (2.11 to 3.68)		
Candidate 9				
PCS of ≤ 40.3 and MCS > 51.5	1086	5.43 (4.37 to 6.48)	1.38	0.042
PCS of ≤ 40.3 and MCS ≤ 51.5	1629	4.05 (3.24 to 4.85)		
a The baseline FFbHR score ranges from 0 to 100, for which a lower score represents greater disability. The baseline MCSs and PCSs range from 0 to 100, for which a lower score represents worse mental and physical functioning.				
b The first row of each candidate subgroup is the selected subgroup with enhanced treatment effect.				

TABLE 22 Candidate subgroups identified by the IPD-SIDES method for the psychological vs. non-active usual care comparison^{a,b}

Subgroups	n	Treatment effect (95% CI)	Interaction effect	Unadjusted p-value
Outcome: short-term RMDQ				
Overall treatment effect (95% CI): 1.40 (0.89 to 1.91)				
Candidate 1				
RMDQ score of > 4	697	1.72 (1.12 to 2.31)	1.07	0.038
RMDQ score of ≤ 4	231	0.65 (−0.11 to 1.40)		
a The baseline RMDQ score ranges from 0 to 24, for which a higher score represents greater disability.				
b The first row of each candidate subgroup is the selected subgroup with enhanced treatment effect.				

TABLE 23 Candidate subgroups identified by the IPD-SIDES method for the sham vs. non-active usual care comparison^{a,b}

Subgroups	n	Treatment effect (95% CI)	Interaction effect	Unadjusted p-value
Outcome: short-term MCS				
Overall treatment effect (95% CI): 2.59 (1.13 to 4.04)				
Candidate 1				
Age ≤ 65 years	705	3.42 (1.80 to 5.04)	4.32	0.019
Age > 65 years	174	−0.90 (−4.16 to 2.35)		
Candidate 2				
PCS of ≤ 42.0	791	3.10 (1.55 to 4.65)	4.99	0.043
PCS of > 42.0	88	−1.89 (−6.07 to 2.28)		
a The baseline PCS ranges from 0 to 100, for which a lower score represents worse physical functioning.				
b The first row of each candidate subgroup is the selected subgroup with enhanced treatment effect.				

Passive physical results compared with non-active usual care results

Short-term Hannover Functional Ability Questionnaire for measuring back pain-related functional limitations outcome

The overall treatment effect for the FFbHR short-term outcome was 9.95 (95% CI 8.80 to 11.11). Four candidate subgroups were identified for the FFbHR short-term outcome. Those with a baseline FFbHR score of ≤ 54.2 had a treatment effect of 12.86 (95% CI 10.81 to 14.91), those with a baseline FFbHR score of ≤ 54.2 and age ≤ 57 years had a treatment effect of 15.86 (95% CI 12.80 to 18.92), those with a FFbHR score of ≤ 54.2 and age ≤ 53 years had a treatment effect of 16.67 (95% CI 13.16 to 20.18) and those with a baseline FFbHR score of ≤ 41.7 had a treatment effect of 15.03 (95% CI 12.06 to 18.01).

- Overall, those with more disability and who are younger are likely to gain a greater benefit on disability from passive physical treatments.

Short-term mental component score of SF-12/36 outcome

The overall treatment effect for the SF-12/36 MCS short-term outcome was 2.96 (95% CI 2.31 to 3.61). Three candidate subgroups were identified for the MCS short-term outcome. Those with a baseline MCS of ≤ 54.3 had a treatment effect of 3.76 (95% CI 2.97 to 4.55), those with a MCS of ≤ 54.3 and PCS of ≤ 43.9 had a treatment effect of 4.27 (95% CI 3.39 to 5.15) and those with a MCS of ≤ 51.3 had a treatment effect of 3.83 (95% CI 2.96 to 4.70).

- These results suggest that those with more psychological distress and worse physical status at baseline will get better outcomes on psychological distress from passive physical treatments.

Short-term physical component score of SF-12/36 outcome

The overall treatment effect for the SF-12/36 PCS short-term outcome was 4.10 (95% CI 3.56 to 4.63). Nine candidate subgroups were identified for the PCS short-term outcome. Those with a baseline PCS of ≤ 43.6 had a treatment effect of 4.39 (95% CI 3.78 to 4.99), those with a baseline PCS of ≤ 43.6 and age ≤ 44 years had a treatment effect of 5.35 (95% CI 4.21 to 6.49), those with a baseline PCS of ≤ 37.8 had a treatment effect of 4.61 (95% CI 3.90 to 5.32), those with a PCS of ≤ 37.8 and age ≤ 62 years had a treatment effect of 5.08 (95% CI 4.21 to 5.94), those with a baseline PCS of ≤ 37.8 and a MCS of > 44.0 had a treatment effect of 5.48 (95% CI 4.55 to 6.41), those with a PCS of ≤ 37.8 and a MCS of > 51.8 had a treatment effect of 5.77 (95% CI 4.63 to 6.91), those with a PCS of ≤ 37.8 , a MCS of > 51.8 and female had a treatment effect of 6.64 (95% CI 5.12 to 8.16), those with PCS ≤ 40.3 had a treatment effect of 4.51 (95% CI 3.85 to 5.16) and, finally, those with a PCS of ≤ 40.3 and a MCS of > 51.5 had a treatment effect of 5.43 (95% CI 4.37 to 6.48). Broadly speaking, these results suggest that:

- Younger patients with worse physical status at baseline will get better outcomes on physical status from passive physical treatments.
- Those with worse physical status but less psychological distress at baseline will get better outcomes on physical status from passive physical treatments.
- Females with worse physical status and less psychological distress at baseline will get better outcomes on physical status from passive physical treatments.

Psychological results compared with non-active usual care results**Short-term Roland–Morris Disability Questionnaire outcome**

The overall treatment effect for the RMDQ short-term outcome was 1.40 (95% CI 0.89 to 1.91). One candidate subgroup was identified for the RMDQ short-term outcome. Those with a baseline RMDQ score of > 4 had a treatment effect of 1.72 (95% CI 1.12 to 2.31).

- This suggests that those with worse disability at baseline gain more benefit from psychological treatment on disability than usual care control.

Sham results compared with non-active usual care results**Short-term mental component score of SF-12/36 outcome**

Two trials were included in the analyses and the sham treatment in both was sham acupuncture. The overall treatment effect for the MCS short-term outcome was 2.59 (95% CI 1.13 to 4.04). Two candidate subgroups were identified for the MCS short-term outcome. Those with age ≤ 65 years at baseline had a treatment effect of 3.42 (95% CI 1.80 to 5.04) and those with a baseline PCS of ≤ 42.0 had a treatment effect of 3.10 (95% CI 1.55 to 4.65). No candidate subgroups were identified for the FFbHR and PCS short-term outcomes.

- This suggests that younger people and those with worse physical status at baseline have a greater benefit from sham treatment on psychological distress than usual care control.

Chapter 8 Methodology and statistical developments 2: subgroup identification using an adaptive refinement by directed peeling algorithm

Background

The adaptive risk group refinement introduced by LeBlanc *et al.*¹⁷² aims to identify subgroups of participants with poor prognosis, whereby the subgroups are defined by cut-offs for the covariates resulting in box-shaped subgroups that are easy to interpret. The approach is based on a so-called 'adaptive refinement by directed peeling' (ARDP) algorithm. Starting with the whole data set, the algorithms peel off fractions of the data in a series of locally optimal steps optimising a prognostic indicator (e.g. median survival in the paper by LeBlanc *et al.*¹⁷²). We aim to identify subgroups of participants who benefit in particular from a specific treatment in that they respond particularly well to the treatment. The approach to subgroup identification presented in this chapter builds on the work by LeBlanc *et al.*¹⁷² and extends it in two ways: (1) the criterion for optimisation is now based on the interaction effects between treatment and subgroup, and (2) data from multiple trials can now be analysed, allowing between-trial heterogeneity in the treatment-by-subgroup interactions thereby generalising the ARDP algorithm from a single-study setting to individual participant data meta-analysis setting. With regard to the latter, this is similar to the IT and SIDES methods (see *Chapter 7*). In the following sections we describe the modified ARDP algorithm for individual participant data meta-analysis.

Adaptive refinement by directed peeling in individual patient data meta-analysis

The ARDP in individual patient data meta-analysis (ARDP-MA) algorithm to construct a region that predicts the best or worst response to treatment consists of the following steps:

1. To determine the covariates to be included and their direction of peeling, run regression analyses on the entire data set to investigate interactions of covariates with treatment. For the identified moderators, the sign of the interaction effect determines the direction of peeling. If larger values of a covariate lead to larger treatment effects then peel off the cases with a smaller value of this covariate. Correspondingly, if smaller values of the covariate lead to larger treatment effects then peel off the larger values of the covariate.
2. Start with a 'subgroup' B^0 that includes all observations, n .
3. The proportion of data to be removed in one step is denoted by α and the minimum number of observations to be peeled off is denoted by n_{min} . For each variable, we move the threshold so that $\max(\alpha n, n_{min})$ observations are removed; the resulting subgroups for the L covariates we denote by B_j^m , $j = 1, \dots, L$. For each subgroup B_j^m calculate the treatment-by-subgroup interaction effect and select the B_j^m , which gives the largest improvement on the interaction effect in comparison with the previous iteration standardised by change in subgroup size. In the setting of data from multiple trials, the interaction effects estimated from the individual trials are combined in a random-effects meta-analysis (two-stage procedure); alternatively an equivalent hierarchical model can be fitted (one-step procedure).
4. The selected subgroup is then called B^{m+1} .
5. Estimate the treatment effects for the outcome of interest for subgroup B^{m+1} .
6. Repeat steps 3–5 until the size of the remaining region is not smaller than r .

Figure 24 illustrates the ARDP algorithm for the identification of subgroups of treatment responders. Expecting a large number of covariates to be included in the analyses, we developed this algorithm earlier on in the project. However, it turned out that situations with a small number of covariates were most relevant for the data sets to be analysed. By restricting the number of covariates to four, we could do far more extensive searches by considering all of the possible combinations of boxes described in the ARDP algorithm above. This allowed us to interrogate the data sets more thoroughly.

Note that this algorithm can be applied to various kinds of end points, as we assume that only appropriate regression models can be fitted to model the outcome. For instance, Gaussian linear models could be applied to continuous outcomes, logistic regression to binomial outcomes, and Cox's proportional hazard models to time-to-event data. No distributional assumption regarding the covariates is required, but they should be ordinal and have a sufficient number of possible outcomes so that the peeling in several steps makes sense. If a covariate is not ordinal then an order could be imposed on it by ordering the outcomes by the regression coefficients estimated in step 1 of the algorithm.¹⁷²

Analyses

The minimum sample size of the subpopulation was defined as $r = 0.10$ of the population analysed. The appeal of the ARDP-MA method is the ability to remove a small proportion of participants at each iteration. Categorical covariates that delineate participants into three or fewer categories would cause the ARDP-MA method to remove a large proportion of participants, an unappealing feature. As all the categorical covariates identified in the analyses of covariance have three or fewer categories, none of them was considered in the ARDP-MA analyses.

Similar to analyses seen in *Chapter 7* (see *Analyses*), two sets of analyses were performed. The first one was to confirm proof of concept, when all interventions (active physical, passive physical and psychological, delivered singly or in combination with the others) were grouped together as being one arm and the non-active usual care grouped with the sham as a control/placebo arm. Analyses were performed for these measurements: average pain, EQ-5D, FFbHR, MCS of SF-12/36, PCS of SF-12/36 and RMDQ score. The outcome was the absolute change from baseline to short-term follow-up. In the second set of analyses, similarly, two treatments are compared and the pairwise comparisons investigated were active physical against non-active usual care, passive physical against non-active usual care, psychological against non-active usual care, and sham against non-active usual care.

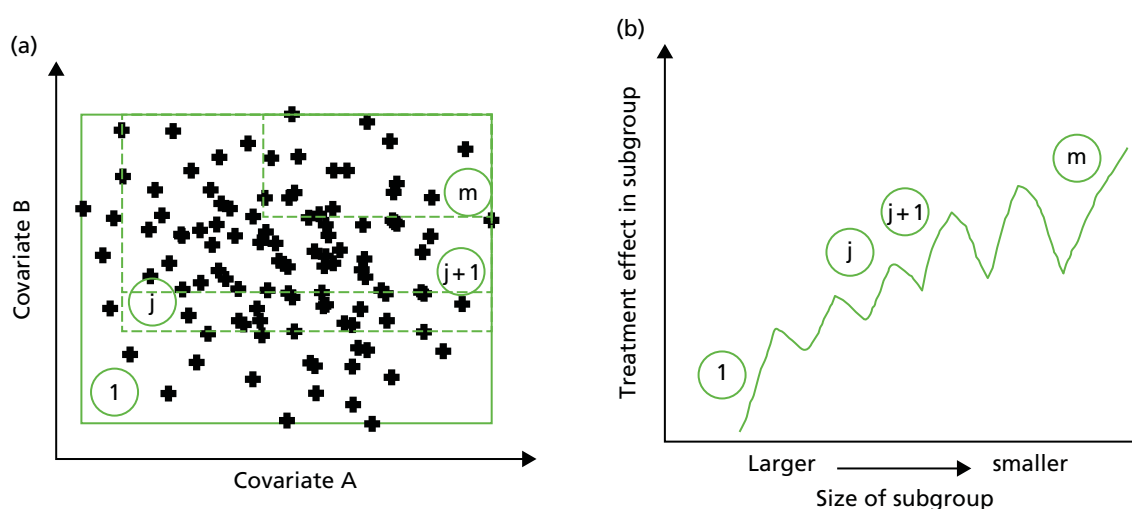


FIGURE 24 Schematic diagram of the ARDP algorithm to identify subgroups of treatment responders. Here the subgroups are defined by thresholds for the two covariates A and B. (a) Shows how two covariates might affect response to treatment. In the upper right hand corner is the group, defined by those two covariates, who have the largest response to treatment; and (b) shows how size of the treatment effect might increase as the size of the subgroup decreases as the peeling processes from bottom left to top right.

Results

We programmed the ARDP-MA method to do a full search, but this limits the number of covariates. As the number of covariates increased, the computational time and resources needed to store the data increased exponentially, causing a massive strain on the system server. Therefore, up to four covariates when necessary were included in the analyses.

Analysis 1: overall comparison treatment compared with control

Table 24 shows the summary of the trials and continuous variables used in the ARDP-MA algorithm to construct a region that predicts the best or worst response for each of the short-term outcome measures.

Short-term average pain outcome

Figure 25 shows the trajectory plot for the treatment effect for the short-term outcome of average pain. The treatment effect increased as more and more participants were excluded from the subgroup. However, Table 25 shows that age and average pain might not be important covariates in improving the treatment effect as their thresholds fluctuate. Of note was that substantial physical limitation (low PCS) seemed to gain benefit in short-term average pain.

TABLE 24 Summary of included trials and variables considered to construct a region that predicts the best or worst response to treatment^a

Outcome ^b	Trials	Variables
Average pain	$m = 3$; $n = 2534$ UK BEAM ³¹ ($n = 926$), BeST ³³ ($n = 498$), Haake ($n = 1110$)	Age, average pain, PCS and MCS at baseline
EQ-5D	$m = 2$; $n = 1365$ UK BEAM ³¹ ($n = 890$), BeST ³³ ($n = 475$)	RMDQ, average pain, PCS and MCS at baseline
FFbHR	$m = 3$; $n = 3718$ Brinkhaus ¹⁰¹ ($n = 284$), Haake ¹³² ($n = 1110$), Witt ⁵⁰ ($n = 2324$)	Age, FFbHR, PCS and MCS at baseline
MCS ^c	$m = 3$; $n = 3,630$ Brinkhaus ¹⁰¹ ($n = 281$), Haake ¹³² ($n = 1110$), Witt ⁵⁰ ($n = 2239$)	Age, FFbHR, PCS and MCS at baseline
PCS ^d	$m = 6$; $n = 5208$ UK BEAM ³¹ ($n = 893$), BeST ³³ ($n = 470$), Brinkhaus ¹⁰¹ ($n = 281$), Haake ¹³² ($n = 1110$), Witt ⁵⁰ ($n = 2248$), YACBAC ¹⁰⁷ ($n = 206$)	Age, PCS and MCS at baseline
RMDQ	$m = 8$; $n = 2675$ UK BEAM ³¹ ($n = 995$), BeST ³³ ($n = 514$), Hancock ¹³¹ ($n = 235$), Kennedy ¹³⁶ ($n = 40$), Pengel ¹⁰³ ($n = 236$), Smeets ⁷⁰ ($n = 212$), VKBIA ¹⁰⁴ ($n = 230$), VKSC2 ¹⁰⁵ ($n = 213$)	Age and RMDQ score at baseline

YACBAC, York Acupuncture Back Pain Trial.

a Any active intervention (active physical, passive physical or psychological delivered either singly or in combination with other intervention) against control/placebo, which is either GP usual care or sham.

b Change from baseline to short-term follow-up (between 2 and 3 months post randomisation or entry to the trial).

c MCS of SF-12/36.

d PCS of SF-12/36.

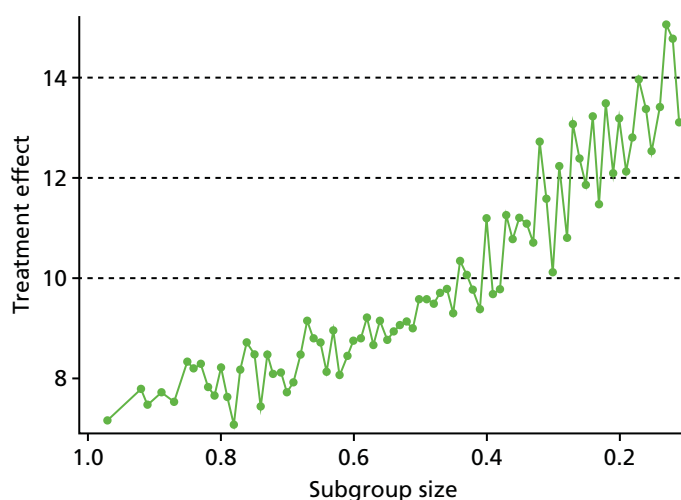


FIGURE 25 Trajectory plot for the treatment effect against the size of the constructed region for the average pain short-term outcome.

TABLE 25 Thresholds for selected size of subgroup for the short-term average pain as seen in *Figure 25*

Subgroup size	Age, years (<)	Pain (>)	PCS ^a (<)	MCS ^b (>)	Treatment effect
0.106 ^c	50	50	33.62	38.21	14.04
0.206	67	50	31.34	28.93	13.18
0.217	67	40	31.34	28.93	12.10
0.227	67	0	31.34	28.93	13.48
0.238	62	50	33.62	28.93	11.49
0.247	91	50	31.34	28.93	13.22
0.255	91	0	31.34	34.18	11.86
0.262	91	40	31.34	28.93	12.38
0.275	91	0	31.34	28.93	13.08
0.285	91	40	31.34	9.46	10.81
0.300	91	0	31.34	9.46	12.23
0.307	67	30	33.62	28.93	10.11
0.402	91	50	35.66	28.93	11.20
0.414	67	50	47.59	38.21	9.39
0.426	91	20	40.45	42.95	9.77
0.434	67	20	43.62	42.95	10.07
0.442	67	0	43.62	42.95	10.34
0.459	91	30	35.66	28.93	9.30
0.501	91	0	43.62	42.95	9.58
0.600	91	0	40.45	34.18	8.76
0.710	67	40	47.59	9.46	7.72
0.804	91	30	47.59	28.93	8.23

a PCS of SF-12/36.

b MCS of SF-12/36.

c For about 10.6% of the population with age < 50 years, average pain score of > 50, SF-12/36 PCS of < 33.62 and SF-12/36 MCS of > 38.21, the treatment effect was 14.04.

Short-term European Quality of Life-5 Dimensions outcome

Figure 26 shows the trajectory plot for the short-term outcome of health utility measured by the EQ-5D. As seen in Table 26, approximately 90% of the initial 1365 participants (corresponding to PCS of < 68 and MCS of < 60, regardless of the average pain and RMDQ score at baseline) had an average treatment effect of 0.073. The treatment effect increased sharply to 0.100 after approximately 30% of the participants were excluded in the model. From then on the treatment effect was quite 'stable' despite a further 40% of participants being excluded from the analysis. There was a marked increase in treatment effect for about 20% of the population (corresponding to PCS of < 31, MCS of < 72, average pain > 0 and RMDQ score of > 6), for whom the average treatment effect was about 0.160.

Short-term Hannover Functional Ability Questionnaire outcome

Figure 27 shows the trajectory plot for the treatment effect against the size of the constructed region for the change of FFbHR score between baseline and short-term follow-up. In the first iteration, approximately 10% of the initial 3718 participants were excluded from the subgroup box and these participants had a high value of PCS at baseline, that is, the remaining 90% in the subgroup correspond to any age, FFbHR score of < 100, PCS of < 48 and MCS of < 72. The average treatment effect was 8.5 (Table 27). The average treatment effect increased as more participants were excluded from the subgroup box. The average treatment effect for the last 10% of the participants (corresponding to any age, FFbHR score of < 29, PCS of < 68 and MCS of < 57) was 16.8. Although an increase of 8 units of the FFbHR score may be of clinical importance, the proportion of participants who would benefit from such improvement is very small. Nevertheless, those with more functional limitation (greater disability) and more psychological distress would benefit more on the FFbHR disability outcome at short term. If we were interested in an improvement from an average of 8.5 to at least 12 then approximately 30% of the participants (age < 67 years, FFbHR score of < 54, PCS of < 40 and MCS of < 72) would benefit more on the disability outcome at short term, a similar result to that observed in the IPD-SIDES Analysis 1, for which participants with FFbHR score of ≤ 54.2 and age ≤ 66 years had an enhanced treatment effect (see Chapter 7, *Subgroups identified by the individual patient data subgroup identification based on a differential effect search method*). It is of note that results from both methods suggest that MCS may not be an essential covariate in improving treatment effect.

- Those with more functional limitation at baseline and who were younger would gain greater improvement in short-term functional ability as measured by the FFbHR.

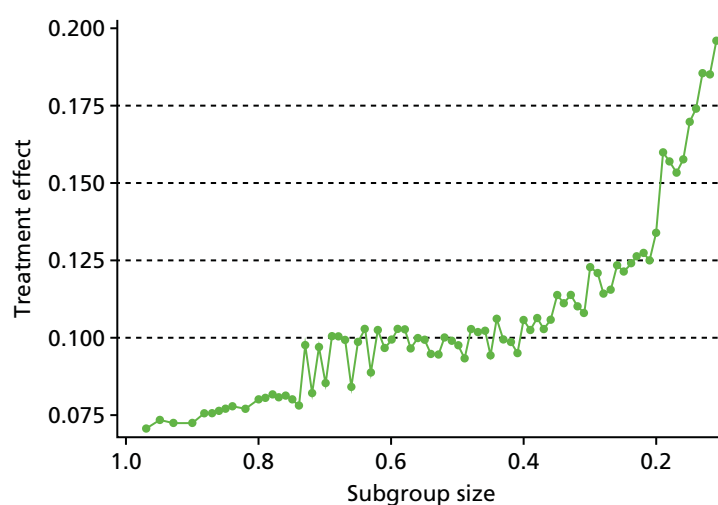


FIGURE 26 Trajectory plot for the treatment effect against the size of the constructed region for the EQ-5D short-term outcome.

TABLE 26 Thresholds for selected size of subgroup for the short-term EQ-5D as seen in *Figure 26*

Subgroup size	PCS ^a (<)	MCS ^b (<)	Pain (>)	RMDQ (>)	Treatment effect
0.101 ^c	35.66	60.35	0.00	14	0.208
0.119	38.01	60.35	0.00	14	0.196
0.127	38.01	72.11	0.00	14	0.185
0.136	47.59	60.35	0.00	14	0.185
0.144	47.59	72.11	0.00	14	0.174
0.151	31.34	56.82	0.00	0	0.170
0.166	31.34	60.35	0.00	6	0.158
0.171	31.34	60.35	0.00	0	0.153
0.188	31.34	72.11	20.00	6	0.157
0.190	31.34	72.11	0.00	6	0.160
0.210	33.62	56.82	0.00	6	0.134
0.219	40.45	47.17	20.00	10	0.125
0.221	40.45	47.17	0.00	10	0.127
0.233	33.62	60.35	0.00	6	0.126
0.244	38.01	47.17	0.00	6	0.124
0.259	33.62	72.11	30.00	6	0.122
0.267	33.62	72.11	0.00	6	0.124
0.303	40.45	47.17	0.00	6	0.123
0.407	67.75	72.11	57.00	0	0.106
0.415	43.62	50.61	20.00	6	0.095
0.429	40.45	56.82	30.00	6	0.099
0.437	38.01	72.11	20.00	6	0.099
0.446	40.45	56.82	0.00	6	0.106
0.451	47.59	50.61	30.00	6	0.094
0.464	47.59	72.11	50.00	6	0.102
0.477	40.45	60.35	20.00	6	0.102
0.482	40.45	60.35	0.00	6	0.103
0.498	43.62	56.82	30.00	6	0.093
0.505	40.45	72.11	30.00	6	0.098
0.512	47.59	56.82	20.00	7	0.099
0.530	40.45	72.11	0.00	6	0.100
0.540	47.59	53.87	0.00	6	0.095
0.541	67.75	60.35	40.00	6	0.095
0.552	47.59	56.82	30.00	6	0.099
0.570	43.62	60.35	0.00	6	0.100
0.574	67.75	56.82	30.00	6	0.097
0.581	47.59	56.82	20.00	6	0.102
0.593	47.59	56.82	0.00	6	0.103

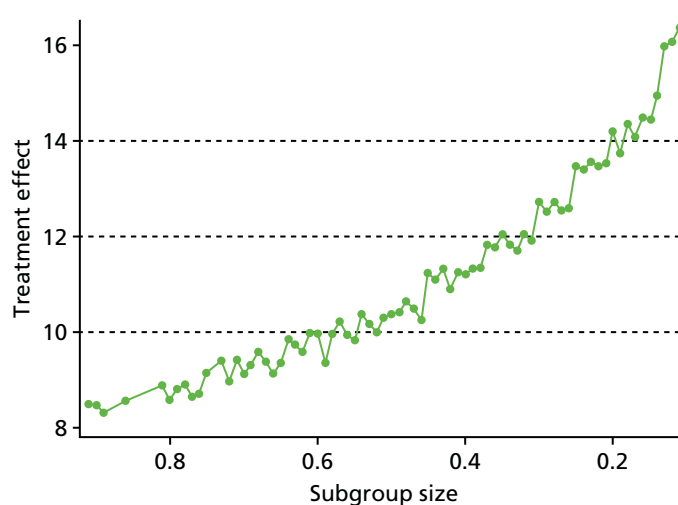
TABLE 26 Thresholds for selected size of subgroup for the short-term EQ-5D as seen in *Figure 26* (continued)

Subgroup size	PCS ^a (<)	MCS ^b (<)	Pain (>)	RMDQ (>)	Treatment effect
0.610	67.75	56.82	20.00	6	0.099
0.704	47.59	60.35	20.00	5	0.085
0.803	47.59	60.35	0.00	0	0.080
0.909	67.75	60.35	0.00	0	0.073

a PCS of SF-12/36.

b MCS of SF-12/36.

c For about 10.1% of the population with SF-12/36 PCS of <35.66, SF-12/36 MCS of <60.35, average pain >0 and RMDQ score of >14, the treatment effect was 0.208.

**FIGURE 27** Trajectory plot for the treatment effect against the size of the constructed region for the FFbHR short-term outcome.**TABLE 27** Thresholds for selected size of subgroup for the short-term FFbHR outcome as seen in *Figure 27*

Subgroup size	Age, years (<)	FFbHR (<)	PCS ^a (<)	MCS ^b (<)	Treatment effect
0.102 ^c	91	29.17	67.75	56.82	16.79
0.118	54	58.33	40.45	47.17	16.35
0.121	54	45.83	67.75	60.35	16.07
0.132	54	45.83	67.75	72.11	15.97
0.150	54	62.50	33.62	72.11	14.92
0.155	54	54.17	40.45	56.82	14.43
0.163	58	45.83	40.45	72.11	14.49
0.171	54	54.17	40.45	60.35	14.06
0.190	54	54.17	40.45	72.11	14.35
0.200	54	54.17	43.62	72.11	13.74
0.206	54	54.17	67.75	72.11	14.18

continued

TABLE 27 Thresholds for selected size of subgroup for the short-term FFbHR outcome as seen in Figure 27 (continued)

Subgroup size	Age, years (<)	FFbHR (<)	PCS ^a (<)	MCS ^b (<)	Treatment effect
0.308	62	54.17	67.75	72.11	12.72
0.314	67	54.17	40.45	60.35	11.90
0.327	62	58.33	40.45	72.11	12.05
0.340	67	54.17	67.75	60.35	11.70
0.345	58	62.50	67.75	72.11	11.82
0.352	67	54.17	40.45	72.11	12.03
0.361	62	58.33	67.75	72.11	11.76
0.378	67	54.17	67.75	72.11	11.82
0.385	91	54.17	40.45	60.35	11.33
0.400	62	70.83	40.45	60.35	11.32
0.402	67	58.33	40.45	72.11	11.20
0.509	67	62.50	67.75	72.11	10.36
0.513	62	100.00	40.45	72.11	10.30
0.528	91	75.00	40.45	56.82	9.99
0.535	91	58.33	67.75	72.11	10.16
0.548	91	62.50	40.45	72.11	10.37
0.553	91	83.33	40.45	56.82	9.82
0.570	67	75.00	40.45	72.11	9.95
0.573	91	70.83	40.45	60.35	10.22
0.582	91	62.50	43.62	72.11	9.96
0.599	67	75.00	47.59	60.35	9.37
0.602	91	75.00	40.45	60.35	9.96
0.702	91	75.00	47.59	60.35	9.14
0.808	91	100.00	47.59	60.35	8.59
0.906	91	100.00	47.59	72.11	8.47

a PCS of SF-12/36.

b MCS of SF-12/36.

c For about 10.2% of the population with age <91 years, FFbHR score of <29.17, SF-12/36 PCS of <67.75 and SF-12/36 MCS of <56.82, the treatment effect was 16.79.

Short-term Short Form questionnaire-12 items/-36 items mental component score outcome

Figure 28 is the trajectory plot for the treatment effect for the short-term outcome of MCS. Table 28 shows a selection of constructed regions and the corresponding thresholds for covariates age, FFbHR score, PCS and MCS. The average treatment effect of approximately 90% of the initial 3630 participants (corresponding with age > 16 years, FFbHR score of < 100, PCS of < 48 and MCS of < 72) was 2.23, and this increased to 5.98 for approximately 10% of the participants (corresponding to age > 16 years, FFbHR score of < 100, PCS of < 29 and MCS of < 51). Approximately 55% of the participants (corresponding to age > 31 years, FFbHR score of < 63, PCS of < 44 and MCS of < 72) had an average treatment effect of 3 units. A smaller region consisting of 30% of the participants (corresponding to age > 54 years, FFbHR score of < 75, PCS of < 44 and MCS of < 57) would gain greater improvement in psychological outcome, that is, an average treatment effect of 4 units. Of interest is the conflicting cut-off suggested by FFbHR and PCS at baseline in constructing these regions, for which the former seemed not to play a critical role and the latter suggested that those with poor physical status would gain greater improvement.

- Those with more psychological distress and who were younger would gain greater improvement in the short-term psychological outcome as measured by the SF-12/36 MCS.

Short-term Short Form questionnaire-12 items/-36 items physical component score outcome

Figure 29 shows the trajectory plot for the treatment effect for the short-term outcome of PCS. Although it shows a general trend of higher treatment effect as subgroups were removed from the initial pool of 5208 participants, the treatment effect increased but was not monotonic and the improvement did not increase very much to warrant a clinical importance. Table 29 shows a selection of constructed regions and the corresponding thresholds for covariates age, PCS and MCS. We thus conclude that there was also no subgroup who would gain benefit in short-term SF-12/36 PCS.

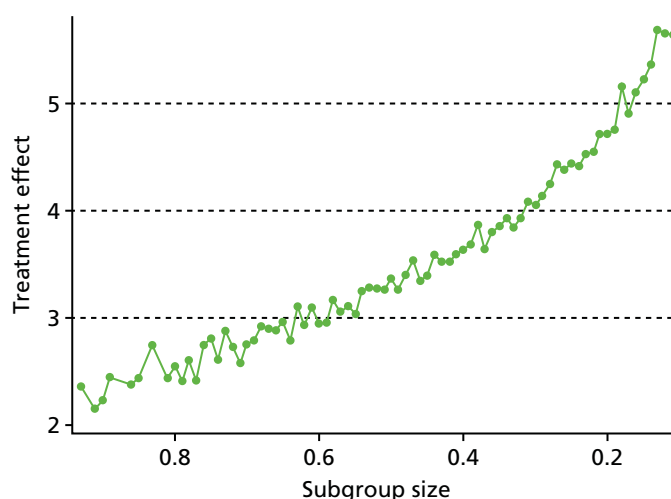


FIGURE 28 Trajectory plot for the treatment effect against the size of the constructed region for the SF-12/36 MCS short-term outcome.

TABLE 28 Thresholds for selected size of subgroup for the short-term SF-12/36 MCS as seen in *Figure 28*

Subgroup size	Age, years (>)	FFbHR (<)	PCS ^a (<)	MCS ^b (<)	Treatment effect
0.108 ^c	16	100.00	28.84	50.61	5.98
0.159	58	75.00	35.66	53.87	5.23
0.163	58	83.33	35.66	53.87	5.11
0.176	58	70.83	38.01	53.87	4.90
0.181	58	75.00	38.01	53.87	5.16
0.194	31	75.00	31.34	53.87	4.76
0.207	31	45.83	43.62	50.61	4.72
0.301	54	75.00	43.62	56.82	4.05
0.317	31	54.17	47.59	53.87	4.08
0.328	31	54.17	40.45	56.82	3.93
0.334	45	62.50	38.01	60.35	3.84
0.341	31	54.17	43.62	56.82	3.93
0.351	31	54.17	67.75	56.82	3.86
0.365	45	62.50	40.45	60.35	3.81
0.373	31	70.83	38.01	53.87	3.64
0.384	45	62.50	43.62	60.35	3.86
0.401	45	62.50	67.75	60.35	3.64
0.505	31	75.00	38.01	60.35	3.37
0.515	45	75.00	67.75	60.35	3.27
0.526	31	83.33	38.01	60.35	3.28
0.535	31	100.00	38.01	60.35	3.29
0.541	31	100.00	67.75	50.61	3.25
0.551	31	62.50	43.62	72.11	3.03
0.568	37	75.00	43.62	60.35	3.10
0.577	31	100.00	47.59	53.87	3.05
0.582	31	70.83	43.62	60.35	3.17
0.597	31	100.00	43.62	56.82	2.96
0.604	45	100.00	67.75	60.35	2.94
0.701	16	75.00	47.59	60.35	2.75
0.807	16	100.00	47.59	60.35	2.55
0.907	16	100.00	47.59	72.11	2.23

a PCS of SF-12/36.

b MCS of SF-12/36.

c For about 10.8% of the population with age > 16 years, FFbHR score of < 100, SF-12/36 PCS of < 28.84 and SF-12/36 MCS of < 50.61, the treatment effect was 5.98.

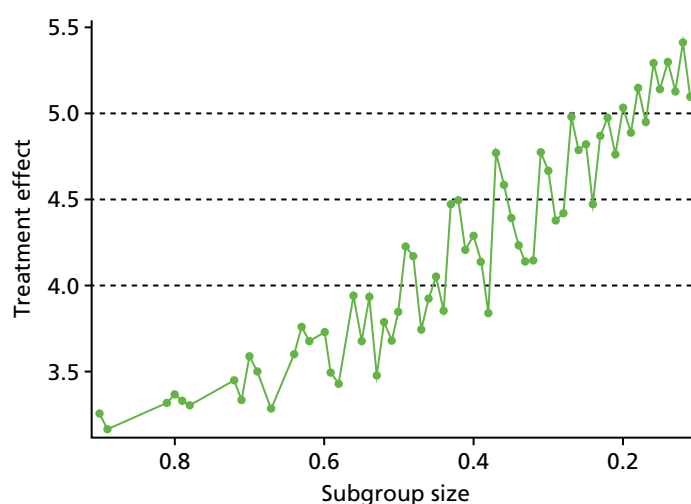


FIGURE 29 Trajectory plot for the treatment effect against the size of the constructed region for the SF-12/36 PCS short-term outcome.

TABLE 29 Thresholds for selected size of subgroup for the short-term SF-12/36 PCS as seen in *Figure 29*

Subgroup size	Age, years (<)	PCS ^a (<)	MCS ^b (>)	Treatment effect
0.110 ^c	54	40.45	56.82	5.30
0.153	54	35.66	47.17	5.14
0.169	67	31.34	47.17	5.29
0.176	91	31.34	50.61	4.95
0.189	67	40.45	56.82	5.15
0.193	67	33.62	50.61	4.89
0.202	91	31.34	47.17	5.03
0.211	58	35.66	42.95	4.76
0.224	62	35.66	47.17	4.98
0.233	67	35.66	50.61	4.87
0.245	62	43.62	53.87	4.47
0.253	67	40.45	53.87	4.82
0.263	58	40.45	47.17	4.79
0.270	67	35.66	47.17	4.98
0.289	67	43.62	53.87	4.42
0.292	91	40.45	53.87	4.38
0.307	62	43.62	50.61	4.67
0.316	67	40.45	50.61	4.78
0.326	67	47.59	53.87	4.15
0.334	54	40.45	34.18	4.14
0.348	62	47.59	50.61	4.23
0.360	58	43.62	42.95	4.39

continued

TABLE 29 Thresholds for selected size of subgroup for the short-term SF-12/36 PCS as seen in *Figure 29 (continued)*

Subgroup size	Age, years (<)	PCS ^a (<)	MCS ^b (>)	Treatment effect
0.366	62	40.45	42.95	4.58
0.372	67	40.45	47.17	4.77
0.385	67	35.66	34.18	3.85
0.391	62	67.75	50.61	4.14
0.409	91	43.62	50.61	4.29
0.413	62	47.59	47.17	4.21
0.427	91	40.45	47.17	4.50
0.430	67	40.45	42.95	4.47
0.443	58	40.45	28.93	3.86
0.459	91	47.59	50.61	4.05
0.467	62	67.75	47.17	3.93
0.471	58	67.75	42.95	3.75
0.486	91	43.62	47.17	4.17
0.496	91	40.45	42.95	4.22
0.508	91	67.75	47.17	3.85
0.609	67	40.45	28.93	3.73
0.703	91	40.45	28.93	3.59
0.802	91	43.62	28.93	3.37
0.903	91	47.59	28.93	3.26

a PCS of SF-12/36.

b MCS of SF-12/36.

c For about 11.0% of the population with age < 54 years, SF-12/36 PCS of < 40.45 and SF-12/36 MCS of > 56.82, the treatment effect was 5.30.

Short-term Roland–Morris Disability Questionnaire outcome

As seen in *Figure 30*, the non-monotonic trajectory plot for the short-term outcome of RMDQ score suggested that there was no subgroup that would gain greater improvement in short-term disability outcome as measured by the RMDQ.

Table 30 shows the selection of a subgroup of participants with thresholds for covariate age and RMDQ score at baseline and their treatment effects.

Analysis 2: pairwise comparisons

Similar to the analyses seen in *Chapter 7* (see *Analysis 2: Pairwise comparisons*), a further examination of the treatment effect between active physical and non-active usual care (usual care/GP or waiting list only), between passive physical and non-active usual care, between psychological and non-active usual care, and between sham and non-active usual care arms, was performed for selected short-term outcomes. *Table 31* summarises the trials and variables considered in the construction of a region that predicts the best or worst response for each pairwise comparison for selected short-term outcome measures.

Active physical versus non-active usual care

Short-term Roland–Morris Disability Questionnaire outcome

Figure 31 shows the trajectory plot for the treatment effect between active physical and non-active usual care for the short-term RMDQ outcome. The figure shows a similar result to the one seen above (see *Analysis 1: overall comparison treatment compared with control/Short-term Roland–Morris Disability Questionnaire outcome*), that is, there was no subgroup that would have a substantial improvement in treatment effect. *Table 32* shows the average treatment effect for selected constructed regions with the corresponding thresholds.

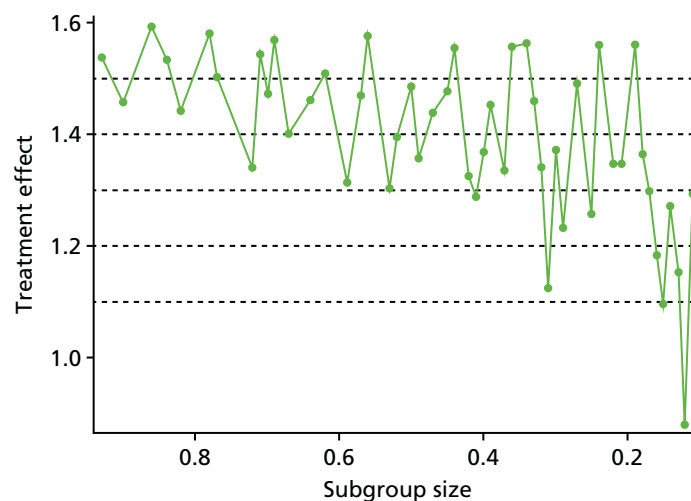


FIGURE 30 Trajectory plot for the treatment effect against the size of the constructed region for the RMDQ short-term outcome.

TABLE 30 Thresholds for selected size of subgroup for the short-term RMDQ outcome as seen in *Figure 30*

Subgroup size	Age, years (<)	RMDQ (<)	Treatment effect
0.110 ^a	45	5	1.13
0.111	41	6	1.29
0.123	31	24	0.88
0.138	37	9	1.15
0.144	45	6	1.27
0.152	37	10	1.10
0.169	54	5	1.18
0.178	45	7	1.30
0.184	50	6	1.36
0.199	37	14	1.56
0.216	37	16	1.35
0.225	50	7	1.35
0.242	37	24	1.56
0.250	58	6	1.26
0.310	50	9	1.37
0.318	91	6	1.13
0.322	45	12	1.34
0.335	41	24	1.46
0.341	62	7	1.56
0.405	50	12	1.37
0.416	54	10	1.29
0.426	58	9	1.33
0.443	45	24	1.55
0.460	50	14	1.48
0.506	50	16	1.48
0.523	62	10	1.39
0.539	91	9	1.30
0.626	54	16	1.51
0.645	58	14	1.46
0.707	58	16	1.47
0.903	91	16	1.46

^a For about 11.0% of the population with age < 45 years and RMDQ score of < 5, the treatment effect was 1.13.

TABLE 31 Summary of included trials and variables considered to construct a region that predicts the best of worst response to treatment for different direct comparisons

Outcome: comparison	FFbHR		RMDQ		MCS ^a		PCS ^b	
	Trials	Variables	Trials	Variables	Trials	Variables	Trials	Variables
Active physical vs. non-active usual care ^c			<i>m</i> = 2; <i>n</i> = 622 UK BEAM ³¹ (<i>n</i> = 465), Smeets ⁷⁰ (<i>n</i> = 157)	Age and RMDQ score at baseline				
Passive physical vs. non-active usual care ^c	<i>m</i> = 3; <i>n</i> = 3272 Brinkhaus ¹⁰¹ (<i>n</i> = 214), Haake ¹³² (<i>n</i> = 734), Witt ⁵⁰ (<i>n</i> = 2324)	Age, FFbHR score, PCS and MCS at baseline			<i>m</i> = 5; <i>n</i> = 3879 UK BEAM ³¹ (<i>n</i> = 479), Brinkhaus ¹⁰¹ (<i>n</i> = 212), Haake ¹³² (<i>n</i> = 734), Witt ⁵⁰ (<i>n</i> = 2248), YACBAC ¹⁰⁷ (<i>n</i> = 206)	Age, PCS and MCS at baseline	<i>m</i> = 5; <i>n</i> = 3879 UK BEAM ³¹ (<i>n</i> = 479), Brinkhaus ¹⁰¹ (<i>n</i> = 212), Haake ¹³² (<i>n</i> = 734), Witt ⁵⁰ (<i>n</i> = 2248), YACBAC ¹⁰⁷ (<i>n</i> = 206)	Age, PCS and MCS at baseline
Psychological vs. non-active usual care ^c			<i>m</i> = 3; <i>n</i> = 957 BeST ³³ (<i>n</i> = 514), VKBIA ¹⁰⁴ (<i>n</i> = 230), VKSC2 ¹⁰⁵ (<i>n</i> = 213)	Age and RMDQ score at baseline				
Sham vs. non-active usual care ^c	<i>m</i> = 2; <i>n</i> = 881 Brinkhaus ¹⁰¹ (<i>n</i> = 144), Haake ¹³² (<i>n</i> = 737)	Age, FFbHR score, PCS and MCS at baseline			<i>m</i> = 2; <i>n</i> = 879 Brinkhaus ¹⁰¹ (<i>n</i> = 142), Haake ¹³² (<i>n</i> = 737)	Age, PCS and MCS at baseline	<i>m</i> = 2; <i>n</i> = 879 Brinkhaus ¹⁰¹ (<i>n</i> = 142), Haake ¹³² (<i>n</i> = 737)	Age, PCS and MCS at baseline

YACBAC, York Acupuncture Back Pain Trial.

^a MCS of SF-12/36.^b PCS of SF-12/36.^c Control treatment is usual care/GP or waiting list.

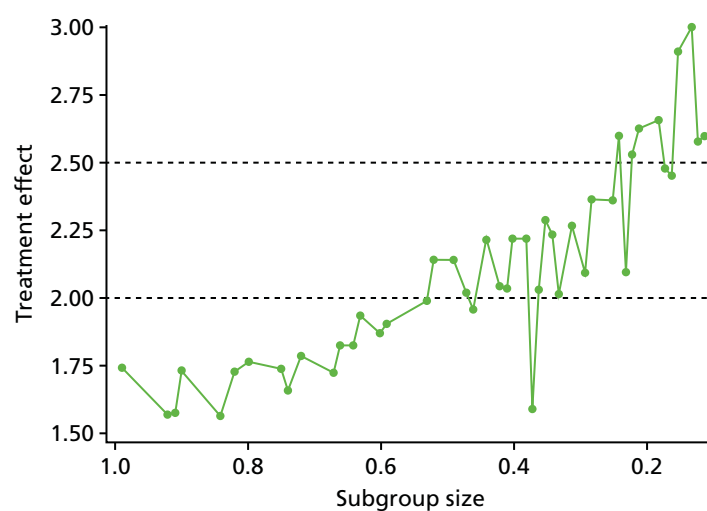


FIGURE 31 Trajectory plot for the treatment effect between active physical and non-active usual care against the size of the constructed region for the RMDQ short-term outcome.

TABLE 32 Thresholds for selected size of subgroup for the short-term RMDQ outcome as seen in *Figure 31*

Subgroup size	Age (>)	RMDQ (>)	Treatment effect
0.109	45	14	3.54
0.190	33	14	2.66
0.211	52	6	2.63
0.291	43	10	2.09
0.314	33	12	2.26
0.405	43	7	2.22
0.495	43	5	2.14
0.527	43	4	2.14
0.592	40	5	1.90
0.605	33	7	1.87
0.807	19	6	1.76
0.908	19 ^a	5	1.73

^a Minimum age = 19 years.

Passive physical care compared with non-active usual care

Short-term Hannover Functional Ability Questionnaire outcome

Figure 32 shows the trajectory plot for the treatment effect between passive physical and non-active usual care against the size of the constructed region for short-term outcome of FFbHR. Table 33 shows that the average treatment effect for approximately 90% of the population (corresponding to a FFbHR score of < 86, regardless of age, PCS and MCS values at baseline) was 10.41, which was slightly higher than the average treatment effect between any therapist-delivered intervention (active, passive, psychological or any combination treatment) and control/placebo (usual care/GP and sham treatment), which was 8.5. Approximately 20% of the population (corresponding to age < 59 years, FFbHR score of < 50, PCS of < 68 and MCS of < 72) gained at least an average treatment effect of 16 units. Younger participants with substantial physical disability (low FFbHR score) gained the most benefit. The PCS and MCS at baseline did not play an influential role in improving treatment effect.

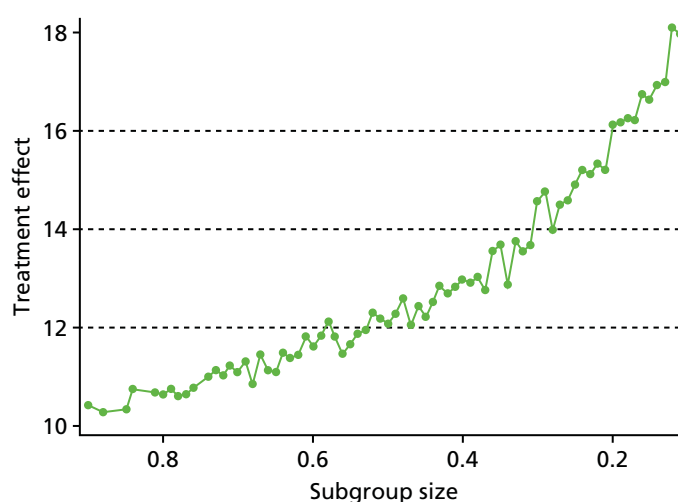


FIGURE 32 Trajectory plot for the treatment effect between passive physical and non-active usual care against the size of the constructed region for the FFbHR short-term outcome.

TABLE 33 Thresholds for selected size of subgroup for the short-term FFbHR outcome as seen in Figure 32

Subgroup size	Age, years (<)	FFbHR (<)	PCS ^a (<)	MCS ^b (<)	Treatment effect
0.101	55	41.67	67.75	72.11	18.42
0.196	68	41.67	67.75	72.11	16.18
0.207	59	50.00	67.75	72.11	16.14
0.306	68	50.00	67.75	72.11	14.57
0.407	91	54.17	40.41	72.11	12.97
0.503	63	86.36	40.41	72.11	12.08
0.602	91	79.17	40.41	60.38	11.62
0.702	68	79.17	47.80	72.11	11.10
0.807	91	100.00	43.73	72.11	10.64
0.904	91	86.36	67.75	72.11	10.41

a PCS of SF-12/36.

b MCS of SF-12/36.

Short-term Short Form questionnaire-12 items/-36 items mental component score outcome

Figure 33 shows the trajectory plot for the treatment effect between passive physical and non-active usual care, which is quite similar to the one seen above (see *Analysis 1: Overall comparison treatment compared with control/Short-term Short Form questionnaire-12 items/-36 items mental component score outcome*) where approximately 90% of the initial 3879 participants (corresponding to age < 68 years, PCS of < 68 and MCS of < 71) had an average treatment effect of 3.06 (Table 34). The treatment effect increased as more participants were excluded from the region to a clinical important difference of 6.3, but this was

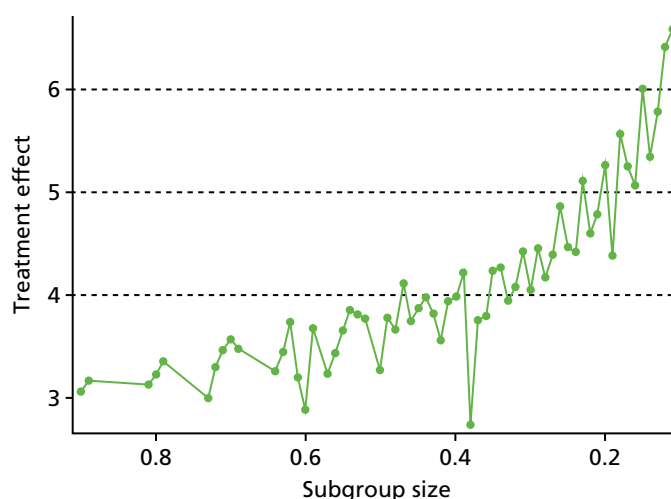


FIGURE 33 Trajectory plot for the treatment effect between passive physical and non-active usual care against the size of the constructed region for the SF-12/36 MCS short-term outcome.

TABLE 34 Thresholds for selected size of subgroup for the short-term SF-12/36 MCS as seen in Figure 33

Subgroup size	Age, years (<)	PCS ^a (<)	MCS ^b (<)	Treatment effect
0.105	51	43.50	37.86	6.33
0.193	68	35.54	47.60	4.38
0.208	63	47.65	37.86	5.26
0.296	91 ^c	67.75	37.86	4.45
0.307	63	43.50	47.60	4.05
0.392	91	43.50	47.60	4.21
0.403	91	37.84	54.15	3.99
0.496	91	67.75	47.60	3.77
0.500	63	47.65	54.15	3.27
0.594	91	67.75	51.02	3.67
0.603	55	67.75	71.32	2.88
0.706	91	43.50	60.37	3.57
0.802	91	47.65	60.37	3.22
0.904	68	67.75	71.32	3.06

a PCS of SF-12/36.

b MCS of SF-12/36.

c Maximum age = 91 years.

applicable to only a small proportion of participants – approximately 10% of them (corresponding to age < 51 years, PCS of < 44 and MCS of < 38), that is, only younger participants with substantial physical limitations and psychological distress would benefit from greater improvement in passive physical treatment against control.

Short-term Short Form questionnaire-12 items/36 items physical component score outcome

The trajectory plot for the treatment effect between passive physical and non-active usual care is shown in *Figure 34*. The trajectory indicates an increase of improvement as regions narrowed but the fluctuation of the treatment effect suggests that there might be no definite subgroup that would gain substantial treatment effect. *Table 35* summarises the average treatment for selected constructed regions with the corresponding thresholds for the comparison seen in *Figure 34*.

Psychological versus non-active usual care

Short-term Roland–Morris Disability Questionnaire outcome

Figure 35 shows the trajectory plot for the treatment effect between psychological and non-active usual care for the short-term RMDQ outcome, and *Table 36* shows the average treatment effect for selected constructed regions with the corresponding thresholds. The results are very similar to those seen above (see *Analysis 1: Short-term Roland–Morris Disability Questionnaire outcome*), that is, there was no subgroup that would gain a substantial improvement in treatment effect.

Sham care compared with non-active usual care

Short-term Hannover Functional Ability Questionnaire outcome

Three trials^{50,101,132} were included in the comparison between passive physical and non-active usual care. All three trials^{50,101,132} had acupuncture as the therapist-delivered intervention; of these, two trials^{101,132} also had sham acupuncture. *Figure 36* shows the trajectory plot for the treatment effect between sham acupuncture and non-active usual care. The average treatment effect was slightly lower between passive physical (acupuncture) and non-active usual care. However, the treatment effect increased as more and more participants were excluded from the ARDP-MA algorithm. *Table 37* shows the average treatment effect between sham acupuncture and non-active usual care for selected constructed regions with the corresponding thresholds.

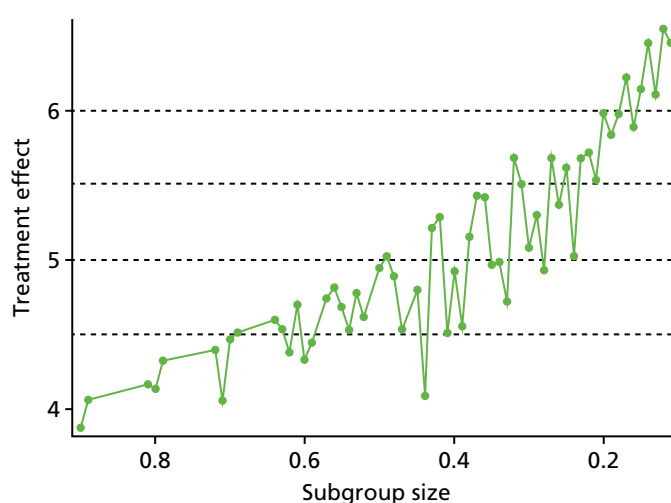


FIGURE 34 Trajectory plot for the treatment effect between passive physical and non-active usual care against the size of the constructed region for the SF-12/36 PCS short-term outcome.

TABLE 35 Thresholds for selected size of subgroup for the short-term SF-12/36 PCS as seen in *Figure 34*

Subgroup size	Age, years (<)	PCS ^a (<)	MCS ^b (>)	Treatment effect
0.107	63	31.19	51.02	6.17
0.192	68	35.54	51.02	5.84
0.205	91 ^c	31.19	43.02	5.99
0.292	68	43.50	51.02	5.30
0.310	55	40.28	33.48	5.09
0.394	68	35.54	28.47	4.56
0.406	91	43.50	47.60	4.93
0.495	91	40.28	37.86	5.02
0.503	68	43.50	37.86	4.95
0.599	91	37.84	9.46	4.45
0.604	91	67.75	43.02	4.33
0.709	68	43.50	9.46	4.47
0.802	91	67.75	33.48	4.14
0.904	68	67.75	9.46	3.88

a PCS of SF-12/36.

b MCS of SF-12/36.

c Maximum age = 91 years.

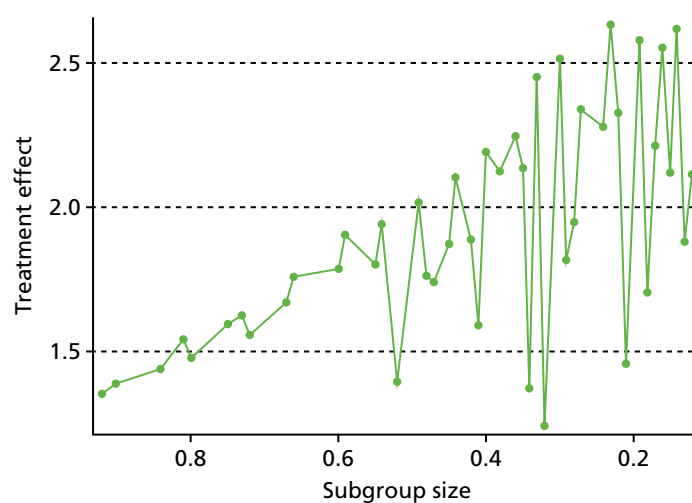
**FIGURE 35** The size of the constructed region for the RMDQ short-term outcome.

TABLE 36 Thresholds for selected size of subgroup for the short-term RMDQ outcome as seen in *Figure 35*

Subgroup size	Age, years (<)	RMDQ (>)	Treatment effect
0.107	41	7	2.84
0.197	49	8	2.58
0.214	69	13	1.46
0.295	45	0	1.81
0.305	49	5	2.52
0.400	52	4	2.19
0.493	56	4	2.02
0.528	85 ^a	8	1.39
0.591	60	4	1.90
0.606	63	5	1.79
0.809	63	0	1.48
0.909	69	0	1.39

^a Maximum age = 85 years.

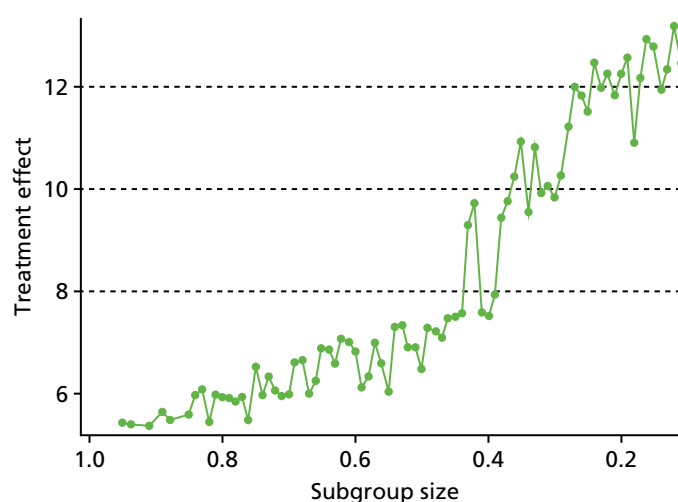
**FIGURE 36** The size of the constructed region for the FFbHR short-term outcome.

TABLE 37 Thresholds for selected size of subgroup for the short-term FFbHR outcome as seen in *Figure 36*

Subgroup size	Age, years (<)	FFbHR (<)	PCS ^a (<)	MCS ^b (<)	Treatment effect
0.103	52	41.67	44.78	51.61	12.64
0.199	52	54.17	60.47	51.61	12.58
0.208	62	45.83	44.78	51.61	12.26
0.301	62	45.83	60.47	72.11	9.85
0.402	52	95.83	60.47	57.68	7.53
0.510	68	58.33	41.50	61.38	6.49
0.605	87 ^c	62.50	41.50	57.68	6.84
0.700	68	66.67	44.78	61.38	6.00
0.806	68	95.83	44.78	72.11	5.95

a PCS of SF-12/36.

b MCS of SF-12/36.

c Maximum age = 87 years.

Short-term SF-12/36 MCS outcome

Figure 37 shows the trajectory plot for the treatment effect between sham and non-active usual care.

The two trials included in this pairwise analysis had sham acupuncture. The figure shows that the average treatment effect did not improve much in the exclusion of the first 70% participants (*Table 38*).

Nevertheless, there was a markedly higher treatment effect which was 6.22 for approximately 20% of the participants (corresponding to PCS of < 36 and MCS of < 39, regardless of age).

Short-term physical component score outcome

The trajectory plot for the treatment effect between sham and non-active usual care is shown in *Figure 38* and *Table 39* summarises the average treatment for selected constructed regions with the corresponding thresholds. There was an increase of improvement as regions narrowed, but the fluctuation of the treatment effect suggests that there might be no definite subpopulation that would gain substantial treatment effect.

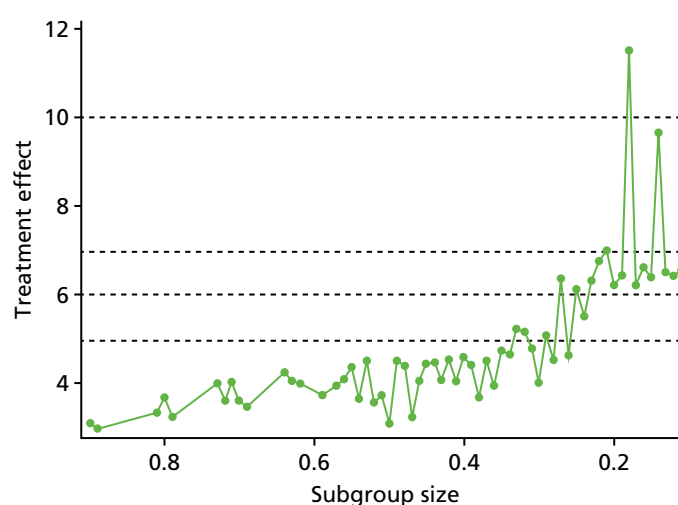
**FIGURE 37** Trajectory plot for the treatment effect between sham and non-active usual care against the size of the constructed region for the SF-12/36 MCS short-term outcome.

TABLE 38 Thresholds for selected size of subgroup for the short-term SF-12/36 MCS as seen in *Figure 37*

Subgroup size	Age, years (<)	PCS ^a (<)	MCS ^b (<)	Treatment effect
0.104	43	36.48	51.97	7.86
0.199	43	39.17	61.54	6.43
0.201	87	36.48	39.07	6.22
0.296	87	57.59	39.07	5.06
0.300	65	42.29	44.25	4.01
0.396	87 ^c	39.17	48.42	4.40
0.410	52	42.29	61.54	4.57
0.501	61	57.59	55.18	3.09
0.709	70	39.17	70.46	3.59
0.809	70	42.29	70.46	3.67
0.902	70	57.59	70.46	3.09

a PCS of SF-12/36.

b MCS of SF-12/36.

c Maximum age = 87.

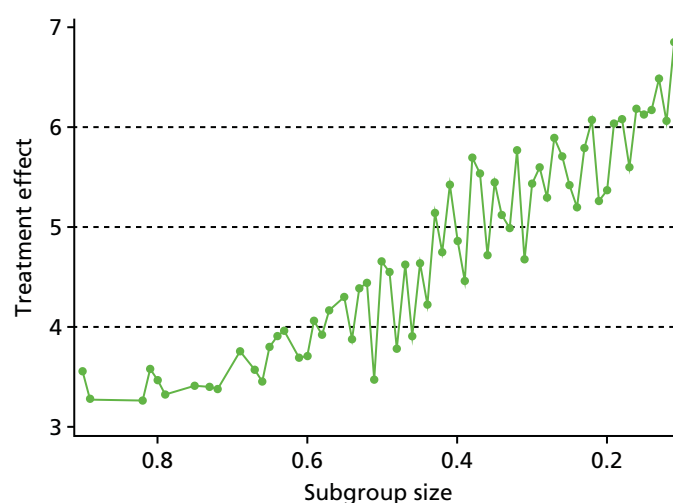
**FIGURE 38** Trajectory plot for the treatment effect between sham and non-active usual care against the size of the constructed region for the SF-12/36 PCS short-term outcome.

TABLE 39 Thresholds for selected size of subgroup for the short-term SF-12/36 PCS as seen in *Figure 38*

Subgroup size	Age, years (<)	PCS ^a (>)	MCS ^b (<)	Treatment effect
0.100	70	39.17	48.42	6.26
0.195	52	32.56	51.97	6.04
0.206	70	36.48	55.18	5.37
0.296	52	30.95	58.10	5.59
0.303	70	30.95	48.42	5.43
0.398	87 ^c	34.31	70.46	4.46
0.403	65	32.56	61.54	4.86
0.495	87	26.96	51.97	4.55
0.503	87	30.95	58.10	4.66
0.598	87	30.95	70.46	4.06
0.602	70	29.16	61.54	3.71
0.801	65	14.41	70.46	3.46
0.902	70	14.41	70.46	3.56

a PCS of SF-12/36.

b MCS of SF-12/36.

c Maximum age = 87 years.

Chapter 9 Methodology and statistical developments 3: identification of cost-effective subgroups by directed peeling

Introduction

The economic analyses sought to identify the most cost-effective treatments for subgroups of patients with LBP. A search algorithm, similar to that used in the previous chapter, was used to identify subgroups to maximise the expected QALY gain from treatment. Although some of the trials in the database provided individual-level data on use of health-care resources, these data were not used in the analyses presented in this chapter. Instead, a threshold approach was used to assess the cost-effectiveness of treatment for defined groups of patients. This was done by comparing estimates of treatment cost from the literature with the maximum cost required to stay below the cost-effectiveness threshold (£20,000–30,000 per QALY, as recommended by NICE), given the estimated QALY gain from treatment.¹⁷³

The use of the QALY outcome reduced the available data for analysis more than for the short-term clinical outcomes in the previous chapter. We therefore used a search algorithm that is suited to data with a lower signal–noise ratio: the directed peeling approach of LeBlanc *et al.*,¹⁷² which works by ‘peeling’ a fraction of patients (with the least favourable effect) from the subgroup in a series of steps. This differs from the full search algorithm described in the previous chapter, as each successively smaller subgroup is constrained to be a subset of the previous one. Both approaches use a ‘directed’ peeling approach, designed to provide simpler descriptions of groups for variables with a monotonic relationship with the outcome of interest. The LeBlanc *et al.*¹⁷² algorithm was developed for analysis of data from a single trial, and so it was adapted here for IPD meta-analysis by incorporating random trial effects into the model.

The analysis was split into four overarching comparisons: all interventions collectively compared with best care; active physical interventions compared with best care; passive physical interventions compared with best care; and active physical interventions compared with passive physical interventions. Psychological interventions were not included in the comparison, as only one trial had the EQ-5D data that was necessary to calculate a QALY and a control arm. Data for comparisons against a ‘sham’ treatment arm were also excluded from this analysis.

Methods

Quality-adjusted life-years

The outcome used for the analysis was the QALY. We calculated QALYs for individuals based on EQ-5D utility scores at baseline and short-, medium- and long-term follow-up (up to 1 year). For trials with SF-36/SF-12 outcomes but no EQ-5D, we used a mapping algorithm¹⁶² to estimate EQ-5D scores. QALYs were estimated using an AUC approach, adjusting for baseline EQ-5D scores (see *Chapter 6, Health-economic outcomes*).

Moderator identification

The specification of the search algorithm required an initial analysis to identify moderating variables, and to determine the direction of peeling. A mixed-effects model was used to identify moderators with a significant interaction with treatment effect on the QALY outcome. The model was specified with moderator, treatment and treatment-by-moderator interaction as fixed effects, and trial and treatment-by-trial interaction as random effects (see *Chapter 6, Outcome variables*). The sign on the

moderator by treatment interaction coefficient dictated whether or not the algorithm should peel from the top or the bottom of the moderator range. A positive relationship with treatment effect suggested that peeling away individuals with lower values of the moderator would yield higher average treatment benefits. A negative relationship suggests that peeling individuals with higher values of the moderator would be best.

Peeling algorithm

The peeling algorithm started by setting the subgroup indicator (B) to '1' for all individuals. Incremental QALY gain from treatment for the whole patient sample was estimated using a mixed-effects model with baseline EQ-5D score and treatment as fixed effects, and trial and treatment-by-trial interaction as random effects.

The algorithm then looped through the following steps until the stopping criteria were met:

- For each moderator, a small proportion of the data was peeled off, taking out the individuals with the highest (lowest) value of the moderator (depending on the direction of the moderator treatment interaction effect). The subgroup indicator (B) was set to '1' for the remaining individuals (the 'in' group) and '0' for the peeled individuals (the 'out' group).
- The difference in incremental QALY gain was estimated for those inside the subgroup compared with those outside using a mixed-effects model: with baseline EQ-5D, treatment effect, subgroup identifier and treatment-by-subgroup interaction as fixed effects, and trial and treatment-by-trial interaction as random effects.
- The magnitude of the treatment by subgroup interaction effect was compared for each moderator. The peel decision was then based on the moderator with the greatest effect.
- Summary statistics were calculated, including the incremental QALY gain within the subgroup, the incremental QALY gain outside the subgroup and the weighted mean incremental QALY across the whole sample.
- If the subgroup contained fewer individuals than a preset minimum number (n_{min}) then the algorithm stopped. Otherwise, the above steps were repeated.

Cost-effectiveness

Individual patient data on health-care resource use was available for some trials in the repository. An initial analysis was conducted using the data from the UK BEAM trial³¹ using individual-level estimates of costs (C) and QALYs (Q) over the 12-month follow-up period. From these data, the net monetary benefit (NMB) was calculated for each individual: $NMB = \lambda \times Q - C$, where λ is a set cost-effectiveness threshold (£20,000 per QALY). This NMB variable was then used as an outcome in the above search algorithm. However, we found that the addition of the cost data increased variation without increasing predictive power. The results of this analysis are not presented here, as one condition of use of the repository data is that all results must include at least two trials to avoid re-analysis of the original trial data. Given that the addition of the individual-level costs was not advantageous in the UK BEAM³¹ analysis, and also the heterogeneity in the resource-use items recorded across those studies with data, we decided to focus on QALYs as the outcome for the economic analysis, and to use a threshold approach to assess cost-effectiveness.

The threshold analysis presents the maximum incremental cost of intervention in order for a treatment subgroup to be deemed cost-effective based on the lower and upper limits of the NICE-recommended threshold (£20,000–30,000 per QALY). For example, if a treatment yields an average incremental QALY gain for a treatment population of 0.05, one would pay up to £1000 ($0.05 \times £20,000$) for the treatment, using the lower threshold or £1500 ($0.05 \times £30,000$) at the upper threshold.

Published literature was used to provide indicative costs of treatment for comparison with the estimated thresholds. The incremental cost of passive treatment over 1 year was estimated at £541 (SD £768) from the UK BEAM³¹ economic analysis: £147 for the intervention and £394 relating to other health-care costs. Estimates for other treatments varied, ranging from £422 (£187 for the intervention, £235 for other health-care costs) for a psychological intervention (BeST³³) to £486 (SD £907) comprised £41 for the intervention and £445 relating to other health-care costs for active therapies (UK BEAM³¹).

Results

Six analyses were run (*Table 40*), dictated by the moderators with significant treatment interaction terms in the QALY ANCOVA. These included the following comparisons: all interventions compared with control; active physical interventions compared with control; passive physical interventions compared with control; and active physical interventions compared with passive physical interventions. As noted above, analyses of psychological intervention and sham were omitted, as in each case only one study provided data for QALY calculation.

As shown in *Table 40*, not all trials had data for all three potential moderators. We therefore conducted three analyses for the intervention against control comparison: the first to include as many trials as possible with QALY data (age and PCS as moderators).

All interventions versus control: moderators – age and physical component score

The algorithm trace is shown in *Figure 39*. The y-axis shows the estimated treatment effect for the subgroup, that is, the 'incremental QALYs' gained from treatment compared with the control arm. The x-axis is the proportion of the starting population peeled away from the treatment group. *Figure 40* shows the mean incremental QALYs for the whole sample, both inside and outside the treatment group. It can be seen that for the full sample, the incremental QALY is declining as a function of the treatment subgroup size. This suggests that those being peeled from the subgroup had a net QALY gain from treatment. However, there is no strong signal in these data. The peeling trace in *Figure 39* shows no notable increase in QALY gain from treatment when up to 80% of the sample are removed from the treatment group. Full details of the peeling trace are available in *Table 41*. Both age and PCS were used for peeling, although over the trace the algorithm favoured peeling that was based on PCS. There is a small rise in QALY gain at the point where 90% of the sample had been removed; the subgroup comprising 10% of the sample included participants aged between 54 and 84 years with a PCS of between 7 and 28. The estimated QALY gain from treating only this subgroup was 0.0852, whereas the estimated mean QALY gain from treating the whole population was lower, at 0.0624.

TABLE 40 Adaptive refinement by directed peeling in IPD meta-analysis: analyses conducted on economic outcomes

Analysis	Outcome variable	Moderators included	Trials included	Sample size: intervention, control
All interventions vs. control				
9.3.1	QALY	Age, PCS ^a	UK BEAM, ³¹ BeST, ³³ YACBAC, ¹⁰⁷ Haake ¹³²	1273, 715
9.3.2	QALY	Age, RMDQ	UK BEAM, ³¹ BeST, ³³ York BP, ¹³³ Smeets ⁷⁰	1092, 422
9.3.3	QALY	Age, PCS, RMDQ	UK BEAM, ³¹ BeST ³³	827, 323
Active physical interventions vs. control				
9.3.4	QALY	Age, RMDQ	UK BEAM, ³¹ York BP ¹³³	232, 264
Passive physical interventions vs. control				
9.3.5	QALY	Age, PCS	UK BEAM, ³¹ YACBAC, ¹⁰⁷ Haake ¹³²	643, 566
Active physical vs. passive physical interventions				
9.3.6	QALY	Age, RMDQ	UK BEAM, ³¹ HullExPro ⁷⁶	232, 288
BEAM, Back pain Exercise And Manipulation; YACBAC, York Acupuncture Back Pain Trial. a PCS of SF-12/36.				

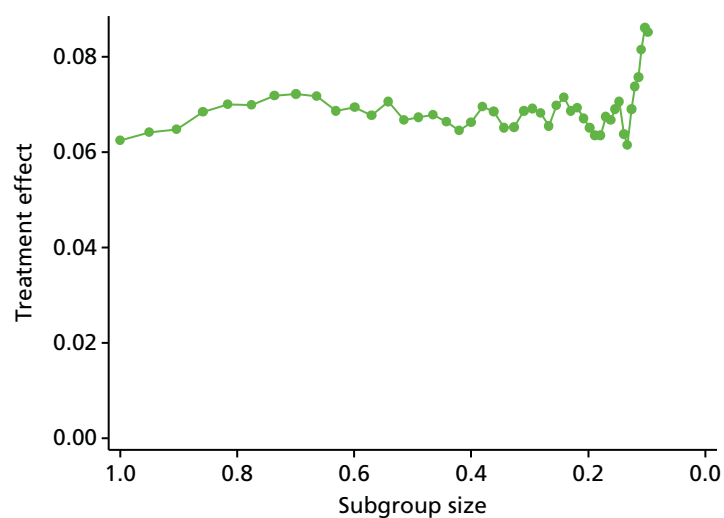


FIGURE 39 Mean treatment effect in subgroup. All interventions vs. control: moderators – age and PCS.

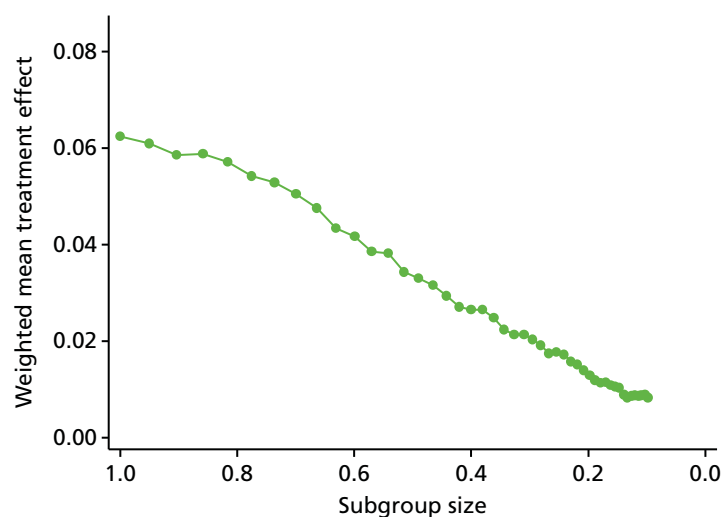


FIGURE 40 Weighted mean treatment effect across treatment subgroup and non-treatment subgroup.

TABLE 41 Algorithm output for analysis 9.3.1 (see Table 40)

Iteration	Moderator	Direction peeled	Proportion in subgroup	≈n	Incremental QALYs		Age		PCS ^a	
					Subgroup	All	Minimum	Maximum	Minimum	Maximum
0	–	–	1.00	1988	0.0624	0.0624	18	87	7	61
1	PCS	Top	0.95	1889	0.0642	0.0610	18	87	7	50
2	Age	Bottom	0.90	1795	0.0648	0.0585	28	87	7	50
3	Age	Bottom	0.86	1706	0.0685	0.0588	32	87	7	50
4	PCS	Top	0.82	1621	0.0700	0.0571	32	87	7	47
5	PCS	Top	0.77	1540	0.0700	0.0542	32	87	7	45
6	PCS	Top	0.74	1463	0.0718	0.0529	32	87	7	43
7	PCS	Top	0.70	1390	0.0722	0.0505	32	87	7	42
8	Age	Bottom	0.66	1319	0.0718	0.0476	34	87	7	42
9	PCS	Top	0.63	1254	0.0688	0.0434	34	87	7	41
10	PCS	Top	0.60	1192	0.0695	0.0417	34	87	7	40
11	PCS	top	0.57	1133	0.0677	0.0386	34	87	7	39
12	PCS	top	0.54	1077	0.0706	0.0383	34	87	7	38
13	PCS	Top	0.52	1024	0.0668	0.0344	34	87	7	38
14	PCS	Top	0.49	973	0.0674	0.0330	34	87	7	37
15	PCS	Top	0.47	925	0.0679	0.0316	34	87	7	36
16	PCS	Top	0.44	879	0.0664	0.0294	34	87	7	36
17	PCS	Top	0.42	836	0.0645	0.0271	34	87	7	35
18	PCS	Top	0.40	795	0.0663	0.0265	34	87	7	35
19	PCS	Top	0.38	756	0.0696	0.0265	34	87	7	34

continued

TABLE 41 Algorithm output for analysis 9.3.1 (see Table 40) (continued)

Iteration	Moderator	Direction peeled	Proportion in subgroup	$\approx n$	Incremental QALYs		Age		PCS ^a	
					Subgroup	All	Minimum	Maximum	Minimum	Maximum
20	Age	Bottom	0.36	719	0.0686	0.0248	36	87	7	34
21	Age	Bottom	0.34	683	0.0652	0.0224	39	87	7	34
22	PCS	Top	0.33	649	0.0652	0.0213	39	87	7	34
23	PCS	Top	0.31	617	0.0688	0.0213	39	87	7	33
24	PCS	Top	0.30	587	0.0691	0.0204	39	87	7	33
25	PCS	Top	0.28	558	0.0682	0.0191	39	87	7	33
26	Age	Bottom	0.27	531	0.0655	0.0175	41	87	7	33
27	Age	Bottom	0.25	505	0.0698	0.0177	43	87	7	33
28	Age	Bottom	0.24	480	0.0716	0.0173	45	87	7	33
29	Age	Bottom	0.23	456	0.0687	0.0158	47	87	7	33
30	Age	Bottom	0.22	434	0.0694	0.0151	49	87	7	33
31	Age	Bottom	0.21	413	0.0671	0.0139	50	87	7	33
32	Age	Bottom	0.20	393	0.0652	0.0129	51	87	7	33
46	PCS	Top	0.10	196	0.0852	0.0084	54	84	7	28

a PCS of SF-12/36.

Depending on the cost of intervention, and NHS 'willingness-to-pay' per QALY, it might be cost-effective for all patients to be offered treatment or for treatment to be limited to a selected subgroup. For example, at a cost-effectiveness threshold of £20,000 per QALY, the maximum that the NHS would pay for the 'intervention' reflected here would be £1248 (per patient over the course of a year) if all patients were to be offered treatment or £1704 if only patients in the 10% subgroup were to be offered treatment. If the threshold of £30,000 was applied then this will be £1872 and £2556, respectively. However, these results do not incorporate any measure of uncertainty and should be considered as only illustrative of the method.

- Older patients with relatively worse physical functioning as measured using the PCS at baseline appear to have moderately better response to treatment.

All interventions versus control: moderators – age and Roland–Morris Disability Questionnaire

Figures 41 and 42 illustrate the peeling trace with moderator's age and RMDQ. The inclusion of the RMDQ limited the sample to four trials (see Table 40). As shown by Figure 41, the peeling algorithm did achieve small but consistent gains in treatment effect within the subgroup, as participants with better (lower) baseline RMDQ scores and who were younger were removed from the treatment group. The algorithm

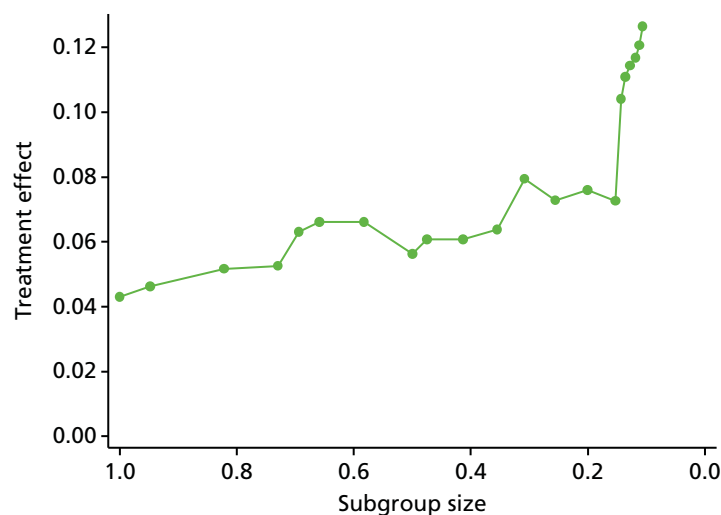


FIGURE 41 Mean treatment effect in subgroup. All interventions vs. control: moderators – age and RMDQ.

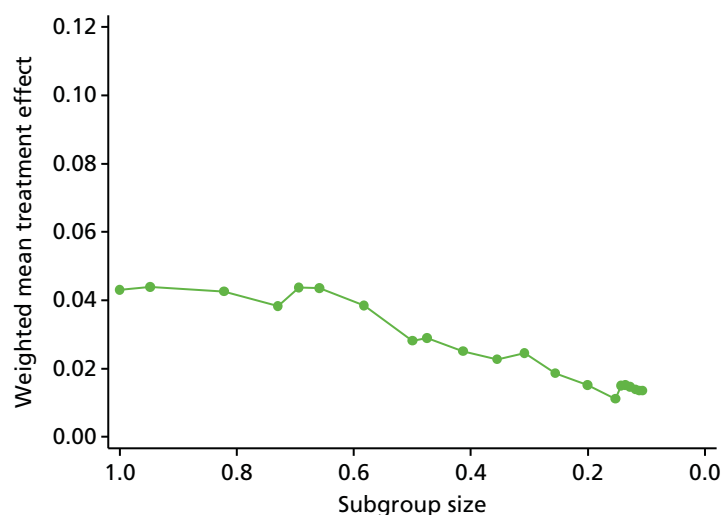


FIGURE 42 Weighted mean treatment effect across treatment subgroup and non-treatment subgroup.

favoured peeling that was based on RMDQ score during the earlier iterations. The apparent monotonicity of RMDQ with respect to treatment effect (as measured in QALYs) is consistent with the regression analysis used for moderator identification (see *Table 17*), as the RMDQ had a more significant relationship with treatment effect compared with age. Owing to some correlation with RMDQ score and age, some older patients were removed from the treatment subgroup as the algorithm peeled based on RMDQ score.

The peeling trace for Analysis 2 is shown in *Table 42*. The subgroup at 20% of the initial sample comprised participants aged > 34 years with a RMDQ score of ≥ 13 . A modest improvement in QALYs gained from treatment can be seen for this subgroup: from 0.043 if the whole population were to be offered treatment to 0.076 for the subgroup. As described previously, the maximum willingness to pay for an intervention yielding these QALY gains would be £860 and £1520, respectively, for the whole population and for the subgroup, where a threshold of £20,000 is applied, or £1290 and £2280, respectively, at a threshold of £30,000 per QALY. As there is no estimation of uncertainty, this result should be seen as illustrative.

- Older patients with worse baseline physical functioning as measured by the RMDQ score at baseline appear to achieve a moderately better response to treatment.

All interventions versus control: moderators – age, physical component score and Roland–Morris Disability Questionnaire

Figures 43 and 44 illustrate the peeling results for the analysis with age, PCS and RMDQ score. As some trials did not have available PCSs and others did not have RMDQ scores, the sample was restricted to two trials.^{31,33} The results of the peeling trace are very similar to those of analysis 2, shown in *Table 42*). The algorithm chose to peel almost exclusively on RMDQ score and age. PCS was used for the first iteration only. As the algorithm reduced the size of the treatment subgroup, the results showed that generally, older patients with worse (higher) RMDQ scores achieved better QALY gains from treatment. Although PCS was not much used for peeling, as the sample size was reduced participants with higher (better) PCSs were removed from the treatment subgroup; this is unsurprising as RMDQ score and PCS are correlated.

As shown in *Table 43* at the point where 19% of the starting sample was left in the treatment subgroup, the subgroup comprised participants aged 44 to 82 years with a RMDQ score over 12 and a PCS of between 7 and 49. At this point the treatment subgroup achieved a QALY gain of 0.0981 from treatment. When the whole population was treated, the mean QALY gain was lower at 0.0504. At a £20,000 per QALY cost-effectiveness threshold, the maximum willingness to pay for an intervention yielding these QALY gains would be £1008 and £1962 for the whole population and the refined subgroup, respectively. At £30,000 per QALY, these figures are £1512 and £2943, respectively. However, as there is no measure of uncertainty reflected in these results, they should only be seen as illustrative.

- Older patients with worse physical functioning as measured using the RMDQ score at baseline appear to have a moderately better response to treatment.

Active physical intervention versus control: moderators – age and Roland–Morris Disability Questionnaire

Analysis so far has pooled all treatment modalities and compared these collectively with control. For analysis 9.3.4 (see *Table 40*) the intervention considered is made up of only active physical interventions, in this case 'exercise'. The comparator arm is still control. This approach limited the data set to two trials.^{31,133} *Figure 45* shows the peeling trace, with RMDQ score and age included as moderators within the algorithm.

The algorithm peeled almost exclusively based on the RMDQ score. As the algorithm reduced the sample size, patients with lower (better) RMDQ scores were removed, suggesting that patients with worse baseline RMDQ scores achieve better treatment outcomes. At iteration 10, age was peeled on, removing patients who were younger.

TABLE 42 Algorithm output for analysis 9.3.2 (see Table 40)

Iteration	Moderator	Direction peeled	Proportion in subgroup	≈n	Incremental QALYs		Age		RMDQ score	
					Subgroup	All	Minimum	Maximum	Minimum	Maximum
0	–	–	1.00	1514	0.0431	0.0431	18	85	0	24
1	RMDQ	Bottom	0.95	1435	0.0463	0.0439	18	85	3	24
2	RMDQ	Bottom	0.82	1245	0.0517	0.0425	19	84	5	24
3	RMDQ	Bottom	0.73	1105	0.0525	0.0383	19	84	6	24
4	Age	Bottom	0.69	1050	0.0631	0.0437	28	84	6	24
5	Age	Bottom	0.66	998	0.0661	0.0436	32	84	6	24
6	RMDQ	Bottom	0.58	881	0.0661	0.0385	32	84	7	24
7	RMDQ	Bottom	0.50	756	0.0563	0.0281	32	84	8	24
8	Age	Bottom	0.47	719	0.0608	0.0289	34	84	8	24
9	RMDQ	Bottom	0.41	625	0.0608	0.0251	34	84	9	24
10	RMDQ	Bottom	0.35	537	0.0639	0.0227	34	84	10	24
11	RMDQ	Bottom	0.31	466	0.0794	0.0244	34	82	11	24
12	RMDQ	Bottom	0.26	387	0.0728	0.0186	34	82	12	24
13	RMDQ	Bottom	0.20	304	0.0760	0.0153	34	82	13	24
14	RMDQ	Bottom	0.15	232	0.0726	0.0111	34	79	14	24
15	Age	Bottom	0.14	217	0.1041	0.0149	38	79	14	24
16	Age	Bottom	0.14	206	0.1109	0.0151	39	79	14	24
17	Age	Bottom	0.13	194	0.1143	0.0146	41	79	14	24
18	Age	Bottom	0.12	179	0.1168	0.0138	44	79	14	24
19	Age	Bottom	0.11	170	0.1206	0.0135	44	79	14	24
20	Age	Bottom	0.11	161	0.1265	0.0134	46	79	14	24

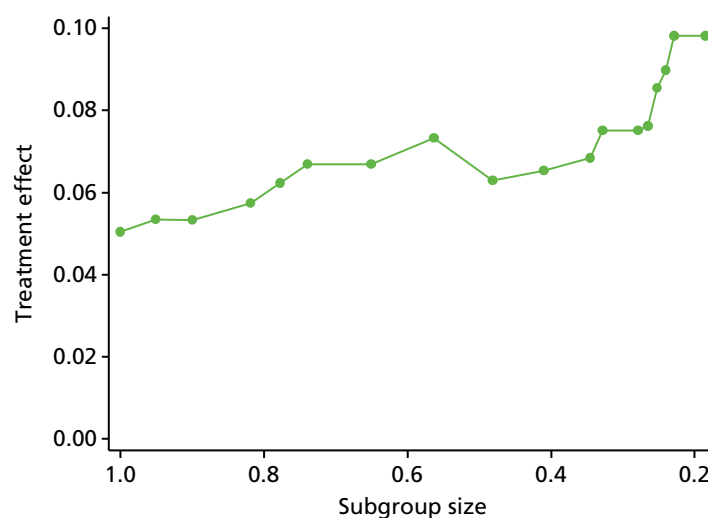


FIGURE 43 Mean treatment effect in subgroup. All interventions vs. control: moderators – age, PCS and RMDQ.

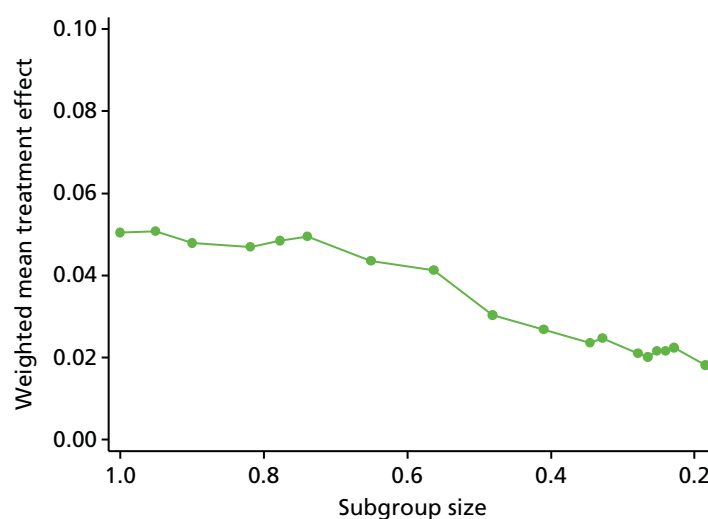


FIGURE 44 Weighted mean treatment effect across treatment subgroup and non-treatment subgroup.

TABLE 43 Algorithm output for analysis 9.3.3 (see Table 40)

Iteration	Moderator	Direction peeled	Proportion in subgroup	$\approx n$	Incremental QALYs		Age		PCS ^a		RMDQ score	
					Subgroup	All	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
0	-	-	1.00	1150	0.0504		18	85	7	61	0	24
1	PCS	Top	0.95	1093	0.0534	-0.0086	18	85	7	51	0	24
2	RMDQ	Bottom	0.90	1034	0.0533	0.0037	18	85	7	51	4	24
3	RMDQ	Bottom	0.82	941	0.0574	-0.0133	19	84	7	51	5	24
4	Age	Bottom	0.78	894	0.0624	0.0187	29	84	7	51	5	24
5	Age	Bottom	0.74	850	0.0669	0.0087	32	84	7	51	5	24
6	RMDQ	Bottom	0.65	748	0.0669	0.0087	32	84	7	51	6	24
7	RMDQ	Bottom	0.56	648	0.0733	0.0100	32	84	7	51	7	24
8	RMDQ	Bottom	0.48	554	0.0629	0.0354	32	84	7	51	8	24
9	RMDQ	Bottom	0.41	472	0.0653	0.0410	32	84	7	51	9	24
10	RMDQ	Bottom	0.35	397	0.0684	0.0438	32	84	7	49	10	24
11	Age	Bottom	0.33	378	0.0751	0.0429	35	84	7	49	10	24
12	RMDQ	Bottom	0.28	321	0.0751	0.0429	35	82	7	49	11	24
13	Age	Bottom	0.27	305	0.0762	0.0394	38	82	7	49	11	24
14	Age	Bottom	0.25	290	0.0855	0.0367	40	82	7	49	11	24
15	Age	Bottom	0.24	276	0.0899	0.0363	42	82	7	49	11	24
16	Age	Bottom	0.23	263	0.0981	0.0343	44	82	7	49	11	24
17	RMDQ	Bottom	0.19	213	0.0981	0.0343	44	82	7	49	12	24

^a PCS of SF-12/36.

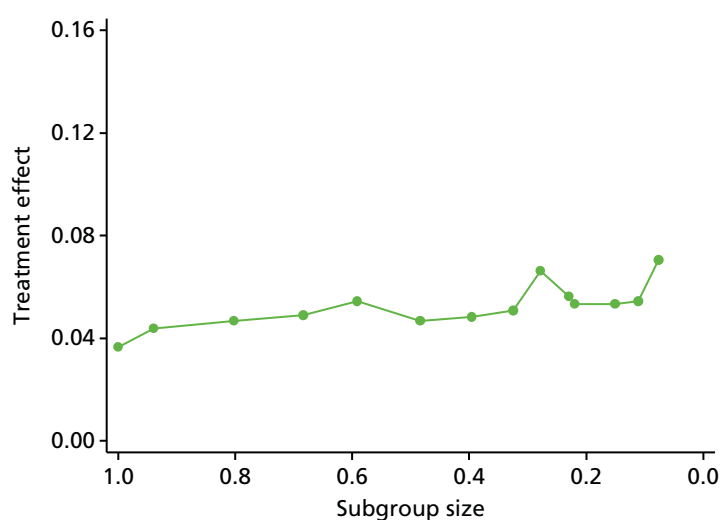


FIGURE 45 Mean treatment effect in subgroup. Active physical intervention vs. control: moderators – age and RMDQ.

As can be seen in *Figure 45*, improvements in the mean incremental treatment effect for the subgroup were very small as no relevant subgroup could be identified from APT in these analyses.

Passive physical intervention versus control: moderators – age and physical component score

Analysis 9.3.5 (see *Table 40*) follows the same approach as analysis 9.3.4 (see *Table 40*); however, in this instance the treatment arm comprised only passive interventions, including manipulation and acupuncture treatments; the comparator remained as a control. These conditions limited the data set to three trials.^{31,105,132} The peeling algorithm was set to peel based on age and PCS. RMDQ score was not available for all of the trials included in this analysis.

As can be seen on *Figure 46*, there was very little change in the incremental treatment effect as the algorithm refined the treatment subgroup. No relevant subgroup could be identified correlating age and/or PCS with above average treatment effect from passive physical treatment in these analyses.

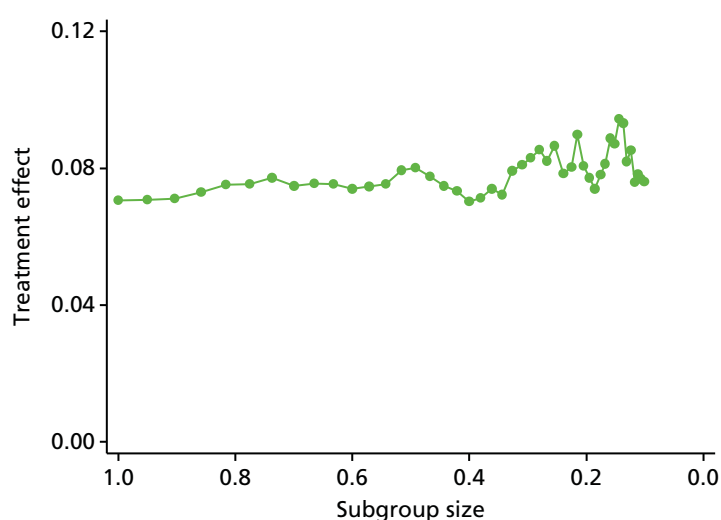


FIGURE 46 Mean treatment effect in subgroup. Passive physical intervention vs. control: moderators – age and PCS.

Adaptive refinement by directed peeling in individual patient data meta-analysis directed peel

Active physical interventions compared with passive physical interventions: moderators – age and Roland–Morris Disability Questionnaire

Analysis 9.3.6 (see *Table 40*) was a comparison of active physical interventions and passive physical interventions. The analysis includes data from two trials. The active treatment was made up of exercise and the passive treatment was made up of manual therapy. For the analysis, passive treatment was considered the reference case for all of the incremental estimates. The peel algorithm was set to refine the subgroup based on the age and RMDQ moderators. The algorithm elected to peel predominantly on the RMDQ score, removing patients with lower (better) RMDQ scores from the treatment group. As can be seen in *Figure 47*, the incremental effect of changing between these two treatment modalities was near zero. The result of the analysis suggests there is no difference in these two treatment modalities across the whole sample, or for any subgroup explored within the analysis of these data.

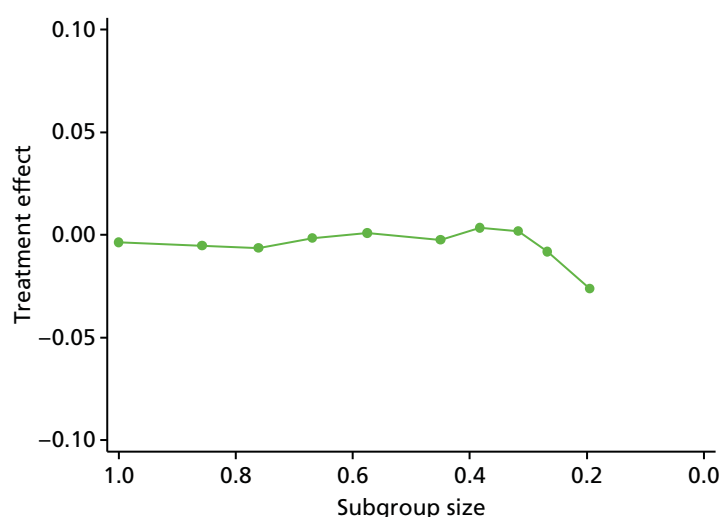


FIGURE 47 Mean treatment effect in subgroup. Active physical interventions compared with passive physical interventions: moderators – age and RMDQ.

Discussion

The application of the peeling algorithm was successful in identifying potentially interesting subgroups for the interventions against control comparison. These subgroups comprised patients who were older, with relatively worse physical functioning at baseline. The gain in treatment effect for the subgroup was small; therefore, given the relatively low cost of the intervention treatment, it is likely to be cost-effective to offer treatment to the whole patient group. The algorithm, however, was not successful in finding any convincing subgroup in the pairwise comparison of active and passive physical treatment. This may be caused by lack of power or simply because there is no subgroup to be found.

The QALY has some key advantages over the other available clinical outcomes. It is a holistic measure of health-related quality of life designed to encompass both physical and mental aspects of a patient's health state. Constructed using EQ-5D responses over time, the QALY also takes account of a patient's recovery profile, integrating short- and long-term treatment response into a single measure. The EQ-5D is scored using the UK social tariff, which is validated and standardised allowing direct comparison of the treatment response for different interventions and diseases. The QALY estimated using the EQ-5D tariff is the accepted measure used by NICE for assessing the cost-effectiveness of new treatments for approval in the NHS. The QALY did, however, raise some particular challenges for the analysis. The use of repeated measures to estimate the QALY restricted the size of the sample, as more observations were lost to missing data when compared with the point estimates used in the clinical analysis. This reduced the power of statistical analyses.

The same approach was taken for moderator identification for the economic component of the analysis as for the clinical analyses. Three potential moderators (age, PCS, RMDQ) of treatment response were identified for the economic analysis. However, the relationship of the QALY with the moderators differed in some cases to that of the clinical outcome measures. For the short-term clinical outcome of PCS, the age-by-treatment interaction was found to be negative and significant ($p < 0.2$), suggesting that younger patients had a better treatment effect. For the outcome of FFbHR, the age-by-treatment interaction was also negative but was just outside the significance threshold of $p < 0.2$. For the other included clinical outcomes, age was not significant. When the QALY was used as the outcome measure, the age-by-treatment interaction was significant at $p < 0.2$ but the relationship was positive, indicating that older patients had a better treatment effect. The EQ-5D at short-term follow-up also exhibited a positive relationship with age, although this relationship was not significant. It may not be surprising that the relationship of the moderators with the different outcomes differed, as they measure different aspects of patient health. Furthermore, the QALY differs by construction from the other outcome measures, as it is calculated as the AUC for a sequence of follow-up points. However, it is also possible that the results are susceptible to missing data bias. Patients with missing EQ-5D data at one or more follow-up points were on average 4 years younger than patients with complete EQ-5D data ($p < 0.05$). One could speculate that younger patients with better expected outcomes might have been excluded from our complete case analysis, as they failed to return follow-up questionnaires. This could bias the treatment response down for younger patients. Four trials had short-term EQ-5D data, comprising 1774 patients (1271 intervention, 503 control) for which there were complete data. Of the 1774 patients, 1467 (1093 intervention, 374 control) had complete data at all of the EQ-5D follow-up points that were necessary to calculate a QALY estimate. This equates to an additional 17% missing data for QALYs compared with short-term outcomes. This might possibly explain the difference in direction of relationship between age and treatment response by outcome measure, as the short-term measures were less prone to missing data than the QALY.

Chapter 10 Methodology and statistical developments 4: subgroup identification with individual participant data indirect network meta-analysis

Background

The recursive partitioning and adaptive peeling approaches described in our analysis plan, although technically of a high standard, failed to identify clinically useful subgroups for whom treatment choices might be prioritised. We therefore also did an exploratory network meta-analysis (NWMA) to identify groups that may gain the greatest benefit from different treatment choices from a Bayesian, rather than a frequentist, perspective.

Methods

We carried out NWMA of the repository trials to explore how the optimal choice of treatment for LBP might vary across subgroups. NWMA is an extension of standard pairwise meta-analysis, applicable in situations in which we have multiple treatments and an evidence base of trials that individually provide evidence on different subsets of all possible pairwise treatment combinations.¹⁷⁴ NWMA involves analysing this network as a whole, by assuming consistency across treatment effects, so that a given pairwise comparison B against C can be derived from trials against a common comparator (A vs. B and A vs. C trials), even if no B versus C trials exist.¹⁷⁵ NWMA has become increasingly popular in decision-making contexts because choosing among more than two treatments requires all pairwise treatment effects to be consistent in this way (the true treatment effects in the decision problem will always be consistent^{176,177}). Given their widespread use in Health Technology Assessment, NWMA commonly uses aggregate data, although there are examples illustrating the value of this approach when IPD are available, particularly in understanding participant-level effect modification.^{178,179}

The standard model for pairwise meta-analysis involving a continuous normally distributed outcome with linear effect modification can be written as equations:

$$y_{it} \sim \text{Normal}(\mu_{it} + \Delta_{it}, \sigma_t) \quad (3)$$

$$\Delta_{it} = I_{it}(d_t + \beta_t(X_{it} - \bar{X})), \quad (4)$$

in which y_{it} is the outcome for participant i in trial t , μ_{it} is the expected outcome for participant i if he/she had been given the control treatment for that trial, Δ_{it} is the expected impact of the treatment participant i received, I_{it} takes value '0' if participant i is in the control arm of trial t and '1' if they are in the intervention arm, d_t is the impact of the intervention for a reference participant, X_{it} is a vector of covariate values for participant i , \bar{X} is a vector of covariate values for the reference participant, and β_t is a vector of coefficients determining how the effect of the intervention evaluated in trial t varies as a function of the covariates of interest. It is possible to further allow for μ_{it} to vary by participants, as shown by Equation 5:

$$\mu_{it} = \mu_t + b(X_{it} - \bar{X}), \quad (5)$$

where μ_t is the expected outcome in the control arm of trial t for the reference participant, and b is a vector of coefficients determining how the control outcome varies as a function of the covariates of interest.

Network meta-analysis extends this analysis by introducing the consistency assumption as shown by Equation 6:

$$d_t = d_{1,active(t)} - d_{1,control(t)}, \quad (6)$$

in which $d_{1,j}$ is defined as the treatment effect of any treatment j in the network compared with a reference treatment (such as standard care), and $active(t)$ and $control(t)$ are the active and control treatments in trial t , respectively. The consistency assumption can further be applied to the β_t parameters as shown by Equation 7:

$$\beta_t = \beta_{1,active(t)} - \beta_{1,control(t)}. \quad (7)$$

We carried out three separate NWMA for the outcomes of interest – short-term change in RMDQ score, short-term change in PCS of SF-12/36 and short-term change in MCS of SF-12/36. All models explore age, sex and baseline PCS/MCS as covariates for both control outcome variation and effect modification. RMDQ models also include baseline RMDQ score for both adjustments. Model estimation involved Bayesian Markov Chain Monte Carlo simulation carried out using WinBUGS 1.4.3 (MRC Biostatistics Unit, Cambridge, UK), using NWMA models that were adapted for IPD analysis from aggregate-data NWMA models that were developed for NICE.¹⁸⁰

Results

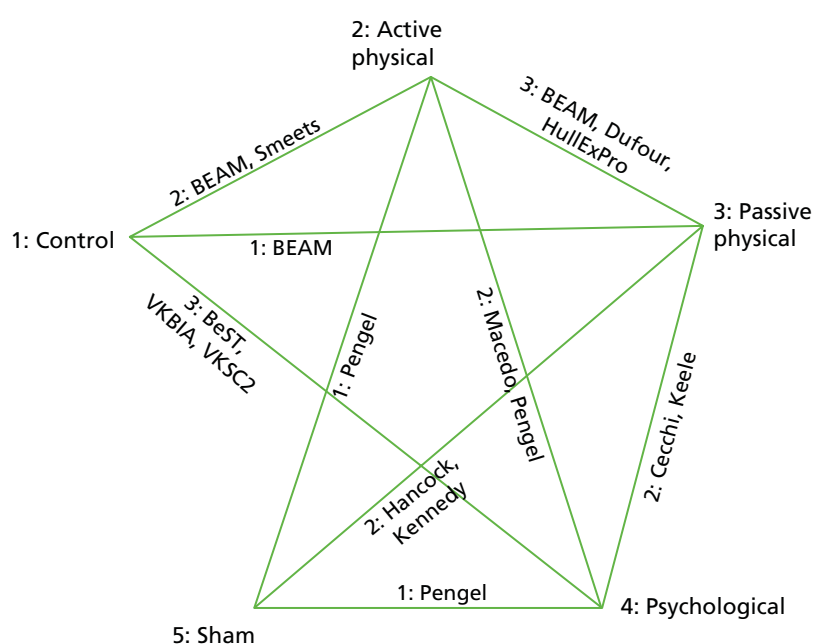
Short-term Roland–Morris Disability Questionnaire outcome

Thirteen trials^{31,33,65,70,76,102–106,131,134,136} ($n = 3447$) in the repository reported this outcome.

The resulting network of evidence is illustrated in Figure 48.

Table 44 gives the predicted treatment effects from the NWMA of these trials for any pairwise comparison of the five treatment classes in the network, assuming a participant profile representing a typical (male) participant. This shows that, for the paradigmatic case of a male aged 50 years, with baseline values of RMDQ = 10, PCS = 40 and MCS = 40, all treatment choices are superior to usual care control treatment. For sham treatment, however, the point estimate for the 95% credible interval for RMDQ does include zero. In addition, the differences between any two treatment approaches can be estimated. For example, in this paradigmatic case there does not seem to be a meaningful difference between sham treatment and psychological treatment.

Table 45 presents coefficient values reflecting the degree of effect modification for the participant characteristics of interest. The evidence for effect modification appears strongest for RMDQ; it is the only characteristic whose coefficient credible intervals for all three treatment serum interventions exclude zero; for sham treatment it does include zero. This analysis suggests that for each 1-point increase in baseline RMDQ score, an additional 0.17- to 0.26-point benefit from active treatments and a 0.43-point benefit from sham treatment will be achieved. However, the 95% credible intervals suggest that the evidence for effect modification related to other covariates is less strong. To quantify the strength of evidence for effect modification, we calculated ‘Bayesian Probabilities of effect modification’ (BP), defined as the greater of two probabilities: that an increase in the characteristic predicts an increase in treatment effect or that it predicts a decrease. A BP of 0.8, for example, suggests that we are 80% sure that a change in the characteristic will increase the effect of treatment. For RMDQ score, the BPs are all > 0.99 (except for sham, with a BP of 0.92) – overwhelming evidence that the effect of treatment depends on baseline scores.



Network of evidence: RMDQ

FIGURE 48 Network of evidence for the short-term RMDQ outcome. Each line denotes the existence of head-of-trials of the two treatments being connected, and the accompanying information denotes the number and names of trials making the comparison.

TABLE 44 Treatment effect with modification (absolute reduction in the short-term RMDQ outcome, mean and 95% credible interval). Coefficients given for individual aged 50 years, male, RMDQ score = 10, PCS = 40 and MCS = 40 at baseline^a

Intervention	Comparator			
	Control	Active physical	Passive physical	Psychological
Active physical	1.94 (1.17 to 2.72)			
Passive physical	2.17 (1.39 to 2.95)	0.23 (−0.61 to 1.07)		
Psychological	1.45 (0.74 to 2.15)	−0.49 (−1.31 to 0.32)	−0.72 (−1.52 to 0.08)	
Sham	1.60 (−1.07 to 4.11)	−0.34 (−2.95 to 2.1)	−0.57 (−3.2 to 1.9)	0.15 (−2.47 to 2.63)

^a Predicted change in condition without treatment adjusted for age, sex and baseline values of RMDQ score, SF-12/36 PCS and SF-12/36 MCS.

TABLE 45 Means, 95% credible intervals and BPs (%) for impact of participant characteristics on effect of treatments (vs. control)

Participant characteristics	Active physical	Passive physical	Psychological	Sham
Age ^a	−0.02 (−0.05 to 0.02) BP = 0.83	0.00 (−0.03 to 0.03) BP = 0.60	−0.02 (−0.05 to 0.01) BP = 0.91	−0.01 (−0.08 to 0.07) BP = 0.56
Sex ^b	−0.22 (−1 to 0.56) BP = 0.71	−0.38 (−1.16 to 0.4) BP = 0.83	−0.01 (−0.78 to 0.77) BP = 0.51	−1.12 (−2.74 to 0.49) BP = 0.91
RMDQ ^a	0.18 (0.06 to 0.31) BP > 0.99	0.26 (0.14 to 0.39) BP > 0.99	0.17 (0.05 to 0.29) BP > 0.99	0.43 (−0.11 to 0.93) BP = 0.92
MCS ^a	−0.01 (−0.06 to 0.05) BP = 0.59	0 (−0.05 to 0.05) BP = 0.51	0.03 (−0.03 to 0.08) BP = 0.85	−0.06 (−0.35 to 0.24) BP = 0.59
PCS ^a	0.05 (−0.03 to 0.13) BP = 0.89	0.04 (−0.04 to 0.12) BP = 0.84	0.03 (−0.04 to 0.11) BP = 0.81	−0.04 (−0.53 to 0.41) BP = 0.52

a Positive value indicates greater reduction in RMDQ from treatment (vs. control) as covariate increases.

b Positive value indicates greater reduction in RMDQ from treatment (vs. control) for females vs. males.

The BPs indicate some, possibly important, differences in benefit by other baseline variables. For example, it is at least 70% likely that men respond more strongly than women to sham treatments and physical treatment but it is equally likely that men respond more or less strongly than women following psychological treatments. On the other hand, baseline MCS has a BP of 85% of positively influencing response to psychological treatments (i.e. those with low levels of psychological distress respond more strongly to psychological treatments than those with high levels of psychological distress), but is almost equally likely to be positively or negatively related to outcomes following physical treatments or sham treatment.

All treatment effects increase, but at different rates, so that the optimal treatment changes as RMDQ score varies. Passive physical therapy is the optimal therapy for the participant as described in *Table 45*, whose RMDQ score is 10. However, sham therapy becomes the optimal treatment if the RMDQ score increases beyond 14 points, whereas APT becomes optimal if the RMDQ score decreases below 7 points.

These thresholds depend on values for other effect modifiers, although their influence is less certain. The only other characteristics with a BP of > 0.90 are age (psychological therapy) and sex (sham therapy). There is evidence, albeit inconclusive, that, as age decreases, active physical and psychological therapies are relatively more effective. *Figures 49* and *50* show how this relationship can be used to define age–RMDQ zones in which each treatment is optimal. Broadly speaking, passive physical therapy is optimal for older participants with a mild to moderate RMDQ score at baseline, APT is optimal for participants with a low RMDQ score at baseline, and sham therapy is optimal for participants with a high RMDQ score at baseline. If we disregard sham treatments as an inappropriate choice for clinical guidelines, passive physical therapies would be optimal for all but the youngest participants with high RMDQ baseline scores (the division would be determined by extending the active–passive equal line into the right-hand side of the graphs). There are no participant profiles for which no intervention is the optimal treatment.

To quantify the strength of evidence for these optimal zones, we calculated the probability that each treatment is optimal for a representative participant profile in each zone. The results (*Table 46*) show that there is considerable uncertainty around the optimal treatment: participant profile 1, for example, is in the passive physical optimal zone, but there is a 54% chance that this is not the optimal treatment for this profile. However, suboptimal treatments can be identified with a greater degree of certainty: psychological therapies, for example, are highly unlikely to be optimal for older participants, or those with a high RMDQ score at baseline (i.e. participant profiles 1, 3, 4 and 6).

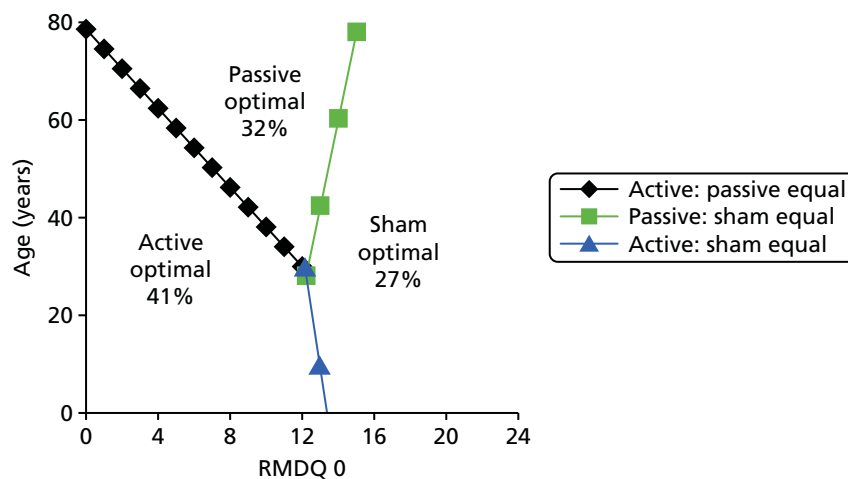


FIGURE 49 Roland-Morris Disability Questionnaire outcome: optimal treatment as a function of RMDQ score at baseline and age for men with MCS = PCS = 40, with proportion of male trial participants whose baseline RMDQ score and age fit into each zone ($n = 721$).

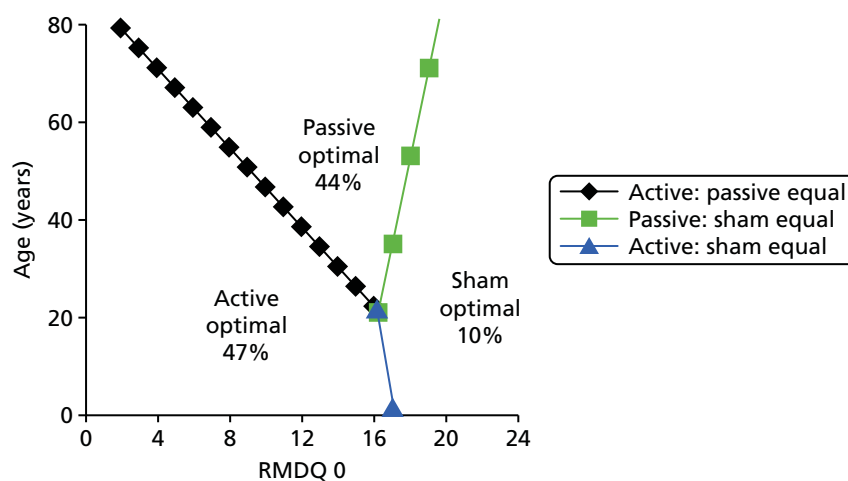


FIGURE 50 Roland-Morris Disability Questionnaire outcome: optimal treatment as a function of RMDQ score at baseline and age for women with MCS = PCS = 40, with proportion of female trial participants whose baseline RMDQ score and age fit into each zone ($n = 1054$).

TABLE 46 Probability that any given treatment is optimal for a range of participant profiles

Participant profile	Probability (%) that treatment is optimal for this participant profile			
	Active physical	Passive physical	Psychological	Sham
1. Male, RMDQ score of 10, age 50 years	18	46	< 1	35
2. Male, RMDQ score of 6, age 30 years	57	11	19%	13
3. Male, RMDQ score of 16, age 40 years	8	34	< 1	57
4. Female, RMDQ score of 14, age 50 years	11	46	2	41
5. Female, RMDQ score of 10, age 30 years	53	14	27	6
6. Female, RMDQ score of 20, age 40 years	8	35	2	54

Short-term Short Form questionnaire-12 items/-36 items physical component summary outcome

Nine trials^{31,33,50,76,101,102,107,132,134} ($n = 5574$) in the repository reported this outcome. The resulting network of evidence is illustrated in Figure 51.

Table 47 gives the predicted treatment effects from the NWMA of these trials for any pairwise comparison of the five treatment classes in the network, assuming a participant profile representing a typical (male) participant. Table 48 presents coefficient values reflecting the degree of effect modification for the participant characteristics of interest. All characteristics, except for age, have at least one effect modification coefficient with a BP of > 0.95 .

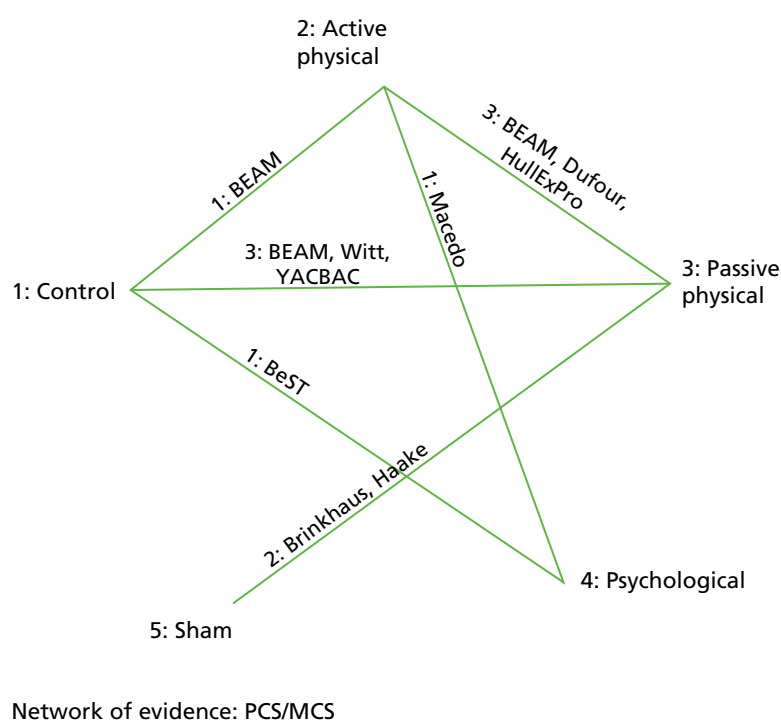


FIGURE 51 Network of evidence for short-term PCS. Each line denotes the existence of head-of-trials of the two treatments being connected, and the accompanying information denotes the number and names of trials making the comparison.

TABLE 47 Treatment effect with modification (absolute increase in short term PCS, mean and 95% credible interval). Coefficients given for individual aged 50 years, male, PCS and MCS = 40, Predicted change in condition without treatment adjusted for age, sex, MCS

Intervention	Comparator			
	Control	Active physical	Passive physical	Psychological
Active physical	3.93 (2.55 to 5.32)			
Passive physical	3.16 (2.4 to 3.92)	-0.77 (-2.13 to 0.58)		
Psychological	2.58 (0.85 to 4.29)	-1.36 (-3.36 to 0.63)	-0.58 (-2.33 to 1.18)	
Sham	1.64 (-0.03 to 3.32)	-2.29 (-4.33 to -0.25)	-1.52 (-3.18 to 0.15)	-0.93 (-3.23 to 1.38)

TABLE 48 Means, 95% credible intervals and BPs (%) for impact of participant characteristics on effect of treatments (vs. control)

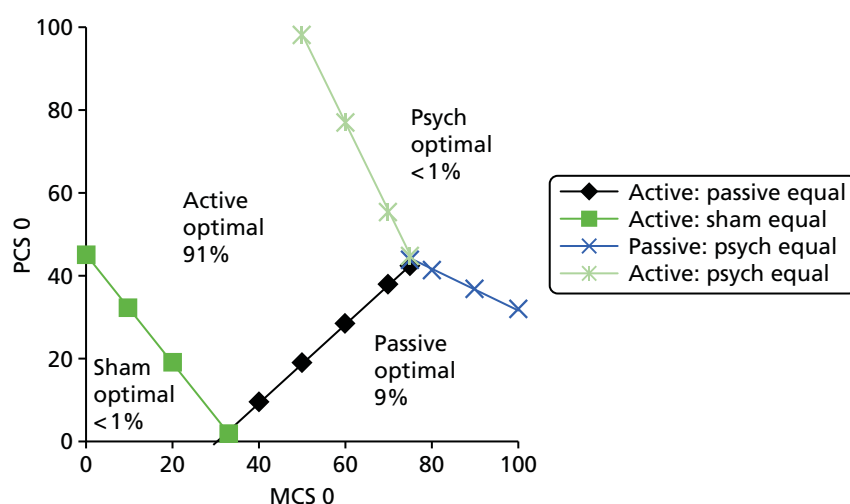
Participant characteristics	Active physical	Passive physical	Psychological	Sham
Age ^a	0.02 (−0.05 to 0.08) BP = 0.68	−0.01 (−0.04 to 0.03) BP = 0.71	−0.04 (−0.1 to 0.03) BP = 0.87	0.00 (−0.06 to 0.06) BP = 0.52
Sex ^b	0.25 (−1.25 to 1.75) BP = 0.63	0.95 (0.04 to 1.87) BP = 0.98	0.29 (−1.43 to 2.01) BP = 0.63	1.55 (−0.15 to 3.23) BP = 0.96
MCS 0 ^a	−0.01 (−0.07 to 0.06) BP = 0.59	0.01 (−0.02 to 0.05) BP = 0.76	0.03 (−0.04 to 0.11) BP = 0.80	−0.07 (−0.14 to 0.00) BP = 0.97
PCS 0 ^a	−0.05 (−0.15 to 0.05) BP = 0.85	−0.07 (−0.13 to −0.02) BP > 0.99	−0.03 (−0.13 to 0.06) BP = 0.76	−0.10 (−0.22 to 0.02) BP = 0.95

a Positive value indicates greater increase in PCS from treatment (vs. control) as covariate increases. MCS 0 and PCS 0 are baseline scores for MCS and PCS, respectively.

b Positive value indicates greater increase in PCS from treatment (vs. control) for females vs. males.

Figures 52 and 53 show how effect modification can be used to define PCS/MCS zones in which each treatment is optimal with short-term PCS as the outcome of interest. Broadly speaking, passive physical therapy is optimal for participants with low PCSs and high MCSs, whereas APT is optimal for participants with high PCSs and low MCSs. Sham appears optimal for participants with low PCSs and MCSs at baseline. If we disregard sham as a valid optimal treatment, the optimal non-sham treatment zones can be identified by extending the active–passive equal line, as with the RMDQ-based zones. Again, there are no participant profiles for which no intervention is optimal.

To quantify the strength of evidence for these optimal zones, we calculated the probability that each treatment is optimal for a representative participant profile in each zone.

**FIGURE 52** Physical component score outcome: optimal treatment as a function of MCS and PCS at baseline for men aged 50 years, with proportion of male participants whose MCSs and PCSs at baseline fit into each zone ($n = 2296$).

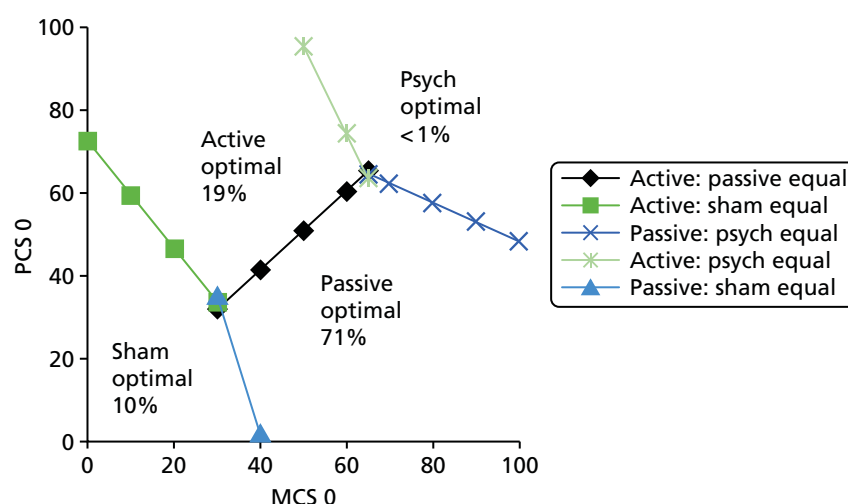


FIGURE 53 Physical component score outcome: optimal treatment as a function of MCS and PCS at baseline for women aged 50 years, with proportion of female participants whose MCSs and PCSs at baseline fit into each zone ($n = 3278$).

The results (*Table 49*) show that, as with the RMDQ score, there is greater certainty around which treatments are suboptimal than around which treatments are optimal. For the paradigmatic cases in *Figures 52* and *53*, it is unlikely that psychological treatments would be the best choice for either gender, but there is a clear indication that there might be differences in proportions who might benefit from active or passive physical treatments if PCS/MCS and sex were the only parameters used for decision-making.

Short-term Short Form questionnaire-12 items/-36 items mental component score outcome

The network of evidence for this outcome is the same as for the SF-12/36 PCS. *Table 50* gives the predicted treatment effects from the NWMA of these trials for any pairwise comparison of the five treatment classes in the network, assuming a participant profile representing a typical (male) participant. *Table 51* presents coefficient values reflecting the degree of effect modification for the participant characteristics of interest. All characteristics, except for sex, have at least one effect modification coefficient with a BP of > 0.95 . It is, perhaps, worth noting here that, for short-term MCS as an outcome, passive physical therapy has the largest effect size for our paradigmatic case. At least for the comparison with active physical, the 95% credibility interval does not cross zero.

TABLE 49 Probability that any given treatment is optimal for a range of participant profiles with PCS as outcome of interest

Participant profile	Probability (%) that treatment is optimal for this participant profile			
	Active physical	Passive physical	Psychological	Sham
1. Male, MCS 40 and PCS 40	81	11	7	< 1
2. Male, MCS 70 and PCS 20	42	43	15	< 1
3. Female, MCS 30 and PCS 50	55	18	6	21
4. Female, MCS 60 and PCS 30	23	68	9	< 1
5. Female, MCS 20 and PCS 20	20	11	1	68

TABLE 50 Treatment effect with modification (absolute change in short-term MCS, mean and 95% credible interval). Coefficients given for individual aged 50 years, male, PCS and MCS = 40. Predicted change in condition without treatment adjusted for age, sex, baseline values of SF-12/36 PCS and MCS

Intervention	Comparator			
	Control	Active physical	Passive physical	Psychological
Active physical	1.53 (0.04 to 3.02)			
Passive physical	3.04 (2.23 to 3.85)	1.50 (0.05 to 2.96)		
Psychological	2.59 (0.80 to 4.39)	1.06 (−1.04 to 3.17)	−0.44 (−2.26 to 1.39)	
Sham	2.13 (0.44 to 3.82)	0.60 (−1.53 to 2.73)	−0.90 (−2.59 to 0.79)	−0.46 (−2.83 to 1.90)

TABLE 51 Mean, 95% credible intervals and BPs (%) for impact of participant characteristics on effect of treatments in the network

Participant characteristics	Active physical	Passive physical	Psychological	Sham
Age ^a	−0.02 (−0.09 to 0.05) BP = 74	−0.03 (−0.07 to 0.01) BP = 93	0.00 (−0.06 to 0.07) BP = 53	−0.09 (−0.15 to −0.03) BP > 99
Sex ^b	0.36 (−1.23 to 1.96) BP = 67	−0.20 (−1.18 to 0.78) BP = 66	−0.47 (−2.26 to 1.34) BP = 70	0.73 (−0.99 to 2.44) BP = 63
MCS ^a	−0.06 (−0.13 to 0.01) BP = 97	−0.10 (−0.14 to −0.06) BP > 99	−0.05 (−0.13 to 0.03) BP > 90	−0.17 (−0.24 to −0.09) BP > 99
PCS ^a	−0.03 (−0.13 to 0.08) BP = 68	−0.08 (−0.14 to −0.02) BP > 99	0.05 (−0.04 to 0.15) BP > 86	−0.15 (−0.27 to −0.03) BP > 99
^a Positive value indicates greater increase in MCS from treatment (vs. control) as covariate increases. ^b Positive value indicates greater increase in MCS from treatment (vs. control) for females vs. males.				

Figures 54 and 55 show how effect modification can be used to define PCS/MCS zones in which each treatment is optimal. Broadly speaking, psychological therapy is optimal for participants with high PCSs (low levels of disability) and moderate to high MCSs (low levels of psychological distress). Passive physical therapy is optimal for participants with low PCSs and high MCSs, and sham therapy is optimal for participants with low PCSs and MCSs (high disability and high levels of psychological distress). If we disregard sham as a feasible recommendation, passive physical therapy becomes optimal for these participants (there are no participant profiles for which no intervention is optimal). To quantify the strength of evidence for these optimal zones, we calculated the probability that each treatment is optimal for a representative participant profile in each zone. The results (Table 52) show that, as with the RMDQ score, there is greater certainty around which treatments are suboptimal than around which treatments are optimal. However, the evidence for effect modification appears strongest on this outcome. It is perhaps of note that for some participant groups (those with high disability and high levels of psychological distress) it appears that sham treatment is highly likely to be the most effective option.

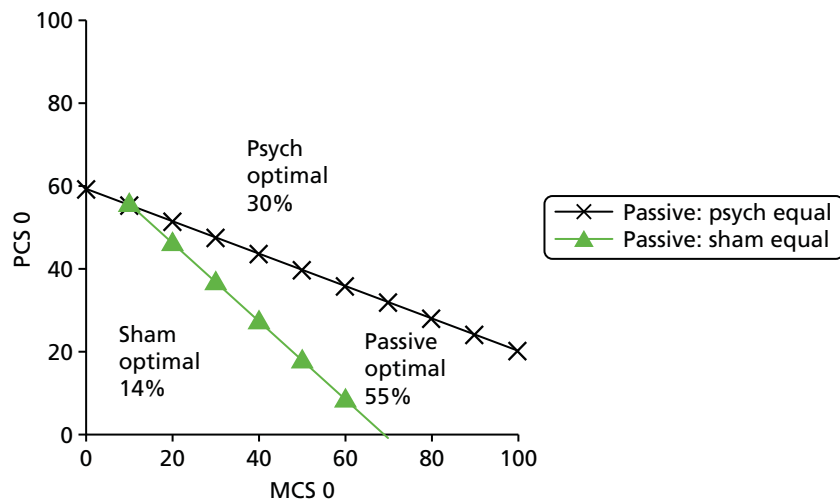


FIGURE 54 Mental component score outcome: optimal treatment as a function of MCS and PCS at baseline for men aged 50 years, with proportion of male participants whose MCSs and PCSs at baseline fit into each zone ($n = 2296$).

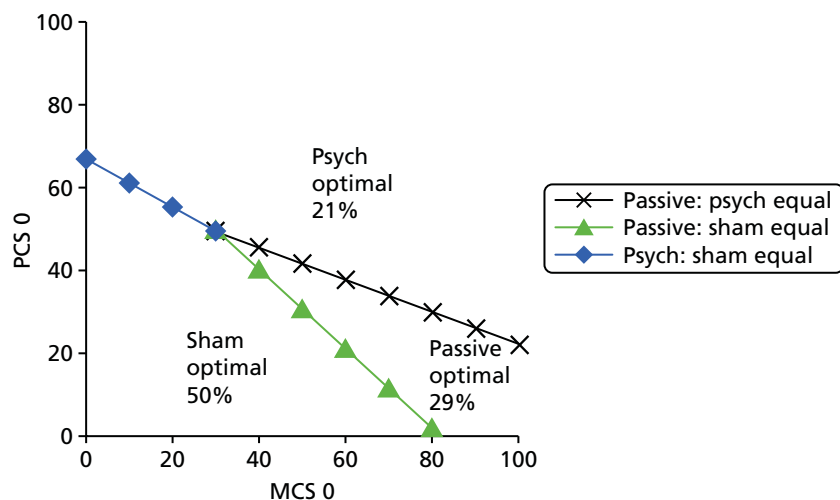


FIGURE 55 Mental component score outcome: optimal treatment as a function of MCS and PCS at baseline for women aged 50 years, with proportion of female participants whose MCSs and PCSs at baseline fit into each zone ($n = 3278$).

TABLE 52 Probability that any given treatment is optimal for a range of participant profiles

Participant profile	Probability (%) that treatment is optimal for this participant profile			
	Active physical	Passive physical	Psychological	Sham
1. Male, MCS 60 and PCS 60	6	< 1	91	< 1
2. Male, MCS 70 and PCS 20	11	65	13	10
3. Male, MCS 30 and PCS 30	< 1	31	< 1	68
4. Female, MCS 60 and PCS 60	12	< 1	82	< 1
5. Female, MCS 80 and PCS 20	26	32	15	11
6. Female, MCS 80 and PCS 20	< 1	13	< 1	87

Chapter 11 Discussion

Introduction

This work is grounded in the pressing need to improve the outcomes for people living with LBP. The targeting of treatments of proven but modest average effectiveness at those who are likely to gain the greatest benefit holds promise. The driver for this research is the considerable uncertainty over which patients are most likely to benefit from which treatment strategy. Improved matching of patients to individual treatments has the potential to improve the overall health gain from, and cost-effectiveness of, treatments for LBP. In particular, how individual patient factors, including duration and severity of the back pain, and physical, social and psychological factors, might affect both adherence and treatment response. There is much published work on predictors of poor outcome for people with LBP, for example the psychosocial 'yellow flags'¹⁸¹ or the STarT Back tool.¹⁸² None, of this work has, however, addressed how these risk factors affect response to treatment. Without explicitly addressing if a particular patient characteristic moderates treatment outcome, targeting treatments at those who are perceived to be at high risk may not be an appropriate choice. During this programme of work we have explored in considerable detail – in two systematic reviews – what is already known about identifying subgroups of people with LBP. This work has demonstrated that the existing work to identify subgroups of patients with LBP within RCTs is generally of a poor methodological quality, and even the high-quality studies do not present evidence to support treatment choices at an individual patient level. Importantly, in this work we have moved beyond using data from single trials and use of single parameters to define subgroups. A large focus of this work has been very technical, on how best to address the challenge of pooling very complex data sets and how best to define subgroups using multiple parameters. To do this we made a series of methodological developments, including three novel methods for subgroup identification: two algorithmic approaches (recursive partitioning, and adaptive risk group refinement) and individual participant data indirect NWMA.

Within the limits of the data that were suitable for pooled analysis, we have identified exploratory subgroups of people who might gain a greater benefit from different treatment approaches in a consistent manner. Interestingly, the groups that we identified as possibly gaining greater benefit from therapist-delivered interventions rather than usual care were typically the converse of expectations. So far as the evidence goes, it seems that younger people with less psychological distress are likely to gain the greatest benefit from these treatments. Although the findings are not strong enough to support these as parameters to prioritise treatment, they do challenge conventional wisdom that people with psychological distress should be targeted for treatment.

Summary of key findings

Systematic reviews (see Chapter 2)

Notwithstanding the perceived importance of performing research to identify subgroups of people living with chronic LBP, there is a paucity of high-quality research in this area. We have identified that nearly all papers reporting analyses of subgroup effects provide no more than exploratory evidence, and that only one study reporting treatment moderation was adequately powered for this analysis. Although it is the identification of differential subgroup effects that is of interest, we failed to identify any robust research that considered subgroups defined by multiple parameters. Rather, we found studies that tested the effect of single potential effect moderators. We have previously found that the available data do not support the use of clinical prediction rules in the management of LBP.⁹³

Age, employment status, education level, back pain status, narcotic use, treatment expectations, moderated treatment effect with $p < 0.05$ in one or more study. The exploratory nature of nearly all of the comparisons, the inconsistent findings across the four included studies and the large number of comparisons made mean that these findings cannot, in themselves, be used to inform management. Notwithstanding the limitations of the existing research we were able to identify some potential moderators to include in our final analyses. The overall weakness of the underpinning data meant that we included potential moderators in our analyses that did not meet conventional criteria for statistical significance. By including moderators found to be significant at the 20% level, our pool of potential moderators became age, gender, employment status, education, back pain status, pain-related disability, narcotic use, treatment expectations, quality of life and psychosocial status.

Analyses of covariance (see Chapter 6)

Our ANCOVAs replicate the conventional approach to moderator identification in a pooled data set. The main purpose of these analyses was to inform selection of potential moderators for our main analysis based on identifying variables significant at the 20% level. In our analyses, we were restricted by the pool of trials using a common set of baseline covariates and outcomes. In this analysis, comparing all intervention groups with all control groups ('non-active usual care plus sham for clinical outcomes' or 'usual care for health-economic outcomes'), we identified some moderators that reached conventional statistical significance for some outcomes. Summarising these findings, these data suggest that those who are worse on a measure of physical function (FFbHR/SF-12/36 PCS) have the most to gain from treatment on physical outcomes and those who are worse on the SF-12/36 MCS at baseline gain the most on this outcome measure. For the outcome of EQ-5D, its baseline value did not moderate treatment response, but pain, physical function (SF-12/36 PCS) and anxiety, which are arguably components of the EQ-5D, did moderate response. The exception to the observation that it is severity at baseline that predicts response to treatment on that measure is that a less favourable baseline FFbHR score moderates outcome on the SF-12/36 MCS. Anxiety – but not catastrophising, coping strategies and depression – moderated treatment response, at $p < 0.05$ in the analyses for the outcome of EQ-5D in which those with lower risk of anxiety had less treatment effect than those with higher risk of anxiety. This is the first meta-analysis to assess effect moderation in the treatment of LBP and hence gives a far more robust assessment than any previous work in this area. The numbers in our analyses mean that if there were true moderation effects in this comparison of all treatments against control then they should have been identified.

Although these observations are of some interest, the main purpose of these analyses was to select potential moderators that were significant at the 20% level to take forward for our main analyses. We were able to take forward FFbHR, RMDQ, SF-12/36 PCS and MCS, age, gender, pain, fear avoidance and coping as variable with a possible signal in one or more analysis.

Recursive partitioning (see Chapter 7)

We successfully adapted two recursive partitioning approaches to identify subgroups in an individual participant data meta-analysis. There are important distinctions in the way they work. The IPD-IT method is seeking to maximise the size of the interaction term when making splits, whereas the IPD-SIDES method is seeking to detect groups with the largest treatment effects.¹⁶⁷ The choice of approach in any future analyses using a recursive partitioning approach will depend on the primary outcome of interest. For our current purpose we prefer the IPD-SIDES approach, as we think it is more likely to identify clinically useful subgroups with large effect sizes. The IPD-IT approach may be more suitable for more exploratory analyses for which maximising any moderation is the outcome of interest. We have presented both analyses here to explore how they perform on a real data set. The IPD-SIDES approach appears to be more sensitive, as it has successfully identified some subgroups within our data, whereas the IPD-IT method did not (Tables 53–57 and see Chapter 7). Our overall analysis of all interventions compared with control (usual care or sham control) provides evidence that the IPD-SIDES method functions well; we found candidate subgroups in a real data set, as well as the simulation in which it was originally tested. For the choice of treatment compared with control (sham plus usual care) using the full data set, there are some clusters of characteristics with different treatment outcomes. For example, for the outcome FFbHR (range of the score

TABLE 53 Overview of results: intervention vs. control (usual care or sham)

OUTCOME ^a							
METHOD (section)	Physical health		Pain: average pain	Mental health: MCS	Quality of life		
	FFbHR	RMDQ			PCS	EQ-5D	QALY ^b
ANCOVA (see Chapter 6, <i>Analyses of covariance</i>)	Positive moderator	None found	Moderate catastrophising; ^c positive fear avoidance ^c	None found	Pain; ^c MCS; ^c moderate fear avoidance ^c	Female; ^c RMDQ ^d pain; ^d moderate fear avoidance ^d	Age; ^c RMDQ ^c
	Negative moderators	Age; ^c FFbHR; ^d PCS ^d	None found	Age; ^c PCS; ^d MCS < 50 ^c	PCS; ^c low anxiety; ^c positive coping ^c	FFbHR; ^d MCS ^d	PCS ^c
Recursive partitioning IPD-SIDES See Chapter 7 (<i>Subgroups identified by the IPD-SIDES method</i>)	Subgroups	Younger with worse FFbHR	None found	1. Better MCS and worse PCS 2. Female with worse PCS	None found	Worse MCS	None found
Directed search^e See Chapter 8 (<i>Analysis 1: Overall comparison treatment compared with control</i>) and Chapter 9 (<i>Results</i>)	Subgroups	1. Younger with worse FFbHR 2. Younger with worse PCS	None found	None found	None found	Younger with worse MCS	1. Older with worse RMDQ 2. Older with worse PCS
a All outcomes measured as change from baseline at short-term follow-up (2–3 months), except for QALY, which is the area under curve for EQ-5D over 12 months.							
b Sham interventions not included in QALY analyses.							
c Variables with $p < 0.2$ and $p > 0.05$ for interactions with treatment effect (FFbHR ANCOVA age $p = 0.2018$).							
d Variables with $p < 0.05$ for interactions with treatment effect.							
e Directed searches to identify subgroups of decreasing size with better expected response to treatment, assuming monotonicity of moderator effects on outcomes. A full search algorithm was used for the short-term outcomes (FFbHR, RMDQ, PCS, pain, MCS, EQ-5D) and a directed peeling search for QALYs.							

^a All outcomes measured as change from baseline at short-term follow-up (2–3 months), except for QALY, which is the area under curve for EQ-5D over 12 months.

^b Sham interventions not included in QALY analyses.

^c Variables with $p < 0.2$ and $p > 0.05$ for interactions with treatment effect (FFbHR ANCOVA age $p = 0.2018$).

^d Variables with $p < 0.05$ for interactions with treatment effect.

^e Directed searches to identify subgroups of decreasing size with better expected response to treatment, assuming monotonicity of moderator effects on outcomes. A full search algorithm was used for the short-term outcomes (FFbHR, RMDQ, PCS, pain, MCS, EQ-5D) and a directed peeling search for QALYs.

TABLE 54 Overview of results: active physical vs. control (usual care)

METHOD (section)	OUTCOME ^a	Physical health					Quality of life	
		FFbHR	RMDQ	PCS	Pain: average pain	Mental health: MCS	EQ-5D	QALY ^b
		None found	None found	None found	None found	None found	None found	None found
Recursive partitioning IPD-SIDES	Subgroups	None found	None found	None found	None found	None found	None found	None found
See Chapter 7 (Subgroups identified by the IPD-SIDES method)								
Directed search^c	Subgroups	None found	None found	None found	None found	None found	None found	None found
See Chapter 8 (Analysis 1: Overall comparison treatment compared with control) and Chapter 9 (Results)								
NWMA^d	Positive moderators	Not conducted	RMDQ; PCS	None found	Not conducted	None found	Not conducted	Not conducted
See Chapter 10 (Results)	Negative moderators	Not conducted	Age	PCS	Not conducted	MCS	Not conducted	Not conducted

^a All outcomes measured as change from baseline at short-term follow-up (2–3 months) except for QALY, which is the area under curve for EQ-5D over 12 months.

^b Sham interventions not included in QALY analyses.

^c Directed searches to identify subgroups of decreasing size with better expected response to treatment, assuming monotonicity of moderator effects on outcomes. A full search algorithm was used for the short-term outcomes (FFbHR, RMDQ, PCS, pain, MCS, EQ-5D) and a directed peeling search for QALYs.

^d Variables with BP of > 0.8 for interactions with treatment effect.

TABLE 55 Overview of results: passive physical vs. usual care control

METHOD (section)	OUTCOME ^a					
	Physical health			Quality of life		
	FFbHR	RMDQ	PCS	Pain: average pain	Mental health: MCS	EQ-5D
Recursive partitioning IPD-SIDES						
Subgroups	Younger with worse FFbHR	None found	1. Younger with worse PCS 2. Worse PCS but better MCS 3. Women with worse PCS and better MCS	None found	Worse MCS and worse PCS	None found
See Chapter 7 (Subgroups identified by the IPD-SIDES method)						
Directed search^c						
Subgroups	Younger with worse FFbHR	Not conducted	None found	Not conducted	Younger with worse PCS and worse MCS	Not conducted
See Chapter 8 (Analysis 1: Overall comparison treatment compared with control) and Chapter 9 (Results)						
NWMA^d						
Positive moderators	Not conducted	Men; RMDQ; PCS	Women	Not conducted	None found	Not conducted
Negative moderators	Not conducted	None found	PCS	Not conducted	Age; PCS; MCS	Not conducted
See Chapter 10 (Results)						

^a All outcomes measured as change from baseline at short-term follow-up (2–3 months) except for QALY, which is the area under curve for EQ-5D over 12 months.

^b Sham interventions not included in QALY analyses.

^c Directed searches to identify subgroups of decreasing size with better expected response to treatment, assuming monotonicity of moderator effects on outcomes. A full search algorithm was used for the short-term outcomes (FFbHR, RMDQ, PCS, pain, MCS, EQ-5D) and a directed peeling search for QALYs.

^d Variables with BP of > 0.8 for interactions with treatment effect.

TABLE 56 Overview of results: psychological vs. usual care control

METHOD (section)	OUTCOME ^a							
	Physical health			Pain: average pain		Mental health: MCS	Quality of life	
	FFbHR	RMDQ	PCS				EQ-5D	QALY ^b
Recursive partitioning IPD-SIDES	Subgroups	None found	Worse RMDQ	None found	None found	None found	None found	None found
See <i>Chapter 7 (Subgroups identified by the IPD-SIDES method)</i>								
Directed search^c	Subgroups	Not conducted	None found	Not conducted	Not conducted	Not conducted	Not conducted	Not conducted
See <i>Chapter 8 (Analysis 1: Overall comparison treatment compared with control) and Chapter 9 (Results)</i>								
NWMA^d	Positive moderators	Not conducted	RMDQ; PCS; MCS	MCS	Not conducted	PCS	Not conducted	Not conducted
See <i>Chapter 10 (Results)</i>	Negative moderators	Not conducted	Age	Age	Not conducted	MCS	Not conducted	Not conducted
^a All outcomes measured as change from baseline at short-term follow-up (2–3 months) except for QALY, which is the area under curve for EQ-5D over 12 months.								
^b Sham interventions not included in QALY analyses.								
^c Directed searches to identify subgroups of decreasing size with better expected response to treatment, assuming monotonicity of moderator effects on outcomes. A full search algorithm was used for the short-term outcomes (FFbHR, RMDQ, PCS, pain, MCS, EQ-5D) and a directed peeling search for QALYs.								
^d Variables with BP of > 0.8 for interactions with treatment effect.								

^a All outcomes measured as change from baseline at short-term follow-up (2–3 months) except for QALY, which is the area under curve for EQ-5D over 12 months.

^b Sham interventions not included in QALY analyses.

^c Directed searches to identify subgroups of decreasing size with better expected response to treatment, assuming monotonicity of moderator effects on outcomes. A full search algorithm was used for the short-term outcomes (FFbHR, RMDQ, PCS, pain, MCS, EQ-5D) and a directed peeling search for QALYs.

^d Variables with BP of > 0.8 for interactions with treatment effect.

TABLE 57 Overview of results: sham vs. control

METHOD (section)	OUTCOME ^a	Physical health					Quality of life	
		FFbHR		Pain: average pain		Mental health: MCS	EQ-5D	
		RMDQ	PCS	None found	None found		None found	QALY
Recursive partitioning IPD-SIDES	Subgroups	None found	None found	None found	None found	Younger with worse PCS	None found	None found
<i>See Chapter 7 (Subgroups identified by the IPD-SIDES method)</i>								
Directed search^b	Subgroups	Younger with either worse FFbHR or PCS	Not conducted	Not conducted	Not conducted	Any age; worse PCS; worse MCS	Not conducted	Not conducted
<i>See Chapter 8 (Analysis 1: Overall comparison treatment compared with control) and Chapter 9 (Results)</i>								
NWMA^c	Positive moderator	Not conducted	Men RMDQ	Women	Not conducted	None found	Not conducted	Not conducted
<i>See Chapter 10 (Results)</i>	Negative moderator	Not conducted	None found	MCS; PCS	Not conducted	Age; PCS; MCS	Not conducted	Not conducted

^a All outcomes measured as change from baseline at short-term follow-up (2–3 months) except for QALY, which is the area under curve for EQ-5D over 12 months.

^b Directed searches to identify subgroups of decreasing size with better expected response to treatment, assuming monotonicity of moderator effects on outcomes. A full search algorithm was used for the short-term outcomes (FFbHR, RMDQ, PCS, pain, MCS, EQ-5D) and a directed peeling search for QALYs.

^c Variables with BP of > 0.8 for interactions with treatment effect.

is from 0 = great limitation to 100 = no limitation) the overall treatment effect of 8.93 (95% CI 7.81 to 10.05) increases to 13.17 (95% CI 10.56 to 15.77) in those with a FFbHR score of ≤ 54.2 and aged ≤ 60 years or for the SF-12/36 PCS (range 0–100 best) the overall treatment effect increases from 3.48 (95% CI 3.01 to 3.96) to 4.89 (95% CI 3.96 to 5.82) in those with a SF-12/36 PCS of ≤ 40.0 and a SF-12/36 MCS of > 54.2 . It is, however, the pairwise comparisons, with usual care control, which might be useable to inform clinical practice.

Passive physical therapy

For passive physical therapy we identified subgroups for the outcomes of FFbHR, plus SF-12/36 MCS/PCS. The results for FFbHR, which represent just acupuncture trials, find a maximal effect of 16.67 (95% CI 13.16 to 20.18) compared with an overall treatment effect of 9.95 (95% CI 8.80 to 11.11) in those aged ≤ 53 years and with a FFbHR score of ≤ 54.2 . Thus acupuncture is likely to be more effective in those with a worse baseline score and who are younger. This finding is probably of little clinical importance, as none of the splits identified a group in which the treatment was ineffective and only 17% of participants (571/3272) were in this group with the largest effect. For the SF-12/36 MCS the maximal effect is seen in those with a low score on both PCS and MCS. In the group with a MCS of ≤ 54.3 and PCS of ≤ 43.9 , the treatment effect increases from 2.96 (95% CI 2.31 to 3.61) to 4.27 (95% CI 3.39 to 5.15). On this occasion, 56% of participants (2171/3898) fall into this group. Again, none of the splits identified a group in which the treatment was not effective, suggesting that it would not be helpful in clinical practice. This could, in any event, be plausibly clinically important only if the outcome of interest was mental health.

For the SF-12/36 PCS, IPD-SIDES found nine candidate models, including one with three splits: baseline PCS, MCS and gender. The final split on gender did not, however, achieve conventional statistical PCS as the first split with either age or MCS as the second split. Treatment was most effective in those with more severe problems and who were younger or had better mental health. There was little to choose from between the added effect from each of the different models with two splits, and no split was found for which the intervention was ineffective. This makes it difficult to suggest a 'best' choice. It is, however, of note that increasing psychological distress appears to make it less likely that passive physical interventions will be effective. This does not support the notion that such treatments should be targeted at those with increased psychological distress.

Active physical therapy

We did not find any subgroups with an enhanced response to active physical therapy.

Psychological therapy

There were fewer participants included in this analysis ($n = 928$) than for passive physical treatments ($n =$ up to 3898), reducing potential for finding subgroups. Nevertheless, the IPD-SIDES method did identify one split for the RMDQ outcome, based on baseline severity as measured using the RMDQ (range 0–24, 0 = best). This split might be of clinical relevance: the 75% (231/928) of participants with a RMDQ score of > 4 gained an additional 1.07 points benefit, taking the average treatment effect from 1.40 (95% CI 0.89 to 1.91) to 1.72 (95% CI 1.12 to 2.31). Furthermore, for the group with a RMDQ score of ≤ 4 the 95% CI for the mean effect included zero (0.65, 95% CI -0.11 to 1.40). This indicates that psychological treatments should be reserved for those with higher RMDQ scores. For the RMDQ score, unlike the other outcome measures reported here, there is an established minimally important change for an individual: 5.0 points.³⁰ The size of the interaction can be interpreted as a small difference, that is, 0.21 of the minimally important change.¹⁸³ It is nevertheless comparable with the overall effect size at 3 months identified in the BeST Trial (1.1 points on the RMDQ, 95% CI 0.38 to 1.71 points), which did not have a lower limit of RMDQ score for study entry.³³ These data can reasonably be used to indicate that psychological treatments should be reserved for those with a RMDQ score of > 4 . Interpretation of the importance of this observation needs to include the important caveat that all of the analyses reported here are exploratory rather than confirmatory. It also fits with the general pattern that treatments tend to have greater effects in those with worse baseline scores on the outcome of interest.

Sham treatment

Interpreting the findings for sham treatments, on this occasion sham acupuncture from two trials^{132,136} on the SF-12/36 MCS is quite challenging. The results of the IPD-SIDES analysis appear to show that for those aged > 65 years, and for those with a SF-12/36 PCS of > 42.0, sham acupuncture is substantially less effective, and that in rest of the population the effect size is enhanced. Although the point estimates indicate harm, the 95% CIs include zero and, at least for SF-12/36, the interaction effect is of borderline statistical significance ($p = 0.043$). It may well be that, for age, we are observing the same phenomena seen for other interventions whereby older people, and those with fewer symptoms, are less likely to benefit. The option of a sham treatment is unlikely to be explicitly offered by the NHS. It could be argued that we do not need to consider this further. On the other hand, any sham intervention includes the potentially very important therapist–patient interaction that is part of all of the interventions we have examined. The differential effects observed might be clinically important in that we have identified subgroups (those aged > 65 years and those with a better PCS > 42.0) who might be harmed by the sham intervention. If this were a true observation it might lead one to question the benefit of offering some therapist-delivered interventions to an older age group or to those with less disability as a consequence of potential adverse effects on their mental health.

Adaptive refinement by directed peeling in individual patient data meta-analysis (see Chapter 8)

We have successfully extended an adaptive risk group refinement method for use in identifying subgroups of patients who may respond better to different treatments. In contrast with the recursive partitioning approaches, adaptive risk group refinement produces multiple solutions representing different-sized proportions of the population, allowing the user to decide at which point on any trajectory plot the additional benefit of selecting subgroups would be clinically worthwhile. This is achieved by repeatedly searching within the data set to identify successively smaller subgroups with larger effects. This approach does not produce the monotonic changes in subgroup specification seen when a peeling approach (see Chapter 9) is used, but may give a better representation of effect for a prespecified size of subgroup.

We were limited, by lack of computational power, to just exploring the effect of four covariates; there is, however, no statistical reason for restricting the covariates used to just four. In this restriction we were able to do a more extensive search by considering all possible combinations of subgroups, thus interrogating the data more thoroughly. It can be seen how this approach can define subgroups in the example of the FFbHR outcome (three acupuncture trials) for all interventions compared with control (usual care and sham) (see Figure 21 and Table 27). Here a clear trajectory, with average effect size increasing from 8.47 to 16.79, is seen. This is largely driven by baseline FFbHR score. In contrast, no such pattern is seen for the RMDQ outcomes (see Figure 25), suggesting that there is not potential for subgroup identification for this group of studies. For the SF-12/36 MCS and PCS outcomes the high variability as subgroup size decreases suggests that it is not possible to define subgroups reliably for these outcomes. Thus for our interpretation of all interventions compared with control (non-active usual care/placebo) is that for the FFbHR outcome younger people with a worse FFbHR score and worse PCS may gain more from treatment and that, for the SF-12/36 MCS outcome, those who are younger and with a worse MCS are likely to gain the greatest benefit. Results from pairwise comparisons between different types of treatment and non-active usual care controls are considered in the following subsections.

Passive physical therapy

We found a similar pattern to the overall comparison for the FFbHR result when passive physical (acupuncture) was compared with non-active usual care, that is, it was more effective for those who were younger with a worse baseline score.

We also found that, for the outcome of SF-12/36 MCS, those who were younger with worse PCS and MCS gained a greater benefit.

Active physical therapy

We did not find any subgroups with an enhanced response to APT. In particular, we did not find that baseline RMDQ score consistently identified subgroups with a better treatment effect.

Psychological therapy

We did not find any subgroups with an enhanced response to psychological therapy.

Sham

We were again able to identify a group that might do better with sham treatment. Its definition was, again, driven by age and baseline severity. Curiously, a worse baseline MCS appears to predict who responds better to sham acupuncture, but not who responds to true acupuncture.

Identification of cost-effective subgroups by direct peeling (see Chapter 9)

The application of the peeling algorithm was successful in identifying potentially interesting subgroups for the interventions against control comparison. These subgroups comprised patients who were older, with relatively worse physical functioning at baseline. The gain in treatment effect for the subgroup was small. Therefore, given the relatively low cost of the intervention, treatment is likely to be cost-effective for the whole patient group. The algorithm, however, was not successful in finding any convincing subgroup in the pairwise comparison of active and passive physical treatment. This may be due to lack of power, or simply that there is no subgroup to be found.

The QALY has some key advantages over the other available clinical outcomes. It is a holistic measure of health-related quality of life, designed to encompass both physical and mental aspects of a patient's health state. Constructed using EQ-5D responses over time, the QALY also takes account of a patient's recovery profile, integrating short- and long-term treatment response into a single measure. The EQ-5D is scored using the UK social tariff; this is validated and standardised, allowing direct comparison of the treatment response for different interventions and diseases. The QALY estimated using the EQ-5D tariff is the accepted measure used by NICE for assessing the cost-effectiveness of new treatments for approval in the NHS. The QALY did, however, raise some particular challenges for the analysis. The use of repeated measures to estimate the QALY restricted the size of the sample, as more observations were lost to missing data when compared with the point estimates used in the clinical analysis. This reduced the power of statistical analyses. For the QALY analyses, the group that had sham treatment were excluded. Although of some interest to explore the effects of sham treatments for clinical outcomes, these are not relevant to an economic analysis.

The same approach was taken for moderator identification for the economic component of the analysis as for the clinical analyses. Three potential moderators (age, PCS, RMDQ score) of treatment response were identified for the economic analysis. However, the relationship of the QALY with the moderators differed in some cases to that of the clinical outcome measures. It was only for the overall comparison of treatment with control that any potential subgroups were identified.

For the short-term clinical outcome of PCS, the age-by-treatment interaction was found to be negative and significant ($p < 0.2$), suggesting that younger patients had a better treatment effect. For the outcome of FFbHR score, the age-by-treatment interaction was also negative but was just outside the significance threshold of $p < 0.2$. For the other included clinical outcomes, age was not significant. When the QALY was used as the outcome measure, the age-by-treatment interaction was significant at $p < 0.2$ but the relationship was positive, indicating that older patients had a better treatment effect. The EQ-5D at short-term follow-up also exhibited a positive relationship with age, although this relationship was not significant. It may not be surprising that the relationship of the moderators with the different outcomes differed, as they measure different aspects of patient health. Furthermore, the QALY differs, by construction from the other outcome measures, as it is calculated as the AUC for a sequence of follow-up points. However, it is also possible that the results are susceptible to missing data bias. Patients with missing EQ-5D data at one or more follow-up points were, on average, 4 years younger than patients with complete EQ-5D data ($p < 0.05$). One could speculate that younger patients with better expected

outcomes might have been excluded from our complete-case analysis, as they failed to return follow-up questionnaires. This could bias the treatment response down for younger patients. Four trials had short-term EQ-5D data, comprising 1774 patients (1271 intervention, 503 control) for which there were complete data. Of the 1774 patients, 1467 (1093 intervention, 374 control) had complete data at all EQ-5D follow-up points necessary to calculate a QALY estimate. This equates to an additional 17% missing data for QALYs compared with short-term outcomes. This might possibly explain the difference in direction of relationship between age and treatment response by outcome measure, as the short-term measures were less prone to missing data than the QALY.

Overall, our interpretation is that those who are older, with worse RMDQ and SF-12/36 PCS are likely to gain a greater benefit on QALY outcomes from treatment. Doing this will not, however, improve overall QALY gain for the whole population, as those outside this subgroup are likely, on average, to benefit from treatment. Treating only this subgroup is very unlikely to be seen as cost-effective, given the relatively low cost of treatment and the NICE threshold of £20,000–30,000 per QALY.

Network meta-analysis (see Chapter 10)

In a further methodological development, we successfully adapted a NWMA approach to identify effect moderators and produce a probability that a particular treatment choice is optimal for individuals with particular profiles. This approach presents the data in a format that is very different from our other approaches to subgroup identification. Analysing the trials as a single network of evidence allows us to detect subgroup effects with greater precision, and the use of Bayesian methods allows quantification of the strength of evidence for alternative modalities. This has allowed us to estimate effect sizes for groups with similar characteristics. See, for example, *Table 44*, which shows that for a paradigmatic case (male, age 50 years, baseline RMDQ score = 10, baseline PCS and MCS both equal 40) active physical, passive physical and psychological treatments are all likely to be effective in reducing RMDQ score compared with control; the credible intervals exclude zero. For sham treatment, the point estimate is consistent with it being effective but the 95% credible interval includes zero. Consistent with the preplanned analyses, baseline severity strongly predicts response to treatment across all interventions (slightly weaker for sham treatment). The effect of age, gender, plus the baseline SF-12/36 PCS and MCS are weaker and are not consistent across modalities. It is this variability that allows tables of probability for a particular treatment choice to be the optimum choice. For our paradigmatic case the probability that passive physical is optimal is 45% and that psychological is optimum is < 1%. These sorts of outputs have the potential to inform clinical decision-making. It should, however, be noted that this approach generates a ranking and that the differences in effect sizes from moderation of the primary outcome by baseline characteristics remains modest. For our paradigmatic case, all treatment options (except sham) have evidence of effectiveness; the 95% credible interval excludes zero. The additional benefit for passive physical treatment over psychological treatment, however, is only 0.72 (95% credible interval –0.08 to 1.52) points on the RMDQ and the 95% credibility interval includes zero. Nevertheless, this approach does have the potential to provide some information, tailored to the individual, which can be used to inform clinical decision-making.

Interpretation

Clinical relevance

In our overall analyses (all interventions compared with control) it appears that women with more severe disability and lower levels of psychological distress are likely to gain the greatest benefit on back pain disability and the PCS of the SF-12/36. For psychological outcomes, as measured by the SF-12/36 MCS, those with poorer baseline psychological health gained the greatest benefit. That those with a less favourable baseline score gain the greatest treatment benefit, on the same measure, may not be surprising, as these are the individuals with the greatest potential for improvement. We have in all of our analyses presented here, and as outlined in our analysis plan, used absolute differences in outcome rather than percentage changes from baseline. In a post hoc analysis we re-ran our initial ANCOVAs, with percentage change from baseline as the dependent variable (data not shown). The apparent significance of any moderator effects was substantially reduced, for example the significance of any moderation of effect of

baseline FFbHR score as the outcome, p -value changed from < 0.0001 to 0.0703 . This suggests that our finding that baseline severity predicts outcome on the same measure might depend on the scale of measurement used for the change.

Our prespecified approaches, recursive partitioning and ARDP did produce identifiable subgroups, the parameter definitions of which were grounded in the data. The differences in effect sizes were generally small, however, and unlikely to be clinically meaningful. The effect sizes in the groups who did less well would still justify the use of these interventions. This overall picture is potentially misleading, however, as the choice is not typically between treatment and no treatment; rather, it is how to select particular treatments for individuals.

Our prespecified analyses give some insights here. For passive physical treatments (acupuncture, manual therapy), those who are younger, with less psychological distress and worse disability were likely to gain the greatest benefit on disability. For psychological treatments, those with more baseline disability were likely to gain a greater benefit on disability. In both of these cases the difference in effect sizes are unlikely to be clinically important. Defining what is clinically important is a challenge for LBP researchers exploring treatment moderation. The authors of the published protocol for an IPD meta-analysis of studies of exercise treatment for LBP have set a minimally clinically important difference for moderation, where the p -value is < 0.05 , to be 20 points on a 100-point scale for pain, and 10 points on a 100-point scale for disability or 'another magnitude deemed clinically important by experts'.¹⁸⁴ Others have argued that, for exercise interventions for LBP, worthwhile between-group differences in pain may be as much as 10 points.¹⁸⁵ None of the subgroups identified by the IPD-SIDES or ARDP-MA method met these criteria.

All of the subgroups identified in this work had quantitative effects when the direction of the treatment effect was in favour of the intervention arm in both subgroups. It is open to debate whether or not a differential subgroup effect that is smaller than a main treatment effect is worthwhile. When the choice is between treatment or no treatment, one might expect that to be clinically meaningful, any moderator effect should be larger than the main effect. Otherwise, as our data show, the overall net benefit from treatment may decrease as it is offered selectively. If the choice is between different treatments with similar main effect sizes, acquisition and opportunity costs, and risk profiles, then quite small moderation effects might increase over treatment effectiveness.

Our health-economic analysis suggests that it is possible to identify groups with better-than-average QALY gain from treatment. Nevertheless, even in the groups with a smaller QALY gain, the incremental cost per QALY gained is sufficiently low that it falls far below the NICE threshold of £20,000. Our analyses show that selecting subgroups of individuals for treatment reduces the overall QALY gain. This means there is not a cost-effectiveness argument for excluding some groups from access to treatments.

On the basis of these analyses we can be confident that the only potentially worthwhile screening tool to select treatments is baseline severity of the measure of interest, although even here those who are less severe will still gain a benefit and we have failed to find evidence that it would be worthwhile offering treatment to selected patients based on baseline severity. We have found that those with higher levels of psychological distress are less likely to benefit from some interventions. Nevertheless, the size of the interaction effect means it is unlikely to serve as a discriminator for selecting treatment approaches as those with *higher* levels of distress may still benefit from treatment.

The importance of these findings is that there is no justification for using higher levels of psychological distress to target treatments. This runs contrary to received wisdom that psychosocial yellow flags could be used to select those who would benefit from treatment.

A RCT of stratified care based on patient prognosis using the STarT Back tool found it to be a very effective and cost-effective approach to managing with LBP.³⁸ The study design, however, did not allow the effect of the stratification tool to be separated out from the effects of therapist selection and the

additional benefits of the customised treatment packages provided after stratification. Thus, although a promising overall approach to targeting back pain treatments, this does not help us to identify differential subgroup effects in this population.

Currently, treatment choices between the types of interventions we have examined here is largely decided by the treating therapist in consultation with the patient. A shared informed decision-making model in which patients are given more information on the evidence for different treatment options and physiotherapists are trained to implement shared informed decision-making does not improve outcomes; indeed it may have an overall harmful effect.¹⁸⁶ An alternative approach of using the output from NWMA to help physiotherapists and their patients choose treatment options could be tested empirically.

Psychological distress as a treatment moderator

That increased psychological distress, as measured by the SF-12/36 MCS, does not appear to increase treatment effect from either passive physical or psychological interventions is an important finding. There is a substantial body of literature suggesting that those with psychological distress should be prioritised for treatment of their LBP because their prognosis is worse.^{187–192} Our data suggest that, for the interventions assessed here, those with higher levels of psychological distress are less likely to benefit. These observations are, of course, limited by the measures we were able to use as potential moderators and that other moderator variables, for example back beliefs or self-efficacy, might have produced different findings. There is some limited evidence ($p < 0.2$) in our overall ANCOVA that catastrophising and fear avoidance might moderate treatment response for the RMDQ outcome when people with a more positive attitude (low scores on catastrophising or low scores on fear avoidance) had greater treatment effect than those with a more negative attitude. In our data set, psychological distress as measured by the MCS is positively correlated with other measures such as fear avoidance (Spearman correlation, $r = 0.064$), depression ($r = 0.137$), and anxiety ($r = 0.151$) (data not presented). This means that it is extremely unlikely that increased values in these scores would have an opposite effect to those we observed for the SF-12/36 MCS.

Thus, taking all of these findings together, a policy of treatment with conventional therapist-delivered interventions focusing on those with higher levels of psychological distress is not sustainable. What these data cannot tell us is whether or not there is a differential effect from a much more intensive treatment programme based on levels of psychological distress at baseline. In the absence of any such evidence, or any reasonable prospect that direct RCT data will become available, one might be able to infer from our findings for less intense interventions that such more intense interventions might be best targeted at those with more severe disability (however defined). This would concur with that which is current practice (where such services are available) and 2009 NICE guidance.

Methodological development

A substantial part of the programme grant was around the development of new approaches to identifying subgroups. From our review of the literature on subgroups, we concluded that the existing methods have a number of problems, including being severely underpowered, able to provide only exploratory or insufficient findings and having rather poor quality of reporting (see *Chapter 2*). Therefore, there is a need to develop new approaches to subgroup identification in back pain research.

We have developed three approaches to subgroup identification:

1. recursive partitioning (IPD-IT and IPD-SIDES method) (see *Chapter 7*)
2. adaptive risk group refinement (see *Chapters 8 and 9*)
3. individual participant data indirect NWMA (see *Chapter 10*).

These new methods challenge the current paradigm for subgroup identification in which single moderator variables are sought. Although such an approach provides a useful first step to exploring subgroups, the outputs have not produced clinically useful data to inform treatment choices for LBP. The more comprehensive methods developed as part of this programme of work use a multiparametric approach to subgroup identification, which gives far greater flexibility and clinical application.

The recursive partitioning and adaptive methods we developed for this work did not allow us to identify clinically relevant subgroups within this data set. We think that this reflects both the limitations of the data set and the likelihood that there are no distinct subgroups that might be identified in this manner. Nevertheless, the techniques performed well on the available data and the different techniques have typically generated consistent outputs. These are important methodological innovations, which we anticipate to have potential across a wide range of clinical areas. Importantly, they both use an approach that examines both the effect of variables and provides cut-off points grounded in the data. In particular, the adaptive methods allow the end-user to judge for themselves the size of any differential subgroup effect (clinical effectiveness or cost-effectiveness) that would be worthwhile and identify the parameters that would define such a group. For our adaptive approaches we have here presented just point estimates, without also ascribing statistical inference to them. This is for the sake of clarity of presentation. We have explored how to add statistical inference to these analyses. This is possible but uses an extremely large amount of computer time and generates little additional information. They are an additional approach that could be used in future analyses.

The development of NWMA to provide individualised advice on which treatment has the highest probability of being optimal for a particular patient profile is extremely exciting. Although no more than exploratory here, as it was not prespecified in our analysis plan, there is potential for this approach to inform clinical decision-making in this and other fields. Analysing the trials as a single network of evidence, and also adopting a Bayesian approach to probability, has provided us with what appears to be useful data to inform clinical decision-making in a field that has previously been devoid of useful information. When evidence is suggestive but not conclusive, Bayesian methods allow this to be quantified in a way that can be incorporated into decision-making by individual clinicians and patients.

We have developed a large and complex data set. This has presented substantial challenges (not fully appreciated at the start of the project) in terms of data management and coding. In contrast with some other areas for which IPD meta-analysis is more common, for example cardiovascular disorders, there is no consistency in how baseline variables or outcomes are measured, and there is the need for a core outcome set in this area. This has meant that we have had to do further methodological development in order to develop a new EAV approach to managing such data sets, which is far more flexible and simple for non-specialist IT staff to adapt as needed. We think that this approach is more robust and flexible than the approach of utilising an Access database used by others doing IPD meta-analysis of back pain trials.¹⁸⁴ This is an important methodological development, which we consider to have utility beyond the scope of this project.

Although not exactly a methodological development, we have examined carefully how one might map between different back pain outcome measures. The finding here that they are neither sufficiently correlated nor sufficiently similar in their responsiveness for data from trials using different outcomes to be pooled is important. This may not be entirely surprising if one examines the time frames over which different measures are considering outcome and the exact content of the measures. We are aware that the National Institutes of Health task force on back pain research identified producing crosswalk values for these 'legacy measures' as priority.¹⁵⁷ Our findings demonstrate that this exercise is not worth pursuing further. These findings also mean that existing meta-analyses of back interventions, for which results from different trials that have used different outcome measures have been pooled, may not be robust. There are multiple examples in the literature of meta-analyses that have either used SMDs or scaled measure to a 0–100 scale. We suggest that all of these reviews need to be interpreted with caution until such time as this issue has been addressed in their analyses. We have also succeeded in developing an approach to judging if different PROMs measuring the same domain can be pooled for meta-analysis that has applicability outside the field of back pain.

It may well be that the lasting legacy and impact of the programme of work reside in the methodological developments needed to do the analyses rather than the outputs of the analyses.

Strengths

This pooled data set of RCTs of therapist-delivered interventions for LBP is a valuable resource for academics and researchers in the field for the future. Such a large data set provides the statistical power needed for subgroup analyses, something that was lacking in many previous studies. This means that negative findings can be taken as absence of effect rather than absence of evidence of effect. In our original proposal we estimated that we needed data on around 3000 participants to do our analyses. Having a pooled data set of 9328 means that we have substantially more statistical power than anticipated. This means that although for many analyses we were able to use only relatively small subsets of the data for which the same outcomes had been used, we were still able to perform robust analyses. Although not being able to pool data from all trials reduces numbers in each analysis, we are confident that in each analysis the same thing is being measured in each trial. This contributes substantially to the strength of our conclusions.

The whole of this programme of work hinges on the strength of the programming and coding of trials, which have enabled the data to be pooled. The data we obtained came from varied and complex data sets using different coding structures. A large amount of work went into standardising the coding. The final database we have developed is probably over-engineered for the analyses that we have conducted. In particular we have included, wherever possible, individual item data rather than scores for any outcome measures. In the end we were not able to use this fine resolution data for our subgroup analyses. Nevertheless, we have created an excellent resource for future researchers to use to explore other research questions. Nearly all of the contributing trialists have indicated that they may be prepared to make the data available for future analyses; we would therefore be keen to encourage back pain researchers to formally bid to access the data. Furthermore, we would like to continue to add data to the repository to increase its future utility and, therefore, we would encourage academics in the field to approach us with data sets that they would like us to include. It is likely that we would need to charge researchers to upload the data to cover the research and programming time. We would therefore encourage researchers to include costs of uploading their final data into this data set in any future grant applications.

The results obtained come from the application of two different frequentist approaches to subgroup identification: recursive partitioning and adaptive risk group refinement. Both approaches yield similar conclusions: that although it is possible to use multiple parameters to describe subgroups, these are unlikely to be clinically important. Additionally, the NWMA has identified the same parameters as being important and with the same directionality (although noting here that for the QALY analysis it is older people who gain a greater benefit). Therefore, as a strength we can be confident that our analyses are robust, yielding the same overall outcome.

Limitations

Our exploratory work on mapping between outcome measures which measure the same domain, to a common scale, led us to conclude that this is not possible and therefore we would be unable to pool outcomes measuring the same domain (see *Chapter 5*). For this reason, despite having a large data set, for some comparisons we had rather fewer data. As the programme was originally produced, we had anticipated using individual item data to help define subgroups. However, as we developed our methodology it became clear that we would not be able to use such a large number of items and obtain meaningful outcomes in a reasonable time frame; such analyses would be beyond capacity of our computing systems. Furthermore, as the work developed, we selected moderators for our analyses grounded in existing data. There is a hazard that we would falsely identify moderators, as data from three of the four studies that informed our choice of potential moderators were included in our analyses here. As we have not identified any large subgroup effects this need not be of great concern. We were able to explore only some of the domains that were identified in our literature review because in many cases only one study had measured that particular variable, and there would be no added value to be gained from running an analysis in the pooled data set.

The interventions used in the trials were trial specific. To enable grouping of interventions, trials were broadly grouped into active physical, passive physical, psychological, sham and control.

Initially, we grouped the sham and control together as a single control group. This was later separated out, based on some exploratory analyses indicating a treatment effect for sham. The sham group is largely made up of participants who received sham acupuncture. Some may argue that our approach to grouping these interventions is not conventional, as every intervention is different, and therefore how can they be grouped and treated as being the same. From a practical perspective of managing the data and using it to do any meaningful analyses it was essential that the data were grouped in some manner. The approach we have taken was carefully considered by the research team, including our lay members, before the final groupings were decided.

Therapist and group effects can also affect the analysis of trials of the types of interventions that we are evaluating here. We did not have enough detail to include these in our pooled analyses. From our experience of the BeST³³ and BEAM³¹ trials, for which we know therapist effects were measured, we have found these to be negligible and therefore are unlikely to be a source of bias.

All of these findings need to be interpreted with some caution. We have undertaken many analyses, meaning that some positive findings might have been observed by chance. In addition, several of the data sets that we included in our analyses were also data sets that were used in other studies to identify our possible moderators, and were the same data sets that we used for our ANCOVAs. This, again, increases the possibility that we might have found a spurious positive result. From our pairwise comparisons, and with these caveats, we failed to identify any clear and consistent differential subgroup effects beyond 'those who have more problems at baseline have more to gain and those, with increased psychological distress, as measured by the SF-12/SF-36 MCS, may gain less benefit'; consequently, this becomes a very strong finding.

Our exploratory analytical approach to identifying subgroups that may do best with different treatment approaches using a Bayesian NWMA has provided some promising results. In this analysis we have not identified subgroups in a conventional manner. Rather, we have used all of the available data to assess the probability that, for a group of patients with a similar profile, a particular treatment choice is the most likely to be effective. For some of our paradigmatic cases there are clear messages as to which treatment types may be more effective. In some cases, sham treatment (typically sham acupuncture) appears to be the preferred choice. As the NHS is unlikely to offer sham treatment as a patient choice, some thought is needed on how to interpret these findings. Perhaps one would choose to offer verum acupuncture, which some argue is no more than a therapeutic placebo.¹⁹³ Even if it is truly a sham treatment, it is one that many have belief in, which could be offered, rather than something that no-one has belief in, such as de-tuned ultrasound. Although of some academic interest to explore how sham treatment could appear to be the optimal treatment, even ahead of the active treatment for which it is the control, this is not of clinical relevance. If this approach to treatment selection was implemented clinically then the option of sham treatment could be removed and the second-choice approach advised.

Meaning of the results and clinical implications

The important clinical implication of the results is that there is very little clinical effectiveness or cost-effectiveness justification for using the baseline characteristics we studied to define groups who might benefit from different back pain treatment. Based on these data, the hypothesis that low-intensity therapist-delivered interventions should be targeted at those with higher levels of psychological distress (as measured by SF-12/36 MCS) is not supported. It is possible that the results of the Bayesian analysis might allow us to give more information which might help improve treatment selection; this will need empirical testing before it can be recommended. Most importantly, we have developed statistical methods for subgroup analysis that move beyond simply looking for interaction effects with single moderator variables. These approaches may have quite wide applicability.

Recommendations for future research

We have made a number of suggestions for further research; however, these are not necessarily in order of priority.

1. Making the data set available to other researchers. We are in the process of updating data sharing agreements to allow us to make our data available to other researchers.
2. Adding additional trial data sets to the repository. We are aware of two other groups that are working on intervention-specific IPD meta-analyses. We are working with them to develop a shared codebook for these trials. A next step would be to develop a user-friendly interface that would allow the original researchers to upload their data into the repository. We are aware of moves to make trial data more freely available for secondary research. Further development of this data set will provide such a resource for the back pain research community.
3. Application of these methods for the identification of subgroups in other clinical areas. We will make our methods freely available to other researchers.
4. Re-analysis of existing meta-analyses of back pain treatments that have pooled different outcome measures. As current Cochrane reviews are updated, it would be possible to group any meta-analyses according to the outcome measure being reported. In the absence of heterogeneity in outcome according to outcome measure used, it may be possible to pool data to give an overall estimate with some caveats as to whether pooling in this manner is robust.
5. Further development of methods and application to the data we already have.
6. Explore the need for a core outcomes set for LBP in the light of existing developments in the area.

Conclusions

The lasting legacy of this work is likely to be the methodological developments needed to do our analyses. We have developed improved systems for storing large, complex data sets; developed methods for assessing comparability of outcome measures, which have demonstrated that different back pain outcome measures cannot be safely pooled for meta-analyses; and we have developed three different approaches to the identification of differential subgroup effects, which provide considerable added values compared with conventional analyses that simply test for interactions.

Using frequentist approaches (recursive partitioning or adaptive approaches) has not allowed the identification of subgroups that might have worthwhile additional benefits from different treatment approaches beyond the potential benefits of being greater in those with more disability at baseline. Importantly, increased psychological distress, as measured using the SF-12/36 MCS, may identify those less likely to benefit from treatment – the opposite of conventional wisdom, which is that this group should be targeted for intervention.

An approach based on Bayesian NWMA offers a potential approach to deciding on optimal therapies. We would suggest that these methods are applied in other clinical areas in which subgroup identification and targeting of treatment may be advantageous.

Our findings do challenge conventional wisdom on who should be prioritised for back pain treatments, that is, those with greater psychological distress. We would not support such an approach until there is evidence to challenge our findings.

Finally, we have developed an important resource for back pain researchers wishing to do further analyses on data from multiple trials.

Acknowledgements

Contributions of authors

Dr Shilpa Patel (Senior Research Fellow, Health Psychology, co-applicant) made a substantial contribution to the design, organisation and conduct of the programme grant, was responsible for negotiating trial data sets and assisted with interpreting data. She contributed substantially to the writing of this report.

Dr Siew Wan Hee (Research Fellow, Statistics) made a substantial contribution to the database development, coding of data, statistical analysis and interpretation of the results. She drafted several sections of this report and commented on other chapters.

Dr Dipesh Mistry (Research Fellow, Statistics) conducted the review on subgroups and contributed to the review on moderators. He developed the recursive partitioning methods during his PhD study (funded through this programme grant) and made a substantial contribution to the application and interpretation of this method to the data sets obtained in the repository. He drafted the literature review section on subgroups and the results chapter for recursive partitioning, as well as commenting on other chapters.

Mr Jake Jordan (Research Associate, Health Economics) made a contribution to the coding of economic data, statistical analysis and interpretation of the results. Together with Dr Joanne Lord, he drafted the relevant health-economic sections of the report and commented on other chapters.

Ms Sally Brown (lay member, co-applicant) made an important contribution to the early design of the programme grant and subsequently assisted with the interpretation of the data, and commented on the overall report.

Dr Melina Dritsaki (Research Fellow, Health Economics) made a substantial contribution to the early thinking and development of coding structures for the economic data. She has commented on, and contributed to, the relevant chapters in this report.

Dr David Ellard (Senior Research Fellow, Health Services Research, co-applicant) substantially contributed to the conduct of the moderators systematic review, including data extraction, interpretation and write-up of the results.

Professor Tim Friede (Professor of Biostatistics, Statistics, co-applicant) made a substantial contribution to the original grant application and development of statistical methods for analysis. He has helped draft various sections of the methods and results chapters, and commented on the overall report.

Professor Sarah E Lamb (Professor of Rehabilitation, Rehabilitation) contributed to the original grant application, assisted with the interpretation of the results and commented on the overall report for intellectual content.

Dr Joanne Lord (Reader, Health Economics, co-applicant) made a substantial contribution to the original grant application and development of statistical methods for analysis of economic data. Together with Mr Jake Jordan, she has drafted the relevant health-economic sections of the report and commented on other chapters.

Dr Jason Madan (Assistant Professor, Health Economics) applied IPD meta-analyses to the pooled data set, interpreted these results and wrote the corresponding chapter for this report.

Dr Tom Morris (Research Fellow, Statistics) contributed substantially to the work on crosswalking in this report. He assisted with drafting and commenting on this chapter. He also made a significant contribution to the coding of clinical data for the pooled repository.

Professor Nigel Stallard (Professor of Medical Statistics, Statistics, co-applicant) made a substantial contribution to the development of statistical methods, analysis and interpretation of data. He has also commented on the overall report.

Mr Colin Tysall (lay member, co-applicant) made an important contribution to the early design of the programme grant, and subsequently assisted with the interpretation of the data and commented on the overall report.

Mr Adrian Willis (Senior Programmer, Programming) developed the programming to enable pooling of large data sets. He contributed to the writing of the database development chapter.

Professor Martin Underwood (Director of Warwick Clinical Trials Unit, Professor of Primary Care Research, chief investigator) was responsible for developing the proposal for funding and had overall responsibility for the conduct of the programme of work. He contributed to all aspects of the programme grant, including interpretation of results and drafting and finalisation of the final report for crucial intellectual content.

Repository Group

Our thanks goes to all the chief investigators and data custodians who agreed to share their trial data with us for this project, including:

- Dr Christer Carlsson (Carlsson)
- Dr Francesca Cecchi (Cecchi)
- Dr Ninna Dufour (Dufour)
- Dr Heinz Endres (Haake)
- Dr Mark Hancock (Hancock)
- Professor Elaine Hay (Keele)
- Dr von Korff (von Korff BIA, von Korff SC2)
- Professor Sarah Lamb (BeST)
- Dr Luciana Macedo (Macedo)
- Dr Hugh MacPherson (YACBAC)
- Professor Chris Maher (Pengle)
- Professor Suzanne McDonough (Kennedy)
- Professor Rob Smeets (Smeets)
- Professor David Torgerson (UK BEAM, HullExPro, York BP)
- Professor Claudia Witt (Witt, Brinkhaus).

Participants

All of the participants who took part in the trials from which we have obtained data.

Other acknowledgements

- Dr Tara Gurung for her input into the moderators systematic review.
- Mr Mark Woolvine for his earlier contributions to the grant.
- Ms Sarah Gunter for her support with the initial grant application.
- BackCare for their earlier contributions to the grant.
- Acupuncture Trialists' Collaboration for assisting with access to trial data sets.
- Lippincott Williams & Wilkins publishers for allowing reproduction of material for this report.

Administrative support

Mr James Crawford for formatting the final report.

Programme Steering Group

Members included Professor Daniëlle van der Windt (chairperson), Professor Claudia Witt, Professor Andrea Manca, Dr Richard Riley, Dr Mindy Cairns, Mr Mike Andrews and Mr Chris Phillips.

Data sharing statement

For a sample data sharing agreement, see *Appendix 4*.

References

1. Andersson GB. Epidemiology of low back pain. *Acta Orthop Scand Suppl* 1998;**281**:28–31.
2. Deyo RA, Cherkin D, Conrad D, Volinn E. Cost, controversy, crisis: low back pain and the health of the public. *Ann Rev Publ Health* 1991;**12**:141–56. <http://dx.doi.org/10.1146/annurev.pu.12.050191.001041>
3. Dionne CE, Dunn KM, Croft PR. Does back pain prevalence really decrease with increasing age? A systematic review. *Age Ageing* 2006;**35**:229–34. <http://dx.doi.org/10.1093/ageing/afj055>
4. Rapoport J, Jacobs P, Bell NR, Klarenbach S. Refining the measurement of the economic burden of chronic diseases in Canada. *CDIC* 2004;**25**:13–21.
5. Palmer KT, Walsh K, Bendall H, Cooper C, Coggon D. Back pain in Britain: comparison of two prevalence surveys at an interval of 10 years. *BMJ* 2000;**320**:1577–8. <http://dx.doi.org/10.1136/bmj.320.7249.1577>
6. Raspe H. Back Pain. In Silman A, Hochberg M, editors. *Epidemiology of the Rheumatic Diseases*. Oxford: Oxford University Press; 2001. pp. 309–38.
7. Raspe H, Hueppe A, Neuhauser H. Back pain, a communicable disease? *Int J Epidemiol* 2008;**37**:69–74. <http://dx.doi.org/10.1093/ije/dym220>
8. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, *et al.* Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012;**380**:2163–96. [http://dx.doi.org/10.1016/S0140-6736\(12\)61729-2](http://dx.doi.org/10.1016/S0140-6736(12)61729-2)
9. Pengel LH, Herbert RD, Maher CG, Refshauge KM. Acute low back pain: systematic review of its prognosis. *BMJ* 2003;**327**:323. <http://dx.doi.org/10.1136/bmj.327.7410.323>
10. Walker BF. The prevalence of low back pain: a systematic review of the literature from 1966 to 1998. *J Spinal Disord* 2000;**13**:205–17. <http://dx.doi.org/10.1097/00002517-200006000-00003>
11. Jeffries LJ, Milanese SF, Grimmer-Somers KA. Epidemiology of adolescent spinal pain: a systematic overview of the research literature. *Spine* 2007;**32**:2630–7. <http://dx.doi.org/10.1097/BRS.0b013e318158d70b>
12. Hoy D, Bain C, Williams G, March L, Brooks P, Blyth F, *et al.* A systematic review of the global prevalence of low back pain. *Arthritis Rheum* 2012;**64**:2028–37. <http://dx.doi.org/10.1002/art.34347>
13. International Association for the Study of Pain (IASP). *IASP Taxonomy: Pain Terms*. Washington, DC; 2014. URL: www.iasp-pain.org/Taxonomy?navItemNumber=576#Pain (accessed 13 October 2014).
14. The British Pain Society (BPS). *FAQs*. London: BPS; 2008. URL: www.britishpainsociety.org/media_faq.htm (accessed 13 October 2014).
15. Savigny P, Watson P, Underwood M. Early management of persistent non-specific low back pain: summary of NICE guidance. *BMJ* 2009;**338**:b1805. <http://dx.doi.org/10.1136/bmj.b1805>
16. Downie A, Williams CM, Henschke N, Hancock MJ, Ostelo RW, de Vet HC, *et al.* Red flags to screen for malignancy and fracture in patients with low back pain: systematic review. *BMJ* 2013;**347**:f7095. <http://dx.doi.org/10.1136/bmj.f7095>
17. Murray CJ, Richards MA, Newton JN, Fenton KA, Anderson HR, Atkinson C, *et al.* UK health performance: findings of the Global Burden of Disease Study 2010. *Lancet* 2013;**381**:997–1020. [http://dx.doi.org/10.1016/S0140-6736\(13\)60355-4](http://dx.doi.org/10.1016/S0140-6736(13)60355-4)

18. Waddell G. *The Back Pain Revolution*. 2nd edn. London: Churchill; 2004.
19. Kent PM, Keating JL. The epidemiology of low back pain in primary care. *Chiropr Osteopat* 2005;**13**:13. <http://dx.doi.org/10.1186/1746-1340-13-13>
20. Steenstra IA, Verbeek JH, Heymans MW, Bongers PM. Prognostic factors for duration of sick leave in patients sick listed with acute low back pain: a systematic review of the literature. *Occup Environ Med* 2005;**62**:851–60. <http://dx.doi.org/10.1136/oem.2004.015842>
21. Thelin A, Holmberg S, Thelin N. Functioning in neck and low back pain from a 12-year perspective: a prospective population-based study. *J Rehabil Med* 2008;**40**:555–61. <http://dx.doi.org/10.2340/16501977-0205>
22. Maniadakis N, Gray A. The economic burden of back pain in the UK. *Pain* 2000;**84**:95–103. [http://dx.doi.org/10.1016/S0304-3959\(99\)00187-6](http://dx.doi.org/10.1016/S0304-3959(99)00187-6)
23. Dagenais S, Caro J, Haldeman S. A systematic review of low back pain cost of illness studies in the United States and internationally. *Spine J* 2008;**8**:8–20. <http://dx.doi.org/10.1016/j.spinee.2007.10.005>
24. Dunn KM, Croft PR. Epidemiology and natural history of low back pain. *Eura Medicophys* 2004;**40**:9–13.
25. Office for National Statistics (ONS). *Full Report: Sickness Absence in the Labour Market, February 2014*. London: ONS; 2014.
26. Ehrlich G, Khaltayev N. *Low Back Pain Initiative*. Geneva: World Health Organization; 1999.
27. National Institute for Health and Care Excellence. *Low Back Pain: Early Management of Persistent Non-Specific Low Back Pain*. Manchester: NICE; 2009.
28. 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. <http://dx.doi.org/10.1097/00007632-198303000-00004>
29. Froud R, Eldridge S, Lall R, Underwood M. Estimating the number needed to treat from continuous outcomes in randomised controlled trials: methodological challenges and worked example using data from the UK Back Pain Exercise and Manipulation (BEAM) trial. *BMC Med Res Methodol* 2009;**9**:35. <http://dx.doi.org/10.1186/1471-2288-9-35>
30. Ostelo RW, Deyo RA, Stratford P, Waddell G, Croft P, Von Korf M, *et al*. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine* 2008;**33**:90–4. <http://dx.doi.org/10.1097/BRS.0b013e31815e3a10>
31. 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. <http://dx.doi.org/10.1136/bmj.38282.669225.AE>
32. Little P, Lewith G, Webley F, Evans M, Beattie A, Middleton K, *et al*. Randomised controlled trial of Alexander technique lessons, exercise, and massage (ATEAM) for chronic and recurrent back pain. *BMJ* 2008;**337**:a884. <http://dx.doi.org/10.1136/bmj.a884>
33. Lamb SE, Hansen Z, Lall R, Castelnuovo E, Withers EJ, Nichols V, *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. [http://dx.doi.org/10.1016/S0140-6736\(09\)62164-4](http://dx.doi.org/10.1016/S0140-6736(09)62164-4)
34. Lamb SE, Lall R, Hansen Z, Castelnuovo E, Withers EJ, Nichols V, *et al*. A multicentred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain. The Back Skills Training (BeST) trial. *Health Technol Assess* 2010;**14**(41). <http://dx.doi.org/10.3310/hta14410>

35. Tilbrook HE, Cox H, Hewitt CE, Kang'ombe AR, Chuang LH, Jayakody S, *et al.* Yoga for chronic low back pain: a randomized trial. *Ann Int Med* 2011;**155**:569–78. <http://dx.doi.org/10.7326/0003-4819-155-9-201111010-00003>
36. Moore A, Derry S, Eccleston C, Kalso E. Expect analgesic failure; pursue analgesic success. *BMJ* 2013;**346**:f2690. <http://dx.doi.org/10.1136/bmj.f2690>
37. Lin CW, Haas M, Maher CG, Machado LA, van Tulder MW. Cost-effectiveness of guideline-endorsed treatments for low back pain: a systematic review. *Eur Spine J* 2011;**20**:1024–38. <http://dx.doi.org/10.1007/s00586-010-1676-3>
38. Hill JC, Whitehurst DG, Lewis M, Bryan S, Dunn KM, Foster NE, *et al.* Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet* 2011;**378**:1560–71. [http://dx.doi.org/10.1016/S0140-6736\(11\)60937-9](http://dx.doi.org/10.1016/S0140-6736(11)60937-9)
39. Gurung T, Ellard DR, Mistry D, Patel S, Underwood M. Identifying potential moderators for response to treatment in low back pain: a systematic review. *Physiotherapy* 2015;**101**:243–51. <http://dx.doi.org/10.1016/j.physio.2015.01.006>
40. Turner JA, Holtzman S, Mancl L. Mediators, moderators, and predictors of therapeutic change in cognitive-behavioural therapy for chronic pain. *Pain* 2007;**127**:276–86. <http://dx.doi.org/10.1016/j.pain.2006.09.005>
41. Kamper SJ, Maher CG, Hancock MJ, Koes BW, Croft PR, Hay E. Treatment-based subgroups of low back pain: a guide to appraisal of research studies and a summary of current evidence. *Best Pract Res Clin Rheumatol* 2010;**24**:181–91. <http://dx.doi.org/10.1016/j.berh.2009.11.003>
42. Lachenbruch PA. A note on sample size computation for testing interactions. *Stat Med* 1988;**7**:467–9. <http://dx.doi.org/10.1002/sim.4780070403>
43. The Cochrane Collaboration (Higgins JP, Green S, editors). *Cochrane Handbook of Systematic Reviews of Intervention*. Version 5.1.0. Oxford: The Cochrane Collaboration; 2011.
44. Pincus T, Miles C, Froud R, Underwood M, Carnes D, Taylor SJ. Methodological criteria for the assessment of moderators in systematic reviews of randomised controlled trials: a consensus study. *BMC Med Res Methodol* 2011;**11**:14. <http://dx.doi.org/10.1186/1471-2288-11-14>
45. Underwood MR, Morton V, Farrin A. Do baseline characteristics predict response to treatment for low back pain? Secondary analysis of the UK BEAM dataset. *Rheumatology (Oxford)* 2007;**46**:1297–302. <http://dx.doi.org/10.1093/rheumatology/kem113>
46. Underwood M, Mistry D, Lall R, Lamb S. Predicting response to a cognitive-behavioural approach to treating low back pain: secondary analysis of the BeST data set. *Arthritis Care Res* 2011;**63**:1271–9. <http://dx.doi.org/10.1002/acr.20518>
47. Witt CM, Schutzler L, Ludtke R, Wegscheider K, Willich SN. Patient characteristics and variation in treatment outcomes: which patients benefit most from acupuncture for chronic pain? *Clin J Pain* 2011;**27**:550–5. <http://dx.doi.org/10.1097/AJP.0b013e31820dfbf5>
48. Sherman KJ, Cherkin DC, Ichikawa L, Avins AL, Barlow WE, Khalsa PS, *et al.* Characteristics of patients with chronic back pain who benefit from acupuncture. *BMC Musculoskelet Disord* 2009;**10**:114. <http://dx.doi.org/10.1186/1471-2474-10-114>
49. Cherkin DC, Sherman KJ, Avins AL, Erro JH, Ichikawa L, Barlow WE, *et al.* A randomized trial comparing acupuncture, simulated acupuncture, and usual care for chronic low back pain. *Arch Intern Med* 2009;**169**:858–66. <http://dx.doi.org/10.1001/archinternmed.2009.65>
50. Witt CM, Jena S, Selim D, Brinkhaus B, Reinhold T, Wruck K, *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. <http://dx.doi.org/10.1093/aje/kwj224>

51. Mistry D, Patel S, Hee SW, Stallard N, Underwood M. Evaluating the quality of subgroup analyses in randomized controlled trials of therapist-delivered interventions for nonspecific low back pain: a systematic review. *Spine* 2014;**39**:618–29. <http://dx.doi.org/10.1097/BRS.0000000000000231>
52. Kraemer HC, Stice E, Kazdin A, Offord D, Kupfer D. How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *Am J Psychiatry* 2001;**158**:848–56. <http://dx.doi.org/10.1176/appi.ajp.158.6.848>
53. Kent P, Keating JL, Leboeuf-Yde C. Research methods for subgrouping low back pain. *BMC Med Res Methodol* 2010;**10**:62. <http://dx.doi.org/10.1186/1471-2288-10-62>
54. Borkan JM, Koes B, Reis S, Cherkin DC. A report from the Second International Forum for Primary Care Research on Low Back Pain. Re-examining priorities. *Spine* 1998;**23**:1992–6. <http://dx.doi.org/10.1097/00007632-199809150-00016>
55. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;**266**:93–8. <http://dx.doi.org/10.1001/jama.1991.03470010097038>
56. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005;**365**:176–86. [http://dx.doi.org/10.1016/S0140-6736\(05\)17709-5](http://dx.doi.org/10.1016/S0140-6736(05)17709-5)
57. Lagakos SW. The challenge of subgroup analyses: reporting without distorting. *N Engl J Med* 2006;**354**:1667–9. <http://dx.doi.org/10.1056/NEJMp068070>
58. Sheets C, Machado LA, Hancock M, Maher C. Can we predict response to the McKenzie method in patients with acute low back pain? A secondary analysis of a randomized controlled trial. *Eur Spine J* 2012;**21**:1250–6. <http://dx.doi.org/10.1007/s00586-011-2082-1>
59. Smeets RJ, Maher CG, Nicholas MK, Refshauge KM, Herbert RD. Do psychological characteristics predict response to exercise and advice for subacute low back pain? *Arthritis Rheum* 2009;**61**:1202–9. <http://dx.doi.org/10.1002/art.24731>
60. Becker A, Leonhardt C, Kochen MM, Keller S, Wegscheider K, Baum E, et al. Effects of two guideline implementation strategies on patient outcomes in primary care: a cluster randomized controlled trial. *Spine* 2008;**33**:473–80. <http://dx.doi.org/10.1097/BRS.0b013e3181657e0d>
61. Cecchi F, Negrini S, Pasquini G, Paperini A, Conti AA, Chiti M, et al. Predictors of functional outcome in patients with chronic low back pain undergoing back school, individual physiotherapy or spinal manipulation. *Eur J Phys Rehabil Med* 2012;**48**:371–8.
62. Cherkin DC, Deyo RA, Battie M, Street J, Barlow W. A comparison of physical therapy, chiropractic manipulation, and provision of an educational booklet for the treatment of patients with low back pain. *N Engl J Med* 1998;**339**:1021–9. <http://dx.doi.org/10.1056/NEJM199810083391502>
63. Cherkin DC, Eisenberg D, Sherman KJ, Barlow W, Kaptchuk TJ, Street J, et al. Randomized trial comparing traditional Chinese medical acupuncture, therapeutic massage, and self-care education for chronic low back pain. *Arch Int Med* 2001;**161**:1081–8. <http://dx.doi.org/10.1001/archinte.161.8.1081>
64. Hansen FR, Bendix T, Skov P, Jensen CV, Kristensen JH, Krohn L, et al. Intensive, dynamic back-muscle exercises, conventional physiotherapy, or placebo-control treatment of low-back pain. A randomized, observer-blind trial. *Spine* 1993;**18**:98–108. <http://dx.doi.org/10.1097/00007632-199301000-00015>
65. Hay EM, Mullis R, Lewis M, Vohora K, Main CJ, Watson P, et al. Comparison of physical treatments versus a brief pain-management programme for back pain in primary care: a randomised clinical trial in physiotherapy practice. *Lancet* 2005;**365**:2024–30. [http://dx.doi.org/10.1016/S0140-6736\(05\)66696-2](http://dx.doi.org/10.1016/S0140-6736(05)66696-2)

66. Juni P, Battaglia M, Nuesch E, Hammerle G, Eser P, van Beers R, *et al.* A randomised controlled trial of spinal manipulative therapy in acute low back pain. *Ann Rheum Dis* 2009;**68**:1420–7. <http://dx.doi.org/10.1136/ard.2008.093757>
67. Karjalainen K, Malmivaara A, Mutanen P, Roine R, Hurri H, Pohjolainen T. Mini-intervention for subacute low back pain: two-year follow-up and modifiers of effectiveness. *Spine* 2004;**29**:1069–76. <http://dx.doi.org/10.1097/00007632-200405150-00004>
68. Kole-Snijders AM, Vlaeyen JW, Goossens ME, Rutten-van Molken MP, Heuts PH, van Breukelen G, *et al.* Chronic low-back pain: what does cognitive coping skills training add to operant behavioral treatment? Results of a randomized clinical trial. *J Consult Clin Psychol* 1999;**67**:931–44. <http://dx.doi.org/10.1037/0022-006X.67.6.931>
69. Roche G, Ponthieux A, Parot-Shinkel E, Jousset N, Bontoux L, Dubus V, *et al.* Comparison of a functional restoration program with active individual physical therapy for patients with chronic low back pain: a randomized controlled trial. *Arch Phys Med Rehabil* 2007;**88**:1229–35. <http://dx.doi.org/10.1016/j.apmr.2007.07.014>
70. Smeets RJ, Vlaeyen JW, Hidding A, Kester AD, van der Heijden GJ, van Geel AC, *et al.* Active rehabilitation for chronic low back pain: cognitive-behavioral, physical, or both? First direct post-treatment results from a randomized controlled trial. *BMC Musculoskelet Disord* 2006;**7**:5. <http://dx.doi.org/10.1186/1471-2474-7-5>
71. Smeets RJ, Vlaeyen JW, Hidding A, Kester AD, van der Heijden GJ, Knottnerus JA. Chronic low back pain: physical training, graded activity with problem solving training, or both? The one-year post-treatment results of a randomized controlled trial. *Pain* 2008;**134**:263–76. <http://dx.doi.org/10.1016/j.pain.2007.04.021>
72. Van der Hulst M, Vollenbroek-Hutten MM, Groothuis-Oudshoorn KG, Hermens HJ. Multidisciplinary rehabilitation treatment of patients with chronic low back pain: a prognostic model for its outcome. *Clin J Pain* 2008;**24**:421–30. <http://dx.doi.org/10.1097/AJP.0b013e31816719f5>
73. Bendix AF, Bendix T, Hastrup C. Can it be predicted which patients with chronic low back pain should be offered tertiary rehabilitation in a functional restoration program? A search for demographic, socioeconomic, and physical predictors. *Spine* 1998;**23**:1775–83, discussion 83–4.
74. Beurskens AJ, de Vet HC, Koke AJ, Lindeman E, Regtop W, van der Heijden GJ, *et al.* Efficacy of traction for non-specific low back pain: a randomised clinical trial. *Lancet* 1995;**346**:1596–600. [http://dx.doi.org/10.1016/S0140-6736\(95\)91930-9](http://dx.doi.org/10.1016/S0140-6736(95)91930-9)
75. Bishop MD, Bialosky JE, Cleland JA. Patient expectations of benefit from common interventions for low back pain and effects on outcome: secondary analysis of a clinical trial of manual therapy interventions. *J Man Manip Ther* 2011;**19**:20–5. <http://dx.doi.org/10.1179/106698110X12804993426929>
76. Carr JL, Klaber Moffett JA, Howarth E, Richmond SJ, Torgerson DJ, Jackson DA, *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. <http://dx.doi.org/10.1080/09638280500030639>
77. Ferreira ML, Ferreira PH, Latimer J, Herbert RD, Maher C, Refshauge K. Relationship between spinal stiffness and outcome in patients with chronic low back pain. *Manual Ther* 2009;**14**:61–7. <http://dx.doi.org/10.1016/j.math.2007.09.013>
78. Glazov G, Schattner P, Lopez D, Shandley K. Laser acupuncture for chronic non-specific low back pain: a controlled clinical trial. *Acupunct Med* 2009;**27**:94–100. <http://dx.doi.org/10.1136/aim.2009.000521>

79. Gudavalli MR, Cambron JA, McGregor M, Jedlicka J, Keenum M, Ghanayem AJ, *et al.* A randomized clinical trial and subgroup analysis to compare flexion-distraction with active exercise for chronic low back pain. *Eur Spine J* 2006;**15**:1070–82. <http://dx.doi.org/10.1007/s00586-005-0021-8>
80. Hsieh LL, Kuo CH, Yen MF, Chen TH. A randomized controlled clinical trial for low back pain treated by acupressure and physical therapy. *Prev Med* 2004;**39**:168–76. <http://dx.doi.org/10.1016/j.ypmed.2004.01.036>
81. Jellema P, van der Windt DA, van der Horst HE, Blankenstein AH, Bouter LM, Stalman WA. Why is a treatment aimed at psychosocial factors not effective in patients with (sub)acute low back pain? *Pain* 2005;**118**:350–9. <http://dx.doi.org/10.1016/j.pain.2005.09.002>
82. Johnson RE, Jones GT, Wiles NJ, Chaddock C, Potter RG, Roberts C, *et al.* Active exercise, education, and cognitive behavioral therapy for persistent disabling low back pain: a randomized controlled trial. *Spine* 2007;**32**:1578–85. <http://dx.doi.org/10.1097/BRS.0b013e318074f890>
83. Kalauokalani D, Cherkin DC, Sherman KJ, Koepsell TD, Deyo RA. Lessons from a trial of acupuncture and massage for low back pain: patient expectations and treatment effects. *Spine* 2001;**26**:1418–24. <http://dx.doi.org/10.1097/00007632-200107010-00005>
84. Mellin G, Hurri H, Harkapaa K, Jarvikoski A. A controlled study on the outcome of inpatient and outpatient treatment of low back pain. Part II. Effects on physical measurements three months after treatment. *Scand J Rehabil Med* 1989;**21**:91–5.
85. Klaber Moffett JA, Carr J, Howarth E. High fear-avoiders of physical activity benefit from an exercise program for patients with back pain. *Spine* 2004;**29**:1167–72, discussion 73. <http://dx.doi.org/10.1097/00007632-200406010-00002>
86. Myers SS, Phillips RS, Davis RB, Cherkin DC, Legedza A, Kaptchuk TJ, *et al.* Patient expectations as predictors of outcome in patients with acute low back pain. *J Gen Int Med* 2008;**23**:148–53. <http://dx.doi.org/10.1007/s11606-007-0460-5>
87. Seferlis T, Nemeth G, Carlsson AM, Gillstrom P. Conservative treatment in patients sick-listed for acute low-back pain: a prospective randomised study with 12 months' follow-up. *Eur Spine J* 1998;**7**:461–70. <http://dx.doi.org/10.1007/s005860050109>
88. Thomas KJ, MacPherson H, Thorpe L, Brazier J, Fitter M, Campbell MJ, *et al.* Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain. *BMJ* 2006;**333**:623. <http://dx.doi.org/10.1136/bmj.38878.907361.7C>
89. Van der Roer N, van Tulder M, Barendse J, Knol D, van Mechelen W, de Vet H. Intensive group training protocol versus guideline physiotherapy for patients with chronic low back pain: a randomised controlled trial. *Eur Spine J* 2008;**17**:1193–200. <http://dx.doi.org/10.1007/s00586-008-0718-6>
90. Vollenbroek-Hutten MM, Hermens HJ, Wever D, Gorter M, Rinket J, Ijzerman MJ. Differences in outcome of a multidisciplinary treatment between subgroups of chronic low back pain patients defined using two multi-axial assessment instruments: the multidimensional pain inventory and lumbar dynamometry. *Clin Rehabil* 2004;**18**:566–79. <http://dx.doi.org/10.1191/0269215504cr772oa>
91. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine: reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007;**357**:2189–94. <http://dx.doi.org/10.1056/NEJMs077003>

92. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess* 2001;**5**(33). <http://dx.doi.org/10.3310/hta5330>
93. Patel S, Friede T, Froud R, Evans DW, Underwood M. Systematic review of randomized controlled trials of clinical prediction rules for physical therapy in low back pain. *Spine* 2013;**38**:762–9. <http://dx.doi.org/10.1097/BRS.0b013e31827b158f>
94. Hancock MJ, Maher CG, Latimer J, Herbert RD, McAuley JH. Independent evaluation of a clinical prediction rule for spinal manipulative therapy: a randomised controlled trial. *Eur Spine J* 2008;**17**:936–43. <http://dx.doi.org/10.1007/s00586-008-0679-9>
95. Brennan GP, Fritz JM, Hunter SJ, Thackeray A, Delitto A, Erhard RE. Identifying subgroups of patients with acute/subacute ‘nonspecific’ low back pain: results of a randomized clinical trial. *Spine* 2006;**31**:623–31. <http://dx.doi.org/10.1097/01.brs.0000202807.72292.a8>
96. Childs J, Fritz J, Flynn T, Irrgang J, Johnson K, Majkowski G, et al. A clinical prediction rule to identify patients with low back pain most likely to benefit from spinal manipulation: a validation study. *Ann Intern Med* 2004;**141**:920–8. <http://dx.doi.org/10.7326/0003-4819-141-12-200412210-00008>
97. Su X, Tsai C-L, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. *JMLR* 2009;**10**:141–58. <http://dx.doi.org/10.2139/ssrn.1341380>
98. Dusseldorp E, Meulman J. The regression trunk approach to discover treatment covariate interaction. *Psychometrika* 2004;**69**:355–74. <http://dx.doi.org/10.1007/BF02295641>
99. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search: a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat Med* 2011;**30**:2601–21. <http://dx.doi.org/10.1002/sim.4289>
100. Underwood MR, Harding G, Klaber Moffett J. Patient perceptions of physical therapy within a trial for back pain treatments (UK BEAM). *Rheumatology (Oxford)* 2006;**45**:751–6. <http://dx.doi.org/10.1093/rheumatology/kei254>
101. Brinkhaus B, Witt CM, Jena S, Linde K, Streng A, Wagenpfeil S, et al. Acupuncture in patients with chronic low back pain: a randomized controlled trial. *Arch Int Med* 2006;**166**:450–7. <http://dx.doi.org/10.1001/archinte.166.4.450>
102. Dufour N, Thamsborg G, Oefelt A, Lundsgaard C, Stender S. Treatment of chronic low back pain: a randomized, clinical trial comparing group-based multidisciplinary biopsychosocial rehabilitation and intensive individual therapist-assisted back muscle strengthening exercises. *Spine* 2010;**35**:469–76. <http://dx.doi.org/10.1097/BRS.0b013e3181b8db2e>
103. Pengel LH, Refshauge KM, Maher CG, Nicholas MK, Herbert RD, McNair P. Physiotherapist-directed exercise, advice, or both for subacute low back pain: a randomized trial. *Ann Int Med* 2007;**146**:787–96. <http://dx.doi.org/10.7326/0003-4819-146-11-200706050-00007>
104. Von Korff M, Balderson BH, Saunders K, Miglioretti DL, Lin EH, Berry S, et al. A trial of an activating intervention for chronic back pain in primary care and physical therapy settings. *Pain* 2005;**113**:323–30. <http://dx.doi.org/10.1016/j.pain.2004.11.007>
105. Moore JE, Von Korff M, Cherkin D, Saunders K, Lorig K. A randomized trial of a cognitive-behavioural program for enhancing back pain self care in a primary care setting. *Pain* 2000;**88**:145–53. [http://dx.doi.org/10.1016/S0304-3959\(00\)00314-6](http://dx.doi.org/10.1016/S0304-3959(00)00314-6)
106. Cecchi F, Molino-Lova R, Chiti M, Pasquini G, Paperini A, Conti AA, et al. Spinal manipulation compared with back school and with individually delivered physiotherapy for the treatment of chronic low back pain: a randomized trial with one-year follow-up. *Clin Rehabil* 2010;**24**:26–36. <http://dx.doi.org/10.1177/0269215509342328>

107. Thomas KJ, MacPherson H, Ratcliffe J, Thorpe L, Brazier J, Campbell M, *et al.* Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain. *Health Technol Assess* 2005;**9**(32). <http://dx.doi.org/10.3310/hta9320>
108. Goldstein MS, Morgenstern H, Hurwitz EL, Yu F. The impact of treatment confidence on pain and related disability among patients with low-back pain: results from the University of California, Los Angeles, low-back pain study. *Spine J* 2002;**2**:391–9; discussion 99–401. [http://dx.doi.org/10.1016/S1529-9430\(02\)00414-X](http://dx.doi.org/10.1016/S1529-9430(02)00414-X)
109. Hagen EM, Eriksen HR, Ursin H. Does early intervention with a light mobilization program reduce long-term sick leave for low back pain? *Spine (Phila Pa 1976)* 2000;**1**;25:1973–6. <http://dx.doi.org/10.1097/00007632-200008010-00017>
110. Sherman KJ, Cherkin DC, Ichikawa L, Avins AL, Delaney K, Barlow WE, *et al.* Treatment expectations and preferences as predictors of outcome of acupuncture for chronic back pain. *Spine (Phila Pa 1976)* 2010;**35**:1471–7. <http://dx.doi.org/10.1097/BRS.0b013e3181c2a8d3>
111. Eisenberg DM, Post DE, Davis RB, Connelly MT, Legedza AT, Hrbek AL, *et al.* Addition of choice of complementary therapies to usual care for acute low back pain: a randomized controlled trial. *Spine (Phila Pa 1976)* 2007;**32**:151–8. <http://dx.doi.org/10.1097/01.brs.0000252697.07214.65>
112. Albaladejo C, Kovacs FM, Royuela A, del Pino R, Zamora J. The efficacy of a short education program and a short physiotherapy program for treating low back pain in primary care: a cluster randomized trial. *Spine (Phila Pa 1976)* 2010;**35**:483–96. <http://dx.doi.org/10.1097/BRS.0b013e3181b9c9a7>
113. Goldby LJ, Moore AP, Doust J, Trew ME. A randomized controlled trial investigating the efficiency of musculoskeletal physiotherapy on chronic low back disorder. *Spine (Phila Pa 1976)* 2006;**31**:1083–93. <http://dx.doi.org/10.1097/01.brs.0000216464.37504.64>
114. Skargren EI, Carlsson PG, Oberg BE. One-year follow-up comparison of the cost and effectiveness of chiropractic and physiotherapy as primary management for back pain. Subgroup analysis, recurrence, and additional health care utilization. *Spine* 1998;**23**:1875–83; discussion 84. <http://dx.doi.org/10.1097/00007632-199809010-00016>
115. Heymans MW, de Vet HC, Bongers PM, Knol DL, Koes BW, van Mechelen W. The effectiveness of high-intensity versus low-intensity back schools in an occupational setting: a pragmatic randomized controlled trial. *Spine (Phila Pa 1976)* 2006;**31**:1075–82. <http://dx.doi.org/10.1097/01.brs.0000216443.46783.4d>
116. Frost H, Lamb SE, Doll HA, Carver PT, Stewart-Brown S. Randomised controlled trial of physiotherapy compared with advice for low back pain. *BMJ* 2004;**329**:708. <http://dx.doi.org/10.1136/bmj.38216.868808.7C>
117. Petersen T, Larsen K, Jacobsen S. One-year follow-up comparison of the effectiveness of McKenzie treatment and strengthening training for patients with chronic low back pain: outcome and prognostic factors. *Spine (Phila Pa 1976)* 2007;**32**:2948–56. <http://dx.doi.org/10.1097/BRS.0b013e31815cda4a>
118. Koes BW, Bouter LM, van Mameren H, Essers AH, Verstegen GJ, Hofhuizen DM, *et al.* A randomized clinical trial of manual therapy and physiotherapy for persistent back and neck complaints: subgroup analysis and relationship between outcome measures. *J Manipulative Physiol Ther* 1993;**16**:211–19.
119. Hagen EM, Odelien KH, Lie SA, Eriksen HR. Adding a physical exercise programme to brief intervention for low back pain patients did not increase return to work. *Scand J Public Health* 2010;**38**:731–8. <http://dx.doi.org/10.1177/1403494810382472>

120. Poole H, Glenn S, Murphy P. A randomised controlled study of reflexology for the management of chronic low back pain. *Eur J Pain* 2007;**11**:878–87. <http://dx.doi.org/10.1016/j.ejpain.2007.01.006>
121. Linton SJ, Andersson T. Can chronic disability be prevented? A randomized trial of a cognitive-behavior intervention and two forms of information for patients with spinal pain. *Spine* 2000;**25**:2825–31; discussion 24. <http://dx.doi.org/10.1097/00007632-200011010-00017>
122. Hurley DA, McDonough SM, Dempster M, Moore AP, Baxter GD. A randomized clinical trial of manipulative therapy and interferential therapy for acute low back pain. *Spine* 2004;**29**:2207–16. <http://dx.doi.org/10.1097/01.brs.0000142234.15437.da>
123. Hondras MA, Long CR, Cao Y, Rowell RM, Meeker WC. A randomized controlled trial comparing 2 types of spinal manipulation and minimal conservative medical care for adults 55 years and older with subacute or chronic low back pain. *J Manipulative Physiol Ther* 2009;**32**:330–43. <http://dx.doi.org/10.1016/j.jmpt.2009.04.012>
124. Berwick DM, Budman S, Feldstein M. No clinical effect of back schools in an HMO. A randomized prospective trial. *Spine* 1989;**14**:338–44. <http://dx.doi.org/10.1097/00007632-198903000-00016>
125. Damush TM, Weinberger M, Perkins SM, Rao JK, Tierney WM, Qi R, *et al.* The long-term effects of a self-management program for inner-city primary care patients with acute low back pain. *Arch Intern Med* 2003;**163**:2632–8. <http://dx.doi.org/10.1001/archinte.163.21.2632>
126. Triano JJ, McGregor M, Hondras MA, Brennan PC. Manipulative therapy versus education programs in chronic low back pain. *Spine* 1995;**20**:948–55. <http://dx.doi.org/10.1097/00007632-199504150-00013>
127. Niemisto L, Lahtinen-Suopanki T, Rissanen P, Lindgren KA, Sarna S, Hurri H. A randomized trial of combined manipulation, stabilizing exercises, and physician consultation compared to physician consultation alone for chronic low back pain. *Spine* 2003;**28**:2185–91. <http://dx.doi.org/10.1097/01.BRS.0000085096.62603.61>
128. Shirado O, Doi T, Akai M, Hoshino Y, Fujino K, Hayashi K, *et al.* Multicenter randomized controlled trial to evaluate the effect of home-based exercise on patients with chronic low back pain: the Japan low back pain exercise therapy study. *Spine (Phila Pa 1976)* 2010;**35**:E811–19. <http://dx.doi.org/10.1097/BRS.0b013e3181d7a4d2>
129. Anema JR, Steenstra IA, Bongers PM, de Vet HC, Knol DL, Loisel P, *et al.* Multidisciplinary rehabilitation for subacute low back pain: graded activity or workplace intervention or both? A randomized controlled trial. *Spine (Phila Pa 1976)* 2007;**32**:291–8; discussion 99–300. <http://dx.doi.org/10.1097/01.brs.0000253604.90039.ad>
130. Alaranta H, Rytokoski U, Rissanen A, Talo S, Ronnema T, Puukka P, *et al.* Intensive physical and psychosocial training program for patients with chronic low back pain. A controlled clinical trial. *Spine* 1994;**19**:1339–49. <http://dx.doi.org/10.1097/00007632-199406000-00007>
131. Hancock MJ, Maher CG, Latimer J, McLachlan AJ, Cooper CW, Day RO, *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* 2007;**370**:1638–43. [http://dx.doi.org/10.1016/S0140-6736\(07\)61686-9](http://dx.doi.org/10.1016/S0140-6736(07)61686-9)
132. Haake M, Muller HH, Schade-Brittinger C, Basler HD, Schafer H, Maier C, *et al.* German Acupuncture Trials (GERAC) for chronic low back pain: randomized, multicenter, blinded, parallel-group trial with 3 groups. *Arch Int Med* 2007;**167**:1892–8. <http://dx.doi.org/10.1001/Archinte.167.17.1892>
133. Moffett JK, Torgerson D, Bell-Syer S, Jackson D, Llewellyn-Phillips H, Farrin A, *et al.* Randomised controlled trial of exercise for low back pain: clinical outcomes, costs, and preferences. *BMJ* 1999;**319**:279–83. <http://dx.doi.org/10.1136/bmj.319.7205.279>

134. Macedo LG, Latimer J, Maher CG, Hodges PW, McAuley JH, Nicholas MK, 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. <http://dx.doi.org/10.2522/ptj.20110290>
135. Carlsson CP, Sjolund BH. Acupuncture for chronic low back pain: a randomized placebo-controlled study with long-term follow-up. *Clin J Pain* 2001;**17**:296–305. <http://dx.doi.org/10.1097/00002508-200112000-00003>
136. Kennedy S, Baxter GD, Kerr DP, Bradbury I, Park J, McDonough SM. Acupuncture for acute non-specific low back pain: a pilot randomised non-penetrating sham controlled trial. *Complement Ther Med* 2008;**16**:139–46. <http://dx.doi.org/10.1016/j.ctim.2007.03.001>
137. Jellema P, van der Roer N, van der Windt DA, van Tulder MW, van der Horst HE, Stalman WA, et al. Low back pain in general practice: cost-effectiveness of a minimal psychosocial intervention versus usual care. *Euro Spine J* 2007;**16**:1812–21. <http://dx.doi.org/10.1007/s00586-007-0439-2>
138. Kainz B, Gulich M, Engel EM, Jackel WH. [Comparison of three outpatient therapy forms for treatment of chronic low back pain: findings of a multicentre, cluster randomized study.] *Die Rehabilitation* 2006;**45**:65–77. <http://dx.doi.org/10.1055/s-2005-915338>
139. Long A, Donelson R, Fung T. Does it matter which exercise? A randomized control trial of exercise for low back pain. *Spine* 2004;**29**:2593–602. <http://dx.doi.org/10.1097/01.brs.0000146464.23007.2a>
140. Von Korff M, Moore JE, Lorig K, Cherkin DC, Saunders K, Gonzalez VM, et al. A randomized trial of a lay person-led self-management group intervention for back pain patients in primary care. *Spine* 1998;**23**:2608–15. <http://dx.doi.org/10.1097/00007632-199812010-00016>
141. Whitehurst DG, Lewis M, Yao GL, Bryan S, Raftery JP, Mullis R, et al. A brief pain management program compared with physical therapy for low back pain: results from an economic analysis alongside a randomised clinical trial. *Arthritis Rheum* 2007;**57**:466–73. <http://dx.doi.org/10.1002/art.22606>
142. Ratcliffe J, Thomas KJ, MacPherson H, Brazier J. A randomised controlled trial of acupuncture care for persistent low back pain: cost-effectiveness analysis. *BMJ* 2006;**333**:626. <http://dx.doi.org/10.1136/bmj.38932.806134.7C>
143. Hancock MJ, Maher CG, Latimer J, Herbert RD, McAuley JH. Can rate of recovery be predicted in patients with acute low back pain? Development of a clinical prediction rule. *Eur J Pain* 2009;**13**:51–5. <http://dx.doi.org/10.1016/j.ejpain.2008.03.007>
144. Roland M, Waddell G, Klaber Moffett J, Burton K, Main C. *The Back Book: the Best Way to Deal with Back Pain; Get Back Active*. London: The Stationery Office; 2002.
145. Vickers AJ, Cronin AM, Maschino AC, Lewith G, MacPherson H, Foster NE, et al. Acupuncture for chronic pain: individual patient data meta-analysis. *Arch Int Med* 2012;**172**:1444–53. <http://dx.doi.org/10.1001/archinternmed.2012.3654>
146. Codd EF. *The Relational Model for Database Management: Version 2*. Boston, MA: Addison-Wesley Longman Publishing Co; 1990.
147. Marengo L, Tosches N, Crasto C, Shepherd G, Miller PL, Nadkarni PM. Achieving evolvable Web-database bioscience applications using the EAV/CR framework: recent advances. *J Am Med Inform Assoc* 2003;**10**:444–53. <http://dx.doi.org/10.1197/jamia.M1303>
148. World Wide Web Consortium (W3C). *XML Technology*. 2010. URL: www.w3.org/standards/xml/ (accessed 29 August 2014).
149. Morris T, Hee SW, Stallard N, Underwood M, Patel S. Can we convert between outcome measures of disability for chronic low back pain? *Spine* 2015;**40**:734–9. <http://dx.doi.org/10.1097/BRS.0000000000000866>

150. Von Korff M, Ormel J, Keefe FJ, Dworkin SF. Grading the severity of chronic pain. *Pain* 1992;**50**:133–49. [http://dx.doi.org/10.1016/0304-3959\(92\)90154-4](http://dx.doi.org/10.1016/0304-3959(92)90154-4)
151. Kohlmann T, Raspe H. [Hannover Functional Questionnaire in ambulatory diagnosis of functional disability caused by backache.] *Die Rehabilitation* 1996;**35**:i–viii.
152. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy* 1980;**66**:271–3.
153. Tait RC, Chibnall JT, Krause S. The Pain Disability Index: psychometric properties. *Pain* 1990;**40**:171–82. [http://dx.doi.org/10.1016/0304-3959\(90\)90068-O](http://dx.doi.org/10.1016/0304-3959(90)90068-O)
154. Stratford P. Assessing disability and change on individual patients: a report of a patient specific measure. *Physiother Can* 1995;**47**:258–63. <http://dx.doi.org/10.3138/ptc.47.4.258>
155. Ware J, Kosinski M, Turner-Bowker D, Gandek B. *How to Score Version 2 of the SF-12® Health Survey (With a Supplement Documenting Version 1)*. Lincoln, RI: QualityMetric Inc; 2002.
156. Ware J, Kosinski M, Dewey J. *How to Score Version 2 of the SF-36 Health Survey*. Lincoln, RI: QualityMetric Inc; 2000.
157. Deyo RA, Dworkin SF, Amtmann D, Andersson G, Borenstein D, Carragee E, et al. Report of the NIH Task Force on research standards for chronic low back pain. *Spine J* 2014;**14**:1375–91. <http://dx.doi.org/10.1016/j.spinee.2014.05.002>
158. Puhan MA, Soesilo I, Guyatt GH, Schunemann HJ. Combining scores from different patient reported outcome measures in meta-analyses: when is it justified? *Health Qual Life Outcomes* 2006;**4**:94. <http://dx.doi.org/10.1186/1477-7525-4-94>
159. Agresti A. *Categorical Data Analysis*. 2nd edn. Hoboken, NJ: Wiley; 2002. <http://dx.doi.org/10.1002/0471249688>
160. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74. <http://dx.doi.org/10.2307/2529310>
161. Dolan P, Gudex C, Kind P. *A Social Tariff for EuroQol: Results from a UK General Population Survey*. York: Center for Health Economics, University of York; 1995.
162. Gray AM, Rivero-Arias O, Clarke PM. Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. *Med Decis Mak* 2006;**26**:18–29. <http://dx.doi.org/10.1177/0272989X05284108>
163. Rowen D, Brazier J, Roberts J. Mapping SF-36 onto the EQ-5D index: how reliable is the relationship? *Health Qual Life Outcomes* 2009;**7**:27. <http://dx.doi.org/10.1186/1477-7525-7-27>
164. Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. *Health Econ* 2005;**14**:487–96. <http://dx.doi.org/10.1002/hec.944>
165. Melzack R. The Short-Form McGill Pain Questionnaire. *Pain* 1987;**30**:191–7. [http://dx.doi.org/10.1016/0304-3959\(87\)91074-8](http://dx.doi.org/10.1016/0304-3959(87)91074-8)
166. Whitehead A. *Meta-Analysis of Controlled Clinical Trials*. In Senn S, Barnett V, editors. Chichester: John Wiley & Sons; 2003.
167. Mistry D. *Recursive Partitioning based Approaches for Low Back Pain Subgroup Identification in Individual Participant Data Meta-analyses*. Coventry: University of Warwick; 2014.
168. Zhang Z, Singer B. *Recursive Partitioning and Applications*. 2nd edn. New York, NY: Springer; 2010. <http://dx.doi.org/10.1007/978-1-4419-6824-1>

169. LeBlanc M, Crowley J. Survival trees by goodness of split. *J Am Stat Assoc* 1993;**88**:457–67. <http://dx.doi.org/10.1080/01621459.1993.10476296>
170. Doyle P. The use of automatic interaction detector and similar search procedures. *Oper Res Quart* 1973;**24**:465–7. <http://dx.doi.org/10.1057/jors.1973.81>
171. Shih Y-S, Tsai H-W. Variable selection bias in regression trees with constant fits. *Comput Stat Data Anal* 2004;**45**:595–607. [http://dx.doi.org/10.1016/S0167-9473\(03\)00036-7](http://dx.doi.org/10.1016/S0167-9473(03)00036-7)
172. LeBlanc M, Moon J, Crowley J. Adaptive risk group refinement. *Biometrics* 2005;**61**:370–8. <http://dx.doi.org/10.1111/j.1541-0420.2005.020738.x>
173. National Institute for Health and Care Excellence. *Guide to the Methods of Technology Appraisal*. London: NICE; 2013.
174. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med* 2002;**21**:2313–24. <http://dx.doi.org/10.1002/sim.1201>
175. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* 2005;**331**:897–900. <http://dx.doi.org/10.1136/bmj.331.7521.897>
176. Cooper NJ, Peters J, Lai MC, Juni P, Wandel S, Palmer S, et al. How valuable are multiple treatment comparison methods in evidence-based health-care evaluation? *Value Health* 2011;**14**:371–80. <http://dx.doi.org/10.1016/j.jval.2010.09.001>
177. Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics* 2008;**26**:753–67. <http://dx.doi.org/10.2165/00019053-200826090-00006>
178. Jansen JP. Network meta-analysis of individual and aggregate level data. *Res Synth Methods* 2012;**3**:177–90. <http://dx.doi.org/10.1002/jrsm.1048>
179. Donegan S, Williamson P, D'Alessandro U, Tudur Smith C. Assessing the consistency assumption by exploring treatment by covariate interactions in mixed treatment comparison meta-analysis: individual patient-level covariates versus aggregate trial-level covariates. *Stat Med* 2012;**31**:3840–57. <http://dx.doi.org/10.1002/sim.5470>
180. Dias S, Welton NJ, Sutton AJ, Ades AE. *NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials*. London: NICE; 2011.
181. Bouter LM, Pennick V, Bombardier C. Cochrane back review group. *Spine* 2003;**28**:1215–18. <http://dx.doi.org/10.1097/01.BRS.0000065493.26069.1C>
182. Hill JC, Dunn KM, Lewis M, Mullis R, Main CJ, Foster NE, et al. A primary care back pain screening tool: identifying patient subgroups for initial treatment. *Arthritis Rheum* 2008;**59**:632–41. <http://dx.doi.org/10.1002/art.23563>
183. Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. *J Clin Epidemiol* 2013;**66**:173–83. <http://dx.doi.org/10.1016/j.jclinepi.2012.08.001>
184. Hayden JA, Cartwright JL, Riley RD, Vantulder MW. Exercise therapy for chronic low back pain: protocol for an individual participant data meta-analysis. *Syst Rev* 2012;**1**:64. <http://dx.doi.org/10.1186/2046-4053-1-64>
185. Ferreira ML, Herbert RD, Crowther MJ, Verhagen A, Sutton AJ. When is a further clinical trial justified? *BMJ* 2012;**345**:e5913. <http://dx.doi.org/10.1136/bmj.e5913>

186. Patel S, Ngunjiri A, Hee SW, Yang Y, Brown S, Friede T, *et al.* Primum non nocere: shared informed decision making in low back pain: a pilot cluster randomised trial. *BMC Musculoskeletal Disord* 2014;**15**:282. <http://dx.doi.org/10.1186/1471-2474-15-282>
187. Hilfiker R, Bachmann LM, Heitz CA, Lorenz T, Joronen H, Klipstein A. Value of predictive instruments to determine persisting restriction of function in patients with subacute non-specific low back pain. Systematic review. *Eur Spine J* 2007;**16**:1755–75. <http://dx.doi.org/10.1007/s00586-007-0433-8>
188. Kent PM, Keating JL. Can we predict poor recovery from recent-onset nonspecific low back pain? A systematic review. *Manual Ther* 2008;**13**:12–28. <http://dx.doi.org/10.1016/j.math.2007.05.009>
189. Wessels T, van Tulder M, Sigl T, Ewert T, Limm H, Stucki G. What predicts outcome in non-operative treatments of chronic low back pain? A systematic review. *Eur Spine J* 2006;**15**:1633–44. <http://dx.doi.org/10.1007/s00586-006-0073-4>
190. Denison E, Asenlof P, Lindberg P. Self-efficacy, fear avoidance, and pain intensity as predictors of disability in subacute and chronic musculoskeletal pain patients in primary health care. *Pain* 2004;**111**:245–52. <http://dx.doi.org/10.1016/j.pain.2004.07.001>
191. Grotle M, Brox JJ, Veierød MB, Glomsrød B, Lønn JH, Vollestad NK. Clinical course and prognostic factors in acute low back pain: patients consulting primary care for the first time. *Spine* 2005;**30**:976–82. <http://dx.doi.org/10.1097/01.brs.0000158972.34102.6f>
192. Henschke N, Maher CG, Refshauge KM, Herbert RD, Cumming RG, Bleasel J, *et al.* Prognosis in patients with recent onset low back pain in Australian primary care: inception cohort study. *BMJ* 2008;**337**:a171. <http://dx.doi.org/10.1136/bmj.a171>
193. Colquhoun D, Novella SP. Acupuncture is theatrical placebo. *Anaesth Analg* 2013;**116**:1360–3. <http://dx.doi.org/10.1213/ANE.0b013e31828f2d5e>

Appendix 1 Review 2: summary of excluded papers

Reproduced from Mistry D, Patel S, Hee SW, Stallard N, Underwood M. Evaluating the quality of subgroup analyses in randomized controlled trials of therapist-delivered interventions for nonspecific low back pain: a systematic review. *Spine* 2014;**39**:618–29; with permission from Lippincott Williams & Wilkins.

Paper	Reason for exclusion
Childs JD, Flynn TW, Fritz JM. A perspective for considering the risks and benefits of spinal manipulation in patients with low back pain. <i>Manual Therapy</i> 2006;11:316-20	Testing a clinical prediction rule
Costa LO, Maher CG, Latimer J, Hodges PW, Herbert RD, Refshauge KM <i>et al.</i> Motor control exercise for chronic low back pain: a randomized placebo-controlled trial. <i>Physical Therapy</i> 2009;89:1275-86.	Look at effect modification over time
Faas A, Chavannes AW, van Eijk JT, Gubbels JW. A randomized, placebo-controlled trial of exercise therapy in patients with acute low back pain. <i>Spine</i> 1993;18:1388-95.	Included patients aged less than 18 years
Faas A, van Eijk JT, Chavannes AW, Gubbels JW. A randomized trial of exercise therapy in patients with acute low back pain. Efficacy on sickness absence. <i>Spine</i> 1995;20:941-7.	Included patients aged less than 18 years and outcome in sub-group analyses not a clinical measure of low back pain (sickness absence)
George SZ, Fritz JM, Childs JD, Brennan GP. Sex differences in predictors of outcome in selected physical therapy interventions for acute low back pain. <i>Journal of Orthopaedic & Sports Physical Therapy</i> 2006;36:354-63.	Pooled datasets of similar trials
George SZ, Zeppieri G, Jr., Cere AL, Cere MR, Borut MS, Hodges MJ <i>et al.</i> A randomized trial of behavioral physical therapy interventions for acute and sub-acute low back pain (NCT00373867). <i>Pain</i> 2008;140:145-57.	Included patients aged less than 18 years and also looked at effect modification over time

Paper	Reason for exclusion
Haas M, Grouppe E, Muench J, Kraemer D, Brummel-Smith K, Sharma R <i>et al.</i> Chronic disease self-management program for low back pain in the elderly. <i>Journal of Manipulative & Physiological Therapeutics</i> 2005;28:228-37.	Intervention not delivered by therapist
Hagen EM, Svensen E, Eriksen HR. Predictors and modifiers of treatment effect influencing sick leave in subacute low back pain patients. <i>Spine</i> 2005;30:2717-23.	Outcome in sub-group analyses not a clinical measure of low back pain (return to work)
Hancock MJ, Maher CG, Latimer J, Herbert RD, McAuley JH. Independent evaluation of a clinical prediction rule for spinal manipulative therapy: a randomised controlled trial. <i>European Spine Journal</i> 2008;17:936-43.	Testing a clinical prediction rule
Jellema P, van der Windt DA, van der Horst HE, Twisk JW, Stalman WA, Bouter LM. Should treatment of (sub)acute low back pain be aimed at psychosocial prognostic factors? Cluster randomised clinical trial in general practice. <i>BMJ</i> 2005;331:84.	Look at effect modification over time
Jellema P, van der Roer N, van der Windt DA, van Tulder MW, van der Horst HE, Stalman WA <i>et al.</i> Low back pain in general practice: cost-effectiveness of a minimal psychosocial intervention versus usual care. <i>European Spine Journal</i> 2007;16:1812-21.	Outcome in sub-group analyses not a clinical measure of low back pain (cost-effectiveness)

Paper	Reason for exclusion
Kool JP, Oesch PR, Bachmann S, Knuesel O, Dierkes JG, Russo M <i>et al.</i> Increasing days at work using function-centered rehabilitation in nonacute nonspecific low back pain: a randomized controlled trial. <i>Archives of Physical Medicine & Rehabilitation</i> 2005;86:857-64.	Outcome in sub-group analyses not a clinical measure of low back pain (days worked over 3 months)
Lamb SE, Lall R, Hansen Z, Castelnuovo E, Withers EJ, Nichols V <i>et al.</i> A multicentred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain. The Back Skills Training (BeST) trial. <i>Health Technology Assessment (Winchester, England)</i> 2014;14:1-253.	HTA report. Secondary sub-groups analyses paper published elsewhere and used instead (Underwood 2011)
Scheel IB, Hagen KB, Herrin J, Oxman AD. A randomized controlled trial of two strategies to implement active sick leave for patients with low back pain. <i>Spine</i> 2002;27:561-6.	Outcome in sub-group analyses not a clinical measure of low back pain (active sick leave)
Skargren EI, Carlsson PG, Oberg BE. One-year follow-up comparison of the cost and effectiveness of chiropractic and physiotherapy as primary management for back pain. Sub-group analysis, recurrence, and additional health care utilization. <i>Spine</i> 1998;23:1875-83.	Looked at an addition disorder (neck pain)
Skargren EI, Oberg BE, Carlsson PG, Gade M. Cost and effectiveness analysis of chiropractic and physiotherapy treatment for low back and neck pain. Six-month follow-up. <i>Spine</i> 1997;22:2167-77.	Looked at an addition disorder (neck pain)

Paper	Reason for exclusion
Staal JB, Hlobil H, Koke AJ, Twisk JW, Smid T, van MW. Graded activity for workers with low back pain: who benefits most and how does it work? <i>Arthritis & Rheumatism</i> 2008;59:642-9.	Outcome in sub-group analyses not a clinical measure of low back pain (return to work)
Steenstra IA, Knol DL, Bongers PM, Anema JR, van MW, de Vet HC. What works best for whom? An exploratory, sub-group analysis in a randomized, controlled trial on the effectiveness of a workplace intervention in low back pain patients on return to work. <i>Spine</i> 2009;34:1243-9.	Outcome in sub-group analyses not a clinical measure of low back pain (return to work)
Thomas KJ, MacPherson H, Ratcliffe J, Thorpe L, Brazier J, Campbell M <i>et al.</i> Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain. <i>Health Technology Assessment (Winchester, England)</i> /1/10;9:iii-iv.	HTA report. Secondary sub-groups analyses paper published elsewhere and used instead (Thomas 2006)
Toda Y. Impact of waist/hip ratio on the therapeutic efficacy of lumbosacral corsets for chronic muscular low back pain. <i>Journal of Orthopaedic Science</i> 2002;7:644-9.	Intervention not delivered by therapist (Corsets given to patients)
van Poppel MN, Koes BW, van der Ploeg T, Smid T, Bouter LM. Lumbar supports and education for the prevention of low back pain in industry: a randomized controlled trial. <i>JAMA</i> 1998;279:1789-94.	Intervention not delivered by therapist (Lumbar supports given to patients)

Reproduced with permission from Lippincott, Williams & Wilkins publishers

Appendix 2 Invitation letter

Warwick
Medical School

CLINICAL TRIALS UNIT

Professor Martin Underwood

Warwick Medical School Clinical Trials Unit

University of Warwick

Coventry

CV4 7AL

[INSERT ADDRESS]

[INSERT DATE]

Study Title: Improving outcomes from the treatment of back pain

Dear [INSERT NAME]

We have successfully obtained funding from the National Institute for Health Research for a programme grant on the management of low back pain. One aspect of this programme is to develop a pooled database of the original data from randomised controlled trials of therapist delivered interventions for low back pain.

The overall aim of our programme grant is to improve the clinical and cost-effectiveness of low back pain treatment by providing patients, their clinical advisors, and health service purchasers with better information about which patients are most likely to benefit from which treatment choices.

By developing this repository of original patient data we hope to conduct pooled secondary analyses. This will help us to determine which patient characteristics, if any, predict clinical response to different treatments for low back pain and/or predict the most cost-effective treatments for low back pain.

We would be very grateful if you would consider sharing the data from your [INSERT STUDY] trial for this important study. If you have any questions or you are interested in sharing this data with us please could you email repository@warwick.ac.uk in the first instance.

We look forward to hearing from you.

Yours sincerely

Martin Underwood
Professor of Primary Care Research

Appendix 3 Information sheet

Low Back Pain Trial Repository Programme Information Sheet for Investigators Programme Summary and Investigator Involvement

Warwick CTU has been funded by UK National Institute of Health Research to do individual patient data meta-analysis of data from trials of low back pain treatments. We are inviting custodians of existing trial datasets to contribute data to this project. There are two stages to this; the first stage is for our currently funded project to explore sub-groups in low back pain (LBP) and the second stage is to maintain a data repository of individual patient data from trials of therapist delivered intervention in low back pain as a resource for the back pain community. The Chief investigator for this project is Martin Underwood.

Stage 1: Improving outcomes from the treatment of back pain

At a population level, we have useful data on the management of LBP. What is not clear is how we can use these data to maximise the treatment benefit for the individual patient i.e. which patients are most likely to benefit from which treatment choices. If we could predict which patients would be most likely to benefit from different treatments, overall effectiveness, and cost-effectiveness, of treatments for Low Back Pain would improve. Any randomised controlled trial (RCT) to directly address this problem would need to be very large.

We have received funding from the NIHR to undertake an individual patient data meta-analysis to identify moderators of treatment effect. From this programme of research, we aim to produce evidence to help patients, their clinical advisors and health service purchasers to select the 'right treatment for the right person at the right time'. We are interested in both clinical and cost-effectiveness.

We have obtained ethical approval for this project from both the University of Warwick's Biological Research Ethics Committee and also a UK National Health Service research ethics committee flagged to assess applications to establish a research database. We have of course considered ethical issues of secondary analysis of data carefully. We will only request and utilise anonymous data and will seek assurance from collaborators that nothing in the original consent process would preclude sharing anonymous data in this way.

In this first stage, once we have sufficient data, we will explore how the complex relationship between demographic factors, patient history and patient characteristics can be used to predict the response to different treatments. We will;

1. estimate within-trial indicators of clinical and economic outcomes at the individual patient level (e.g. health care costs and QALYs over the trial period),
2. statistically analyse the RCT dataset to identify moderators that could contribute to a practical Clinical Prediction Rule that can be used to inform LBP management.

We would like you to share data for this work. Ideally we want to include individual item responses to outcome measures rather than summary values in order that we can ensure consistency in how summary scores are calculated. However, we would like to stress that you

are under no obligation to send us any data you wouldn't wish to share. If you only have summary measures available we would still be delighted to have your data. We are particularly interested in any data that will inform our cost-effectiveness analyses.

We would like you to share the following data with us:

- Participant characteristics and baseline measurements
- Assigned intervention(s)
- Intervention(s) received
- Recorded outcomes at each time point (during the intervention and follow-up) including
 - Values of individual items from all the questionnaires
 - Health economic/utility measurements (e.g. EQ5D or SF6D items)
- Recorded use of health services and related expenditure for patients (during intervention and follow-up)
- Anonymised data allowing us to measure any clustering by therapist or site

If they are available and you are happy to share them with us then copies of the following documents:

- The final protocol
- Case report forms (CRF)
- Coding manual for the CRF codes

We are aware that these documents may not be available – for example we know of one large study that lost all its archived material in a flood. Whatever you have available would be very helpful to the team.

Upon receiving the dataset we will run a validity and quality check to ensure data integrity. A validity-quality report will be sent to you for comment and/or feedback. We aim to resolve any inconsistencies in the data before integrating the dataset with the rest of the dataset in the repository. Once the dataset has been integrated into the repository, the original dataset from you will be destroyed.

We have established secure methods to transfer anonymous data sets and will send you full details when appropriate. We are only too aware of how hard it was to collect these data in the first place; will handle them very carefully!

At present we are asking for data sharing agreements for this study only. We will produce a new data sharing agreement for stage two of the project.

All research teams who contribute to the project will be acknowledged in any publications. Where possible, we will do this by including one member of each trial team as a named member of the collaborative group who have supported this programme; you may choose whom is acknowledged. This may be a different person for each set of trial data you share with us. This will ensure your contribution will be recognised by PubMed and citation tracking. We will give you the opportunity to comment on any papers that have used your data prior to submission. You will not, however, be obliged to comment.

Stage 2: Future use of the repository

Once developed, we would like to maintain this pooled data set as a resource for the research community as we anticipate that there will be many future research questions to be asked from this data set. Therefore any shared data sets will need to be as complete as possible as we will only be able to put each study into the repository once; this is why we are asking for such a detailed dataset for stage one of the project.

We will establish a governance structure including an independent steering committee to oversee fair access to the data by ourselves and others in the future. As a collaborator we would welcome any application to utilise this data (subject to steering committee approval). I do not anticipate needing to charge for access to these data. We will be seeking additional funding to maintain and add to the pooled dataset as a resource for the back pain research community.

We will be looking for additional funding to continue supporting the database and adding further trial data sets in the future.

We will ask for separate and additional consent from you to include your data in phase 2. If you do not wish any of your data to be used in any subsequent analyses, you will be able to specify this at this point. Please be assured that we will not use your data for any other analyses than those stipulated by you and those which have received approval from the steering committee.

Thank you for taking the time to read this information and we hope that you will consider our request to share your data and contribute to this valuable programme.

Repository Programme Team:
Professor Martin Underwood (Chief Investigator)
Professor of Primary Care Research
University of Warwick
Warwick Clinical Trials Unit
Gibbet Hill Campus
Coventry
CV4 7AL
Tel: XXXX
Email: [XXXX](#) or [XXXX](#)

Professor Nigel Stallard, Professor of Medical Statistics, University of Warwick

Professor Tim Friede, Professor of Biostatistics, University Medical Center Göttingen

Professor Sallie Lamb, Professor of Rehabilitation, Warwick Clinical Trials Unit

Dr Shilpa Patel, Research Fellow, University of Warwick

Dr Joanne Lord, Reader Health Economics Research Group, Brunel University

Dr David Ellard, Senior Research Fellow, University of Warwick

Appendix 4 Sample data sharing agreement

Data sharing agreement

Standard template

Research project title: Improving outcomes from the treatment of back pain

Reference: RP-PG-0608–10076

1.0 Organisations

This Data Sharing Agreement is drawn up between:

Professor Martin Underwood

Warwick Clinical Trials Unit

University of Warwick

Gibbet Hill

Coventry

CV4 7AL

and:

[INSERT DETAILS]

2.0 Period of agreement

This agreement commences on [INSERT DATE] and will terminate on [INSERT DATE] unless extended by mutual agreement of both parties in writing, at which point an Amendment will be issued by University of Warwick to replace this document.

3.0 Data required

[INSERT INSTITUTION NAME] will supply all anonymous trial data from [INSERT TRIAL NAME].

Data required:

Individual patient data with descriptions of variable coding

AND/OR

Scored variable databases with descriptions of variable coding

We will require confirmation from the Chief Investigator that patients in the original trial have given informed consent.

4.0 Permissions

The data will come from completed randomised controlled trials. All data will be anonymous and no patient identifiable information will be shared.

Approval to obtain data will be obtained from the University of Warwick's Biological Research Ethics Committee and the Oxford 'C' NHS REC.

5.0 Purpose for which the data are to be used

The data will be used to develop a repository of IPD on potential moderators, health outcomes, and health-care resource use and costs, from RCTs testing therapist-delivered interventions for low back pain. We will conduct statistical and health-economic analyses on this pooled data set.

We will not reanalyse any trial data already published.

Data access is restricted to those named in *Table A* of this agreement. Any changes will be notified to [INSERT INSTITUTION NAME].

6.0 User obligations

The University of Warwick formally wishes to acknowledge its explicit commitment to maintaining the confidentiality, safety, security and integrity of all data to which the organisation is privy and which may be held under its guardianship.

The University of Warwick continues to legitimately enter into formal agreement and/or implicit undertaking with all its clients, staff, visitors, suppliers and others, in recognition of the fact that the data are held under the guardianship of the University of Warwick which is pertinent to the individual client, staff member, visitor, supplier and/or other, will only be used for the explicit agreed purpose or purposes for which it has been provided, and that there will be no unlawful disclosure or loss of the same.

TABLE A Individuals who will have access to and use of the repository

Permitted users	Job title – organisation they work for – where they will access data
Martin Underwood	Chief investigator based at Warwick CTU – Medical School, data will be accessed within the university only
Shilpa Patel	Study Manager based at Warwick CTU – Medical School, data will be accessed within the university only
Sallie Lamb	Co-investigator based at Warwick CTU – Medical School, data will be accessed within the university only
Nigel Stallard	Statistical lead based at Warwick CTU – Medical School, data will be accessed within the university only
Tim Friede	Statistical advisor based at Göttingen University, data will be accessed within their institution
Statistician (Research Fellow)	Statistics Research based at Warwick CTU – Medical School, data will be accessed within the university only
Joanne Lord	Health Economist lead based at Brunel University, data will be accessed within their institution
Health Economist (Research Fellow)	Health Economist Research Fellow based at Brunel University, data will be accessed within their institution
Dipesh Mistry	PhD student based at Warwick CTU – Medical School, data will be accessed within the university only
Programming Team	Programming team based at Warwick CTU – Medical School, data will be accessed within the university only
Claire Daffern	Quality Assurance Manager at Warwick CTU – Medical School, data will be accessed within the university only

Users of the data supplied are obliged to fully comply with The Data Protection Act 1998, together with all other related and relevant legislation and Department of Health directives covering issues of data sharing and including:

- British (International) Standard ISO 27001
- The Caldicott Report 1997
- The Freedom of Information Act 2000
- Section 251 of the Health and Social Care Act 2006
- Confidentiality: NHS Code of Practice 2003
- NHS Records Management Code of Practice (Part 1, 2006 & Part 2, 2009)
- The NHS Information Security Management Code of Practice 2007
- The Computer Misuse Act 1990
- The Electronic Communications Act 2000
- The Regulation of Investigatory Powers Act 2000
- The Copyright, Designs and Patents Act 1988
- The Re-Use of Public Sector Information Regulations 2005
- The Human Rights Act 1998.

7.0 Transfer of data from [INSERT INSTITUTION NAME] and the University of Warwick

Anonymous data will be obtained from [INSERT INSTITUTION NAME]. Data will be encrypted and sent to the University of Warwick by [INSERT INSTITUTION NAME] via the University's file transfer application.

Once the data have been received, the original source will be moved to an encrypted drive. A processed copy of the data will be imported into a secure database.

Together with the encrypted data [INSERT INSTITUTION NAME] will provide a detailed description of the variables.

8.0 Storage of data

The original data source will be temporarily stored on a file server directory that is only accessible to the chief investigator and study manager until it is moved to an AES 256 encrypted volume. Data will be processed and imported from the encrypted volume into a Microsoft 2005 SQL Server database hosted in the University of Warwick's data centre. The data will be regularly replicated on to a failover server and routinely backed up to a Storage Area Network.

9.0 Data retention

The intention is to keep the repository once it has been developed and make it available to other researchers. An independent steering committee will be convened to assess applications for the repository.

If the repository is deemed to be no longer required, all data will be deleted from the servers. Deletion of data is irreversible and involves the database being disconnected and all data and transaction files being destroyed using a secure deletion application.

The WCTU may invoke the right to implement the research exemption clause of the Data Protection Act in order to retain the data for future research activities.

10.0 Agreement signatures

For and on behalf of:

Warwick Clinical Trials Unit

Signed:

Print name: Professor Martin Underwood

Post/title: Head of Division of Health Sciences, Warwick Medical School

Date:

For and on behalf of:

[INSERT INSTITUTION NAME]

Signed:

Print name:

Post/title:

Date:

Appendix 5 Instruction on secure data transfer

Repository programme

Instructions for transferring data sets to the University of Warwick

- Please ensure your data sets are anonymised.
- Compress/encrypt your data set using an open-source compression software programme (e.g. 7Zip)
- Follow this link: <https://files.warwick.ac.uk/repositorylbpdata/sendto>

WARWICK

Files.Warwick » Repository lbp data

The owner of the file space "Repository lbp data" will receive an email notification when you send them the file.

Your name

Your email

Message for the owner of the file space "Repository lbp data"

File - Maximum 250 MB

- Please fill in the boxes as required:
 - Your name
 - Your e-mail; and
 - Any message (e.g. name of the trial, contact telephone number)
- Click on the 'Browse' button
- Choose the file to upload
- Click on the 'Upload and send file' button.

A member of the Repository team will send an e-mail confirming that the data set has been uploaded successfully. We will also call you to obtain the password required to decrypt the file.

Thank you.

Appendix 6 Excluded studies

Paper	Trial	Number of participants
Brinkhaus B, Witt CM, Jena S, Linde K, Streng A, Irnich D, <i>et al.</i> Interventions and physician characteristics in a randomized multicenter trial of acupuncture in patients with low-back pain. <i>J Altern Complement Med</i> 2006; 12 :649–57	Brinkhaus	301
Hancock MJ, Maher CG, Latimer J, McLachlan AJ, Cooper CW, Day RO, <i>et al.</i> 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. <i>Lancet</i> 2007; 370 :1638–43	Hancock	240
Hancock MJ, Maher CG, Latimer J, Herbert RD, McAuley JH. Can rate of recovery be predicted in patients with acute low back pain? Development of a clinical prediction rule. <i>Eur J Pain</i> 2009; 13 :51–5	Hancock	240
Härkäpää K, Järvikoski A, Mellin G, Hurri H. A controlled study on the outcome of inpatient and outpatient treatment of low back pain. Part I. Pain, disability, compliance, and reported treatment benefits three months after treatment. <i>Scand J Rehabil Med</i> 1989; 21 :81–9	Härkäpää	459
Härkäpää K, Mellin G, Järvikoski A, Hurri H. A controlled study on the outcome of inpatient and outpatient treatment of low back pain. Part III. Long-term follow-up of pain, disability, and compliance. <i>Scand J Rehabil Med</i> 1990; 22 :181–8	Härkäpää	476
Hurwitz EL, Morgenstern H, Harber P, Kominski GF, Belin TR, Yu F, <i>et al.</i> A randomized trial of medical care with and without physical therapy and chiropractic care with and without physical modalities for patients with low back pain: 6-month follow-up outcomes from the UCLA low back pain study. <i>Spine</i> 2002; 27 :2193–204	Hurwitz	681
Hurwitz EL, Morgenstern H, Harber P, Kominski GF, Belin TR, Yu F, <i>et al.</i> The effectiveness of physical modalities among patients with low back pain randomized to chiropractic care: findings from the UCLA low back pain study. <i>J Manipulative Physiol Ther</i> 2002; 25 :10–20	Hurwitz	681
Hurwitz EL, Morgenstern H, Chiao C. Effects of recreational physical activity and back exercises on low back pain and psychological distress: findings from the UCLA Low Back Pain Study. <i>Am J Public Health</i> 2005; 95 :817–1824	Hurwitz	681
Hurwitz EL, Morgenstern H, Kominski GF, Yu F, Chiang LM. A randomized trial of chiropractic and medical care for patients with low back pain: eighteen-month follow-up outcomes from the UCLA low back pain study. <i>Spine (Phila Pa 1976)</i> 2006; 31 :611–21, discussion 22	Hurwitz	681
Koes BW, Bouter LM, van Mameren H, Essers AH, Verstegen GM, Hofhuizen DM, <i>et al.</i> A blinded randomised clinical trial of manual therapy and physiotherapy for chronic back and neck complaints: physical outcome measures. <i>J Manipulative Physiol Ther</i> 1992; 15 :16–23	Koes	256
Koes BW, Bouter LM, van Mameren H, Essers AH, Verstegen GM, Hofhuizen DM, <i>et al.</i> Randomised clinical trial of manipulative therapy and physiotherapy for persistent back and neck complaints: results of one year follow up. <i>BMJ</i> 1992; 304 :601–5	Koes	256
Koes BW, Bouter LM, van Mameren H, Essers AH, Verstegen GM, Hofhuizen DM, <i>et al.</i> The effectiveness of manual therapy, physiotherapy, and treatment by the general practitioner for nonspecific back and neck complaints. A randomized clinical trial. <i>Spine</i> 1992; 17 :28–35	Koes	256
Kominski GF, Heslin KC, Morgenstern H, Hurwitz EL, Harber PI. Economic evaluation of four treatments for low-back pain: results from a randomized controlled trial. <i>Med Care</i> 2005; 43 :428–35	Hurwitz	681
Lamb SE, Lall R, Hansen Z, Castelnovo E, Withers EJ, Nichols V, <i>et al.</i> A multicentred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain. The Back Skills Training (BeST) trial. <i>Health Technol Assess</i> 2010; 14 (41)	BeST	701

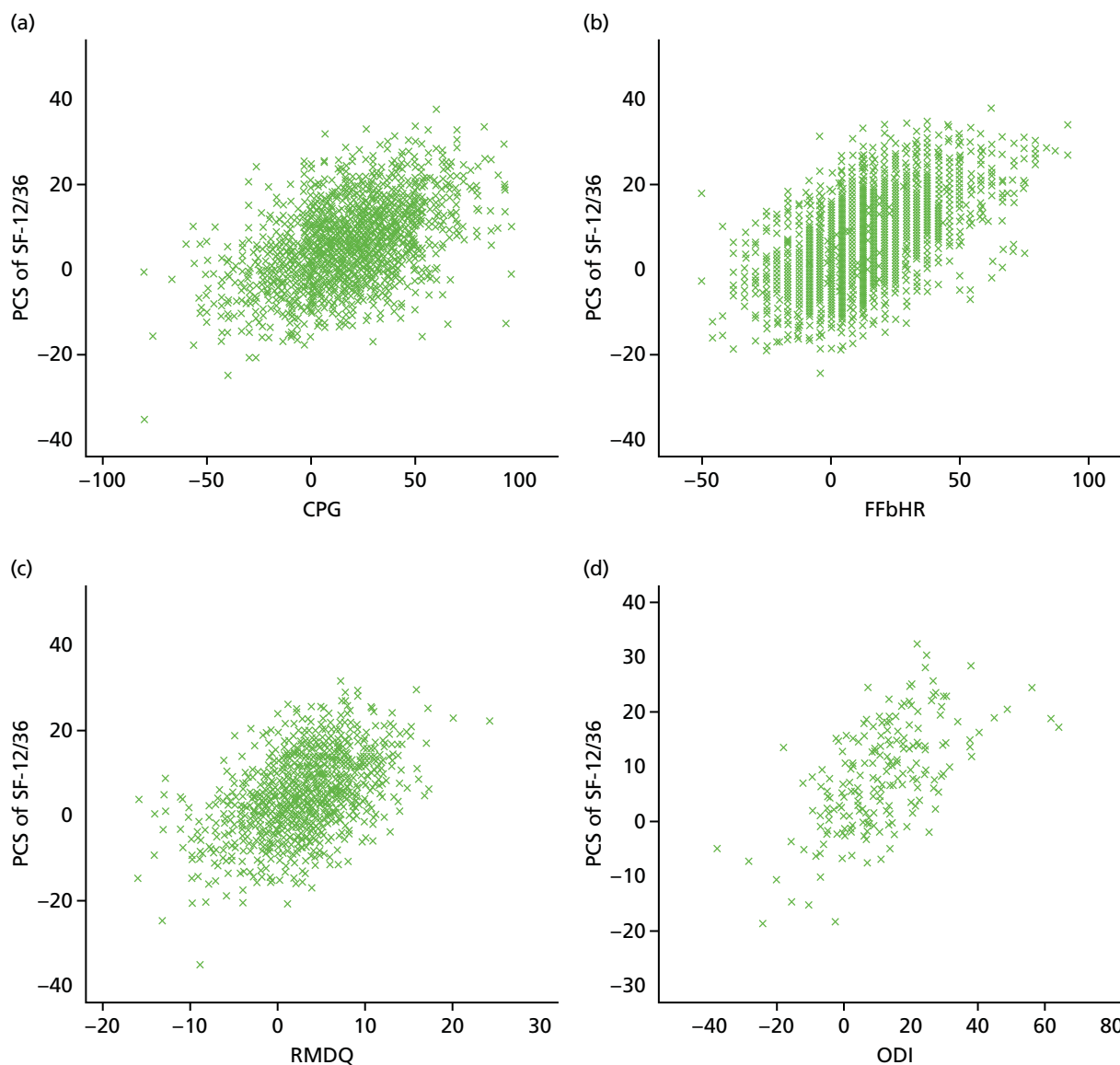
Paper	Trial	Number of participants
Mellin G, Hurri H, Harkapaa K, Jarvikoski A. A controlled study on the outcome of inpatient and outpatient treatment of low back pain. Part II. Effects on physical measurements three months after treatment. <i>Scand J Rehabil Med</i> 1989; 21 :91–5	Härkäpää	459
Myers SS, Phillips RS, Davis RB, Cherkin DC, Legedza A, Kaptchuk TJ, <i>et al.</i> Patient expectations as predictors of outcome in patients with acute low back pain. <i>J Gen Intern Med</i> 2008; 23 :148–53	Myers	444
Sherman KJ, Cherkin DC, Ichikawa L, Avins AL, Barlow WE, Khalsa PS, <i>et al.</i> Characteristics of patients with chronic back pain who benefit from acupuncture. <i>BMC Musculoskelet Disord</i> 2009; 10 :114	Sherman	638
Skargren EI, Oberg BE, Carlsson PG, Gade M. Cost and effectiveness analysis of chiropractic and physiotherapy treatment for low back and neck pain. Six-month follow-up. <i>Spine</i> 1997; 22 :2167–77	Skargren	323
Skargren EI, Oberg BE. Predictive factors for 1-year outcome of low-back and neck pain in patients treated in primary care: comparison between the treatment strategies chiropractic and physiotherapy. <i>Pain</i> 1998; 77 :201–7	Skargren	323
Smeets RJ, Maher CG, Nicholas MK, Refshauge KM, Herbert RD. Do psychological characteristics predict response to exercise and advice for subacute low back pain? <i>Arthritis Rheum</i> 2009; 61 :1202–9	Smeets	259
Steenstra IA, Knol DL, Bongers PM, Anema JR, van Mechelen W, de Vet HC. What works best for whom? An exploratory, sub-group analysis in a randomized, controlled trial on the effectiveness of a workplace intervention in low back pain patients on return to work. <i>Spine (Phila Pa 1976)</i> 2009; 34 :1243–9	Steenstra	196
Rivero-Arias O, Grey A, Frost H, Lamb SE, Stewart-Brown S. Cost-utility analysis of physiotherapy treatment compared with physiotherapy advice in low back pain. <i>Spine (Phila Pa 1976)</i> 2006; 31 :1381–7	Rivero-Arias	286
Underwood MR, Morton V, Farrin A. Do baseline characteristics predict response to treatment for low back pain? Secondary analysis of the UK BEAM dataset. <i>Rheumatology (Oxford)</i> 2007; 46 :1297–302	BEAM	1,334
Whitehurst DG, Lewis M, Yao GL, Bryan S, Raftery JP, Mullis R, <i>et al.</i> A brief pain management program compared with physical therapy for low back pain: results from an economic analysis alongside a randomized clinical trial. <i>Arthritis Rheum</i> 2007; 57 :466–73	Whitehurst	402

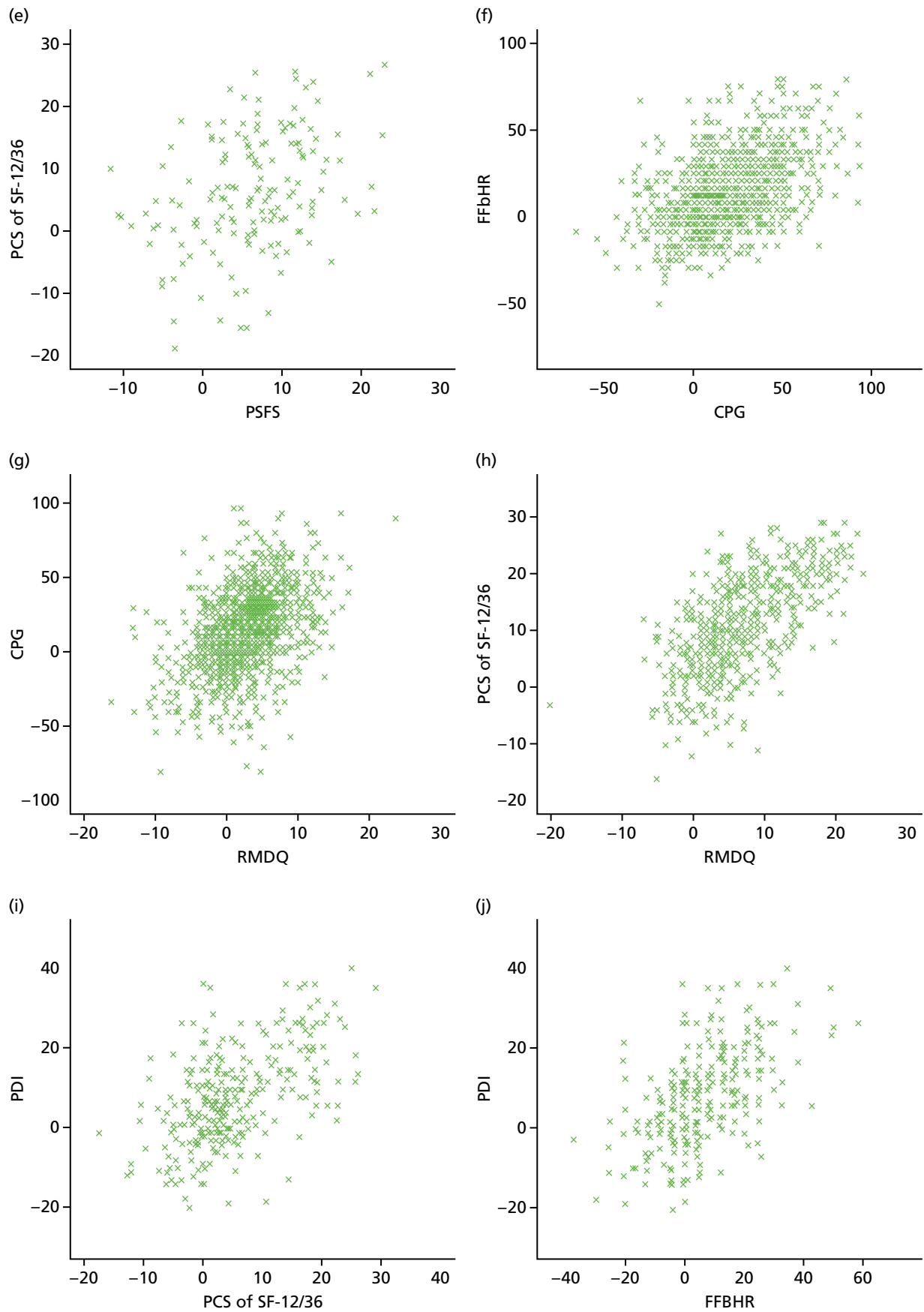
Appendix 7 Trials unavailable

Full reference	Number of participants
Alaranta H, Rytokoski U, Rissanen A, Talo S, Ronnema T, Puukka P, <i>et al.</i> Intensive physical and psychosocial training program for patients with chronic low back pain. A controlled clinical trial. <i>Spine</i> 1994; 19 :1339–49	193
Albaladejo C, Kovacs FM, Royuela A, del Pino R, Zamora J. The efficacy of a short education program and a short physiotherapy program for treating low back pain in primary care: a cluster randomized trial. <i>Spine (Phila Pa 1976)</i> 2010; 35 :483–96	348
Anema JR, Steenstra IA, Bongers PM, de Vet HC, Knol DL, Loisel P, <i>et al.</i> Multidisciplinary rehabilitation for subacute low back pain: graded activity or workplace intervention or both? A randomized controlled trial. <i>Spine (Phila Pa 1976)</i> 2007; 32 :291–8, discussion 99–300	196
Berwick DM, Budman S, Feldstein M. No clinical effect of back schools in an HMO. A randomized prospective trial. <i>Spine</i> 1989; 14 :338–44	222
Cherkin DC, Deyo RA, Battie M, Street J, Barlow W. A comparison of physical therapy, chiropractic manipulation, and provision of an educational booklet for the treatment of patients with low back pain. <i>N Engl J Med</i> 1998; 339 :1021–29	321
Cherkin DC, Eisenberg D, Sherman KJ, Barlow W, Kaptchuk TJ, Street J, <i>et al.</i> Randomized trial comparing traditional Chinese medical acupuncture, therapeutic massage, and self-care education for chronic low back pain. <i>Arch Intern Med</i> 2001; 161 :1081–8	262
Cherkin DC, Sherman KJ, Avins AL, Erro JH, Ichikawa L, Barlow WE, <i>et al.</i> A randomized trial comparing acupuncture, simulated acupuncture, and usual care for chronic low back pain. <i>Arch Intern Med</i> 2009; 169 :858–66	638
Damush TM, Weinberger M, Perkins SM, Rao JK, Tierney WM, Qi R, <i>et al.</i> The long-term effects of a self-management program for inner-city primary care patients with acute low back pain. <i>Arch Intern Med</i> 2003; 163 :2632–8	211
Eisenberg DM, Post DE, Davis RB, Connelly MT, Legedza AT, Hrbek AL, <i>et al.</i> Addition of choice of complementary therapies to usual care for acute low back pain: a randomized controlled trial. <i>Spine (Phila Pa 1976)</i> 2007; 32 :151–8	444
Frost H, Lamb SE, Doll HA, Carver PT, Stewart-Brown S. Randomised controlled trial of physiotherapy compared with advice for low back pain. <i>BMJ</i> 2004; 329 :708	286
Goldby LJ, Moore AP, Doust J, Trew ME. A randomized controlled trial investigating the efficiency of musculoskeletal physiotherapy on chronic low back disorder. <i>Spine (Phila Pa 1976)</i> 2006; 31 :1083–93	346
Goldstein MS, Morgenstern H, Hurwitz EL, Yu F. The impact of treatment confidence on pain and related disability among patients with low-back pain: results from the University of California, Los Angeles, low-back pain study. <i>Spine J</i> 2002; 2 :391–9; discussion 99–401	681
Hagen EM, Eriksen HR, Ursin H. Does early intervention with a light mobilization program reduce long-term sick leave for low back pain? <i>Spine (Phila Pa 1976)</i> 2000; 25 :1973–6	457
Hagen EM, Odelien KH, Lie SA, Eriksen HR. Adding a physical exercise programme to brief intervention for low back pain patients did not increase return to work. <i>Scand J Public Health</i> 2010; 38 :731–8	246
Heymans MW, de Vet HC, Bongers PM, Knol DL, Koes BW, van Mechelen W. The effectiveness of high-intensity versus low-intensity back schools in an occupational setting: a pragmatic randomized controlled trial. <i>Spine (Phila Pa 1976)</i> 2006; 31 :1075–82	299
Hondras MA, Long CR, Cao Y, Rowell RM, Meeker WC. A randomized controlled trial comparing 2 types of spinal manipulation and minimal conservative medical care for adults 55 years and older with subacute or chronic low back pain. <i>J Manipulative Physiol Ther</i> 2009; 32 :330–43	240
Hurley DA, McDonough SM, Dempster M, Moore AP, Baxter GD. A randomized clinical trial of manipulative therapy and interferential therapy for acute low back pain. <i>Spine</i> 2004; 29 :2207–16	240

Full reference	Number of participants
Johnson RE, Jones GT, Wiles NJ, Chaddock C, Potter RG, Roberts C, <i>et al.</i> Active exercise, education, and cognitive behavioral therapy for persistent disabling low back pain: a randomized controlled trial. <i>Spine (Phila Pa 1976)</i> 2007; 32 :1578–85	196
Koes BW, Bouter LM, van Mameren H, Essers AH, Verstegen GJ, Hofhuizen DM, <i>et al.</i> A randomized clinical trial of manual therapy and physiotherapy for persistent back and neck complaints: sub-group analysis and relationship between outcome measures. <i>J Manipulative Physiol Ther</i> 1993; 16 :211–19	256
Linton SJ, Andersson T. Can chronic disability be prevented? A randomized trial of a cognitive-behavior intervention and two forms of information for patients with spinal pain. <i>Spine</i> 2000; 25 :2825–31, discussion 24	243
Mellin G, Harkapaa K, Hurri H, Jarvikoski A. A controlled study on the outcome of inpatient and outpatient treatment of low back pain. Part IV. Long-term effects on physical measurements. <i>Scand J Rehabil Med</i> 1990; 22 :189–94	459
Niemisto L, Lahtinen-Suopanki T, Rissanen P, Lindgren KA, Sarna S, Hurri H. A randomized trial of combined manipulation, stabilizing exercises, and physician consultation compared to physician consultation alone for chronic low back pain. <i>Spine</i> 2003; 28 :2185–91	204
Petersen T, Larsen K, Jacobsen S. One-year follow-up comparison of the effectiveness of McKenzie treatment and strengthening training for patients with chronic low back pain: outcome and prognostic factors. <i>Spine (Phila Pa 1976)</i> 2007; 32 :2948–56	260
Poole H, Glenn S, Murphy P. A randomised controlled study of reflexology for the management of chronic low back pain. <i>Eur J Pain</i> 2007; 11 :878–87	243
Sherman KJ, Cherkin DC, Ichikawa L, Avins AL, Delaney K, Barlow WE, <i>et al.</i> Treatment expectations and preferences as predictors of outcome of acupuncture for chronic back pain. <i>Spine (Phila Pa 1976)</i> 2010; 35 :1471–7	447
Shirado O, Doi T, Akai M, Hoshino Y, Fujino K, Hayashi K, <i>et al.</i> Multicenter randomized controlled trial to evaluate the effect of home-based exercise on patients with chronic low back pain: the Japan low back pain exercise therapy study. <i>Spine (Phila Pa 1976)</i> 2010; 35 :E811–19	201
Skargren EI, Carlsson PG, Oberg BE. One-year follow-up comparison of the cost and effectiveness of chiropractic and physiotherapy as primary management for back pain. Sub-group analysis, recurrence, and additional health care utilization. <i>Spine</i> 1998; 23 :1875–83, discussion 84	323
Triano JJ, McGregor M, Hondras MA, Brennan PC. Manipulative therapy versus education programs in chronic low back pain. <i>Spine</i> 1995; 20 :948–55	209

Appendix 8 Scatterplots of raw change scores of outcome measures





Appendix 9 Statistical analysis plan

IMPROVING OUTCOMES FROM THE TREATMENT OF BACK PAIN

STATISTICAL ANALYSIS PLAN

Version	1.0
Effective date	9 December 2013
Prepared by	Siew Wan Hee Jake Jordan
Approved by	Team of Low Back Pain Repository Members of Repository Oversight Committee

Contents

	List of Abbreviations	3
1	Background	5
2	Aims of the Analysis	6
3	Quality Control	7
4	Outcome Variables	8
5	Moderator Variables	18
6	Treatment Arms	22
7	Follow-up Time Points	23
8	Datasets	23
9	Statistical Analysis	25
10	Reporting of the Results	29
	References	30
A	Appendix A: Project Specific Guide: Transfer, Query, Map, Report and Upload Data to the Repository	34

List of Abbreviations

ALBPSQ	Acute Low Back Pain Screening Questionnaire
ANCOVA	Analysis of covariance
AUC	Area under the curve
BBQ	Back Beliefs Questionnaire
BDI	Beck Depression Inventory
BMI	Body mass index
CES-D	Center for Epidemiologic Studies Depression
CPG	Chronic Pain Grade Scale
CSQ	Coping Strategy Questionnaire
DASS	Depression Anxiety and Stress Scale
DRAM	Distress and Risk Assessment Method
FABQ	Fear-Avoidance Beliefs Questionnaire
FFbHR	Hannover Functional Ability Questionnaire for Measuring Back Pain-Related Functional Limitations (Funktionsbeeinträchtigung durch Rückenschmerzen)
GP	General practitioner
HADS	Hospital Anxiety and Depression Scale
INMB	Incremental net monetary benefit
IPD	Individual patient data
LBP	Low back pain
MAR	Missing at random
MCS	Mental Component Scale
MI	Multiple imputation
MNAR	Missing not at random
MSPQ	Modified Somatic Perception Questionnaire
MZDI	Modified Zung Depression Index

Effective: 9 December 2013

Version 1.0

NICE	National Institute for Health and Clinical Excellence
NMB	Net monetary benefit
ODI	Oswestry low back pain Disability Questionnaire
PCS	Physical Component Scale
PDI	Pain Disability Index
PI	Principal investigator
PRSS	Pain-Related Self Statement
PSEQ	Pain Self-Efficacy Questionnaire
PSFS	Patient Specific Functional Scale
QALY	Quality-Adjusted Life Year
QoL	Quality of Life
RCT	Randomized controlled trials
RMDQ	Roland-Morris Disability Questionnaire
SES	Pain Experience Scale (Schmerzempfindungsskala)
TENS	Transcutaneous electrical nerve stimulation
TSK	Tampa Scale for Kinesiophobia
VAS	Visual analogue scale

1. Background

1.1 Summary

The aim of the Low Back Pain Repository is to develop a repository of individual patient data (IPD) from randomized controlled trials (RCT) testing therapist-delivered interventions for low back pain (LBP). Principal investigators (PI) whose trials satisfy the inclusion criteria (Table 1.1) are approached to share their anonymized data with us. Datasets from them are then queried and validated before they are uploaded to the standardized repository database.

The primary objective of this study is to determine which patient characteristics at baseline predict clinical response to different treatments and the most cost-effective treatments for low back pain.

1.2 Design of the programme

Development of the data repository

The flow diagram of the development of the data repository is shown in Figure 1.1.

Identification of treatment moderators

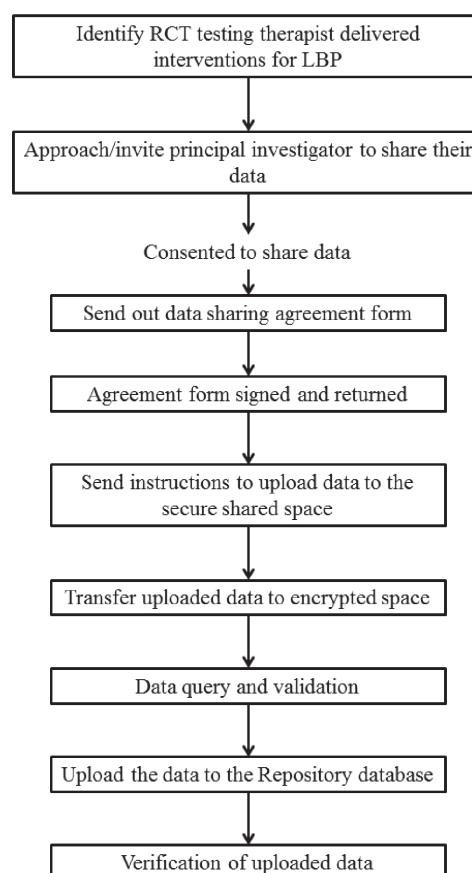
A systematic review was performed to search for RCT of therapist delivered interventions for LBP that identified patient characteristics at baseline that might predict the response to treatments. Variables that were identified from this review are entered into the pool of potential moderators to inform the final analysis.

1.3 Timing of analysis and reporting

The timeline for the data collection, analysis and reporting is shown in Table 1.2. All the investigators who have consented to share their data uploaded their data to the secure shared space before 28 February 2013.

Table 1.1 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Randomized controlled trials for non-specific low back pain	Non-randomized controlled trials (for example, observational, cohort, retrospective study)
Therapist delivered interventions trials (including psychological interventions and intensive rehabilitation programmes)	Pharmacotherapy trials
Participants aged ≥ 18	



Abbreviations: RCT, randomized controlled trials; LBP, low back pain.

Figure 1.1 Flow diagram of the development of the data repository

2. Aims of the analysis

The primary aim of the analysis is to identify a combination of patient characteristics at baseline to recommend a particular therapist delivered intervention to a subpopulation where it would be optimal to and are associated with the endpoints of interest, namely, disability (Section 4.1), pain (Section 4.2), psychological distress (Section 4.3), non-utility quality of life (Section 4.4), health utility (Section 4.5) and cost-effectiveness (Section 4.6).

Table 1.2 Timing of analysis and reporting

	2013												2014			
	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
1. Freeze collection of data																
2. Query, validate and upload all data obtained to the Repository database																
3. Map the network diagram																
4. Develop statistical models for clinical analysis																
5. Develop the models for economic analysis																
6. Analyse the data with models developed in (4) and (5)																
7. Refine the predictor model																
8. Test and validate the refined predictor model																
9. Result report																
10. Final report																
11. Dissemination and publication																

3. Quality control

3.1 Data query

Data query is performed on all data uploaded to the secure shared space. Any inconsistency, for example, out-of-range values, inconsistent dates, is resolved before being uploaded to the standardized repository database.

3.2 Extract, transform and load

A technical guideline (Appendix A) gives a detailed procedure to transfer, query, map, report and load the shared trial data to the repository database.

3.3 Verification of uploaded data to the repository database

Once the original data have been uploaded to the repository database, the data are verified manually to ensure that the process of uploading did not compromise the data integrity.

4. Outcome variables

This section describes the derivations of the scoring and scales for the measurements of the outcomes of interest. Clinical outcomes are classified broadly into physical disability (Section 4.1), pain (Section 4.2), psychological distress (Section 4.3) and non-utility quality of life (Section 4.4). The health utility and cost-effectiveness outcomes are presented in Sections 4.5 and 4.6.

As there is no single instrument that was used by all trials, the methodology in either selecting an instrument or scaling each instruments to one standard measurement will be discussed within each subsection; section 4.1.2 for physical disability, section 4.2.2 for pain and section 4.3.2 for psychological distress.

4.1 Physical disability

According to the definition from the World Report on Disability by World Health Organization (2011), disability refers to difficulties arising from any or all three of these conditions; impairments, activity limitations and participation restrictions. It is not merely a health problem but arises from the interaction between the health condition(s) and environmental and personal factors.

4.1.1 Instruments

Benefits of treatments

Some RCTs might have a single standalone instrument that asked the participant to rate the benefit of the treatment they have received. It is usually presented as a numerical rating scale with “substantial benefit” on one end, “substantial harm” on the other end, and a “no benefit” in between.

Chronic Pain Grade Scale

The Chronic Pain Grade Scale (CPG) is an instrument to grade chronic pain status (Von Korff *et al.*, 1992). It has two dimensions, namely, disability and pain intensity scores. It used with different durations recall, and may refer to all pain or specifically to low back pain. The disability score is made up of three items:

- In the past XX months/weeks, how much has (back) pain interfered with your daily activities rated on a 0-10 scale where 0 is 'no interference' and 10 is 'unable to carry on any activities'?
- In the past XX months/weeks, how much has (back) pain changed your ability to take part in recreational, social and family activities where 0 is 'no change' and 10 is 'extreme change'?
- In the past XX months/weeks, how much has (back) pain changed your ability to work (including housework) where 0 is 'no change' and 10 is 'extreme change'?

The disability score is derived as followed,

$$\text{Disability score} = \text{mean}(\text{of the three items}) \times 10.$$

The range of the score is from 0 to 100 where the higher score means more severe disability.

Hannover Functional Ability Questionnaire for Measuring Back Pain-Related Functional Limitations (Funktionsbeeinträchtigung durch Rückenschmerzen)

The Hannover Functional Ability Questionnaire for measuring back pain-related functional limitations (FFbHR) is a self-administered questionnaire developed to assess the functional limitations in daily living activities (Kohlmann and Raspe, 1996). There are 12 items and participants are instructed to tick if they could perform the activity (Yes, final score 2), could perform but with difficulty (Yes but with difficulty, final score 1) or not (No or with external help, final score 0).

$$\text{FFbHR score} = (\text{sum of all items})/24 \times 100.$$

The range of the score is from 0 (great limitation) to 100 (no limitation).

Oswestry Disability Index

The Oswestry low back pain Disability Questionnaire (ODI) is made up of 10 sections that are found to be most relevant to people suffering from low back pain (Fairbank *et al.*, 1980). It aims to assess the limitations of various activities of daily living. The activities are pain intensity, person care, lifting, walking, sitting, standing, sleeping, sex life, social life and travelling. Each section is scored between 0 and 5 (greatest disability) and the final score is

$$\text{ODI score} = \text{Total score from all sections} / \text{Total possible score} \times 100.$$

For example, if all 10 sections were completed and the total score was 16, then ODI score was $16/50 \times 100 = 32$. However, if one section was missing or not applicable and the total score was also 16 then ODI score was $16/45 \times 100 = 35.5$. The range of the score is from 0 (no disability) to 100 (greatest disability).

Pain Disability Index

The Pain Disability Index (PDI) is a measurement of the degree to which pain interferes with functioning in family/home responsibilities, recreation, social activity, occupation, sexual behaviour, self-care, and life-support activities (Tait *et al.*, 1990). Each item score ranges from 0 (no disability) to 10 (worst disability).

$$\text{PDI score} = \text{sum of all seven items}.$$

The range of the score is from 0 (no disability) to 70 (worst disability).

Patient Specific Functional Scale

The Patient Specific Functional Scale (PSFS) is an instrument that requires participants to identify up to 5 important activities that they are unable to perform or have difficulty with because of their low back pain (Stratford *et al.*, 1995). Participants are also asked to rate the level of difficulty, from 0 (unable to perform activity) to 10 (able to perform activity at preinjury level) associated with each activity. Participants are reminded of these activities at subsequent follow-ups and rate the level of difficulty.

Roland-Morris Disability Questionnaire

The Roland-Morris Disability Questionnaire (RMDQ) is a measurement for low back pain function in primary care trials (Roland and Morris, 1983). Participants are instructed to tick the statement that describes them on the day of completing the questionnaire. Item that is ticked is represented numerically by 1 and by 0, otherwise.

RMDQ score = sum of all items that are ticked.

The range of the score is from 0 (no disability) to 24 (severe disability).

SF-12/SF-36

The standard (4-week recall) and acute (1-week recall) of SF-12 (versions 1 and 2) and SF-36 (version 1 and 2) are 12- and 36-item generic measurements of quality of life, respectively (Ware *et al.*, 2002; and Ware *et al.*, 2000). The 12 items in the SF-12 measure eight scales, namely, physical functioning, role physical, bodily pain, general health, vitality, social functioning, role emotional and mental health. The 36 items in the SF-36 measure the same eight scales and an additional scale, health transition. Each of the scale is transformed and standardized to compute physical (PCS) and mental (MCS) summary measures. The steps for scoring and standardized transformation are available in the manuals. The standardized and norm-based scales are necessary for direct interpretation.

The PCS component is of interest as a measurement disability measurement. The range of the score is from 0 (substantial limitations) to 100 (no physical limitations).

Troublesomeness

This is a 6-point Likert item to ascertain the troublesomeness of LBP symptom. It is rated as “no pain experienced” (score of 1) to “extremely troublesome” (score of 6) (Parsons *et al.*, 2006).

4.1.2 Selection of instrument

All the trials had used either FFbHR, RMDQ or Von Korff as their disability outcome. An exploratory research will be performed to map FFbHR, RMDQ and Von Korff into quality-adjusted life years (QALY) or health utility outcome. The analysis is then based on the QALY/utility outcome.

In the event that it is not possible to map any of the instruments' scores to one common outcome, trials will be grouped by common outcome and analyses for these trials will be based on that common outcome.

4.2 Pain

4.2.1 Instruments

Chronic Pain Grade Scale

The Chronic Pain Grade Scale (CPG) is an instrument to grade chronic pain status (Von Korff *et al.*, 1992). It has two dimensions, namely, disability and pain intensity scores. It used with different

Effective: 9 December 2013

Version 1.0

durations recall, and may refer to all pain or specifically to low back pain. The pain intensity score is made up of three items:

- How would you rate your (back) pain on a 0-10 scale at the present time, that is, right now, where 0 is 'no pain' and 10 is 'pain as bad as could be'?
- In the past XX months/weeks, how intense/bad was your worst pain rated on a 0-10 scale where 0 is 'no pain' and 10 is 'pain as bad as could be'?
- In the past XX months/weeks, on the average, how intense/bad was your pain rated on a 0-10 scale where 0 is 'no pain' and 10 is 'pain as bad as could be'?

The pain intensity score is derived as followed,

$$\text{Pain score} = \text{mean}(\text{of the three items}) \times 10.$$

The range of the score is from 0 to 100 where the higher score means more severe pain. Underwood *et al.* (1999) modified the CPG pain intensity scale to be more specific for low back pain. However, the scoring for pain intensity remains the same.

McGill Pain Questionnaire (VAS)

The long (Melzack, 1975) and short (Melzack, 1987) forms of the McGill Pain Questionnaire aim to quantify the sensory, affective and evaluative dimensions of pain experience and are commonly used in diagnosis. The short form also has a visual analogue scale (VAS) that anchors with “no pain” at the left pole and “worst possible pain” at the right pole.

SF-12/SF-36

As described in Section 4.1.1, the SF-12/36 is made up of eight scales, namely, physical functioning, role physical, bodily pain, general health, vitality, social functioning, role emotional and mental health. One of them, bodily pain, is of interest as a measurement for pain. The range of the score is from 0 (very severe and extremely limiting pain) to 100 (no pain or limitations due to pain).

Visual Analogue Scale

Most RCTs might have a single standalone instrument that asked the participant to either rate or mark in an analogue scale that describes their average/worst pain at the present time or over the past XX months/weeks. The VAS is usually presented as a line that anchors with “no pain” at one end and “worst possible pain” at the other end. The line could be either horizontal or vertical.

4.2.2 Selection of instrument

There exist slight differences between average pain and worst pain. The recall period asked in each instrument and between trials may also differ slightly and this may have an impact in the analyses. Thus, analyses will be performed for the following pain outcomes:

- Average pain today
- Average pain over the past 1 week
- Average pain over the past 1 month

Effective: 9 December 2013

Version 1.0

- Average pain over the past 3 months
- Worst pain today
- Worst pain over the past 1 week
- Worst pain over the past 1 month
- Worst pain over the past 3 months

For all analyses, individual VAS will be the primary pain outcome. Where a numerical rating scale (range, 0 to 10) is used it will be scaled to an analogue scale that gives a range from 0 to 100.

If VAS was not available from a trial, the following instruments will be used (in descending order):

- The CPG pain intensity score is an average of the three possible questions that are usually asked in VAS. Thus, if scoring from individual items were available then the scoring of the individual item that is equivalent to the VAS item will be used and scaled to an analogue scale to give a range from 0 to 100. However, if only the CPG pain intensity score is available then the summary score will be used.
- The bodily pain domain of SF-12/36.

4.3 Psychological distress

4.3.1 Instruments

Beck Depression Inventory

The Beck Depression Inventory (BDI) is an instrument used to assess the intensity of depression in psychiatrically diagnosed patients and also to detect depression in normal population (Beck *et al.*, 1961 and 1979). It is made up of 21 items (symptoms) and the intensity is rated from 0 (neutral) to 3 (maximum severity).

BDI score = sum of all 21 items.

The range of the score is from 0 to 63 where the higher score means severe depression. The classification (for those diagnosed with affective disorder) (Beck *et al.*, 1988):

None or minimal depression	< 10
Mild to moderate depression	10 - 18
Moderate to severe depression	19 - 29
Severe depression	30 - 63

Center for Epidemiological Studies Depression Scale

The Center for Epidemiologic Studies Depression Scale (CES-D Scale) is an instrument to measure current level of depressive symptomatology in normal population (Radloff, 1977). There are 20 items in the list that the participant might have felt or behaved during the past week. There are four possible frequency of occurrence for each symptom (item), namely, less than 1 day, 1 to 2 days, 3 to 4 days and 5 to 7 days. The response is subsequently scored from 0 to 3 where a score of 0 represents less than 1 day and a score of 3 represents the highest frequency.

CES-D score = sum of all 20 items.

Effective: 9 December 2013

Version 1.0

The range of the score is from 0 to 60 where the higher score indicates more symptoms. A score of 16 or higher is an indicator of high depressive symptoms (Radloff, 1977).

Depression Anxiety Stress Scales

The Depression Anxiety and Stress Scale (DASS) is an instrument that measure depression, anxiety and stress in diverse settings (Lovibond and Lovibond, 1995). The full version of DASS consists of 42 items whereas the short-form version, DASS-21, consists of 21 items taken from the full version (Henry and Crawford, 2005). Each item asks the participant how much the statement applies to them over the past week and is scored from 0 (did not apply at all) to 3 (very much or most of the time).

$$\text{DASS-42}_{\text{depression/anxiety/stress}} = \text{sum of all the corresponding items.}$$

$$\text{DASS-21}_{\text{depression/anxiety/stress}} = \text{sum of all the corresponding items} \times 2.$$

The range for each subscale is from 0 to 42 with higher score indicates severity. The classification:

	Depression	Anxiety	Stress
Normal	0 - 9	0 - 7	0 - 14
Mild	10 - 13	8 - 9	15 - 18
Moderate	14 - 20	10 - 14	19 - 25
Severe	21 - 27	15 - 19	26 - 33
Extremely severe	≥ 28	≥ 20	≥ 34

Distress and Risk Assessment Method

The Distress and Risk Assessment Method (DRAM) is constructed from Modified Somatic Perception Questionnaire (MSPQ) and Modified Zung Depression Index (MZDI) (Main *et al.*, 1992). It identifies four types of patients, namely, normal (N), at risk (R), distressed-depressive (DD) and distressed-somatic (DS). The cut-offs for classification:

Type N	MZDI < 17
Type R	17 – 33 MZDI and MSPQ < 12
Type DD	MZDI > 33
Type DS:	17 – 33 MZDI and MSPQ ≥ 12.

EuroQol (Anxiety/Depression)

The descriptive system of EQ-5D-3L consists of five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) (EuroQol Group, 1990). Only the anxiety/depression dimension is of interest here. The dimension has three severity levels indicating no problem (level 1), moderate (level 2) and extreme (level 3) problems.

Hospital Anxiety and Depression Scale

The hospital anxiety and depression scale (HADS) is an instrument to detect anxiety and depression (Snaith, 2003). Each dimension consists of seven items and each item is rated from 0 to 3.

- Anxiety = sum(of items 1, 3, 5, 7, 9, 11, 13).
- Depression = sum(of items 2, 4, 6, 8, 10, 12, 14).

Therefore, the possible score for anxiety is from 0 to 21, and similarly, for depression, 0 to 21. The classification:

Normal	0 - 7
Possible presence of respective state	8 - 10
Presence of respective state	≥ 11

Table 4.1 Dimensions of psychological distress and the instruments used to measure them.

Dimensions	Instruments
Depression	DASS-42/21 _{depression} , DRAM, EuroQol (Anxiety/Depression), HADS _{depression} , MZDI, MCS of SF-12/36
Anxiety	DASS-42/21 _{anxiety} , EuroQol (Anxiety/Depression), HADS _{anxiety} , MCS of SF-12/36

Modified Zung Depression Index

The Modified Zung Depression Index (MZDI) is an instrument that could recognise depressive features and has been highly associated with participant's level of disability (Main *et al.*, 1992). It consists of 23 items and participant is to rate how frequent they experience each of the statement recently. The scoring for each item ranges from 0 (less than 1 day per week) to 3 (5 to 7 days per week). The scoring for items 2, 6, 7, 12, 14, 16, 18, 20, 21 and 23 is reversed

$$\text{MZDI score} = \text{sum of all items.}$$

The range of the score is from 0 to 69 where higher score indicates more depressed.

SF-12/SF-36

As described in Section 4.1.1. The MCS component is of interest as a psychological distress measurement. The range of the score is from 0 (substantial social and role disability due to emotional problems) to 100 (absence of psychological distress).

4.3.2 Selection of instrument

There are two dimensions of psychological distress that are of particular interest, namely, depression and anxiety. Table 4.1 shows the instruments that are used to measure these dimensions. Within each instrument there is usually a classification system that is widely used to classify patients into ordinal category, for example, with minimal, moderate, or severe level of anxiety/depression. Therefore, all the instruments will be mapped into a single ordinal categorical variable. The scores will be categorized by the 33.33rd and 66.67th percentile or by the instrument's cut-off that discriminate the low and high risk from the moderate risk group.

4.4 Quality of life

SF-12/SF-36

As described in Section 4.1.1. Both the PCS and MCS components are considered in the quality of life measurement. The range of the score is from 0 (substantial limitations/frequent psychological distress) to 100 (no physical limitations/absence of psychological distress).

4.5 Health utility

4.5.1 Utility measures hierarchy (EQ-5D – SF-12/36)

One of the challenges with the economic analysis is differing Quality of Life (QoL) instruments being used to estimate patient utility across the different trials. As the primary measure to estimate utility we will use the EQ-5D. If the data from the EQ-5D were not collected, the SF-12/36 will be used and a mapping process applied to convert the SF-12/36 results to EQ-5D dimension scores and utility estimates.

EuroQol

The EQ-5D-3L is a standardized measurement of health status for clinical and economic appraisal (Brooks, 1996; Dolan, 1997). It incorporates the description and valuation of health status into a single package with two components. One component is a standardized multi-dimensional descriptive system of general health. The second is a ready-to-use preference-based value set obtained from the general population. The descriptive system of EQ-5D-3L consists of five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression), and each dimension has three severity levels indicating no problem (level 1), moderate (level 2) and extreme (level 3) problems. The patient's health status can be described and defined by filling in the descriptive system. Once the health status has been identified, an attached preference-based value can be calculated from the value set, which will serve as the quality adjustment weight for calculating quality-adjusted life years (QALYs). The UK Social Tariff value set will be used to calculate the quality adjustments (utility).

SF-12/SF-36

As described in Section 4.1.1. Both the PCS and MCS components are considered in the quality of life measurement. The range of the score is from 0 (substantial limitations/frequent psychological distress) to 100 (no physical limitations/absence of psychological distress).

4.5.2 Mapping SF-12/36 to EQ-5D

Mapping is an approach to derive an estimate of health state utility for one survey from scores elicited using another survey. The EQ-5D will be the primary instrument used to estimate utility. For trials with no EQ-5D data, the SF-12/36 will be used and a mapping process applied to convert the SF-12/36 results to EQ-5D dimension scores and utility estimates.

It is possible to use an algorithm (Sheffield) to convert the SF-12/36 into an SF-6D and assign utility values, however studies (Brazier and Roberts, 2004) have demonstrated these may not be directly comparable with those from the EQ-5D tariff.

Effective: 9 December 2013

Version 1.0

There are several methods available to map the SF-12/36 to the EQ-5D. Firstly, a choice must be made to map the SF-12/36 to the EQ-5D index score, or to map to the EQ-5D individual dimensions. The advantage of mapping to the dimension score is that the data used to define the mapping algorithm is not country specific, whereas the index score is based on the country specific tariffs and limits the generalizability of the algorithm. This will not be an issue, as we are only considering utility from a UK valuation perspective. The disadvantage of mapping to the individual dimensions is added complexity without necessarily increased predictive power (Rowen *et al.*, 2009).

Once we have decided whether to map to the index value or the dimension score, we have our dependant variable. Second there is a choice as to how we estimate the relationship between the SF-12/36 (our explanatory variable) and the EQ-5D (dependant variable). The first choice is to use existing estimates generated from existing algorithms based on large national datasets. The alternative is to generate our own estimates of the relationship using the trials with SF-12/36 data and EQ-5D data. We would generate these estimates using an existing, validated econometric approach. Literature has shown (Rowen *et al.*, 2009) that heterogeneity across populations can lead to different mapping estimates being generated. This suggests applying existing estimates to our trial data may not be appropriate if the characteristics of our trial data differ from the original study. However, the differences in estimates may be small and outweighed by the added simplicity of the approach.

In addition, for the benefits of generating new mapping estimates to be realised, those studies used to generate the new estimates (studies with both SF-12/36 & EQ-5D data) must be of a large sample which is homogenous with the studies the mapping is applied to (studies with only SF-12/36 data). If new estimates are generated to support the mapping process, there is the added complexity of suitable validation of the estimates and approach. This is required as advised by the NHS DSU TSD guidelines (Longworth and Rowen, 2013). With an existing algorithm and estimates, this validation should have already occurred.

With each of the mapping approaches discussed there exists the risk of bias being introduced into the results. Rowen *et al.* (2009) found each of these methods would overestimate the Health State Utility for patients with worse health states. For this reason, which ever approach is used, validation against those trials with both SF-12/36 and EQ-5D data is paramount to minimize this risk of bias.

In the first instance a simple approach will be applied using existing estimates and mapping algorithm to estimate the EQ-5D utility index for the trials with only SF-12/36 data. For validation purposes this will also be applied to trials with both SF-12/36 & EQ-5D. The accuracy of the estimates can then be compared directly. More complex mapping methods, as described, will be explored as necessary.

4.5.3 Derivation of QALYs

Quality adjusted life years (QALYs) are a standardized measure of a patient's health status. The EQ-5D is a method of estimating a patient's utility level at a given point in time. In order to turn this into

a QALY it must be integrated over time. For example, an EQ-5D utility score of 1, held by a patient for a 6 month period would equate to a QALY of 0.5. In this way QALYs can be calculated as the area under the curve (AUC), where time is on the horizontal axis and utility is measured on the vertical axis. Where EQ-5D data is not directly available, the mapped EQ-5D scores will be used and an AUC will be generated from the mapped utility scores. The AUC will be calculated for each patient, providing a QALY score as measured over a 1 year time horizon.

Under perfect conditions an exact continuous curve could be estimated for each patient, giving an unbiased estimate of their QALY score over 1 year. In practice this is not feasible. As an alternative, a discrete approximation method is used, called discrete or numerical integration. The AUC is divided up into a series of trapezoids from which the area is then calculated. For a curve concave to the origin this has the effect of slightly underestimating the true area, for a convex function the area will be slightly overestimated.

The more data points (in our case EQ-5D follow up points) the better the accuracy of the numerical estimation method. This does lead to a further issue. The trials within this study have different numbers of follow up points. This suggests that for those with more follow up points a more accurate (less biased) estimate of their QALYs will be achieved. In practice this is unlikely to cause a material difference.

4.6 Cost-effectiveness

4.6.1 Cost

Cost of treatment is made up of the cost of the intervention and the cost of healthcare resource use following the intervention. Unit costs will be identified for all healthcare resource use items from English national sources (NHS reference costs, PSSRU). The trials included in this study have varying levels of detail on healthcare resource usage. For trials with recorded resource use data, total costs per patient will be generated by multiplying the amount of resource use by its associated unit cost and adding the cost of the intervention itself. Costs will be calculated over a 1 year time horizon. Costs will be presented as a total cost per patient from an NHS perspective.

Primary analysis will include trials with both health outcomes and resource use data from which a cost of treatment can be estimated. Trials with extensive missing resource use data may also need to be excluded if the missing data cannot be imputed in a robust and stable way (see Section 8.3).

For trials lacking resource use data, costs cannot be calculated directly. Where this is the case, costs will be estimated indirectly as a function of the health outcomes. Using data from trials with both resource use and health outcome a regression model will be estimated. The specification of the model will be dictated by the data. A mixed effects model controlling for clustering by trial and intervention

with costs as the dependant variable will be assumed. Health outcomes will be the main independent variable, with demographics and baseline data included as covariates to control for heterogeneity across trial. The purposes of the model will be to estimate the relationship between the health outcomes, other covariates (primarily demographic data) and the total cost of treatment. If the model does not have suitable predictive power it will not be appropriate to include those trials without resource use in the full economic analysis.

4.6.2 Net monetary benefit

Using the methods described above, QALYs/effects (E) and costs (C) will be estimated for each patient over a 1 year time horizon. The cost effectiveness analysis will be formed of three parallel streams. Firstly, to maximize QALYs (irrespective of costs), secondly to minimize costs (irrespective of QALYs) and finally to maximize expected net monetary benefit (NMB). The expected NMB is calculated as a function of the QALYs, costs and the societal willingness to pay per QALY gained (λ) as shown above. In this way, the expected NMB accounts for both costs and QALYs simultaneously. The NMB will be calculated using a threshold willingness to pay of £30k per QALY gained, as per National Institute for Health and Clinical Excellence (NICE) guidelines.

5. Moderator variables

This section defines the explanatory variables that may potentially be treatment moderators. The moderators are made up of participant characteristics/demographics (Section 5.1), employment and work status (Section 5.2), and baseline clinical data (Sections 4.1, 4.2, 4.3, 4.4 and 5.3)

5.1 Participant characteristics and demographic data

Variables collected at baseline:

- Age
- Sex
- Ethnicity
- Education
- BMI
- Previous treatment(s)

5.2 Employment and work status

The employment and work status are collected at baseline.

5.3 Baseline clinical data

This section describes the derivations of the scoring and scales of the instruments used to measure clinical outcomes at baseline. The outcomes are classified broadly into disability (Section 4.1), pain (Section 4.2), psychological distress (Section 4.3), quality of life (Section 4.4), fear avoidance and

beliefs (Section 5.3.1), catastrophizing (Section 5.3.2), coping (Section 5.3.3), sensory and affective perception (Section 5.3.4) and benefits of treatment (Section 5.3.5).

5.3.1 Fear avoidance and beliefs

Acute Low Back Pain Screening Questionnaire

The Acute Low Back Pain Screening Questionnaire (ALBPSQ) is a biopsychosocial screening instrument with 24 items (Linton and Hallden, 1998). Three items asked for year of birth (age), sex and nationality, and the other 21 are scored from 0 to 10 that contribute to the ALBPSQ score.

$$\text{ALBPSQ score} = \text{sum of all items.}$$

The total score ranges from 0 to 210. However, only the following three items are used to measure the fear-avoidance beliefs:

- Physical activity makes my pain worse.
- An increase in pain is an indication that I should stop what I am doing until the pain decreases.
- I should not do my normal work with my present pain.

The scores for these items will be summed up.

Back Beliefs Questionnaire

The Back Beliefs Questionnaire (BBQ) is an instrument that measures a participant's beliefs about their LBP and the inevitable future as the consequence of LBP (Symonds *et al.*, 1996). It consists of nine inevitability statements and five "distracting" statements. Participant is to rate each item with score from 1 (completely disagree) to 5 (completely agree). The BBQ scale is computed by reversing the scoring for items 1, 2, 3, 6, 8, 10, 12, 13, and 14 (the inevitability statements), and then, summing them up. The total score ranges from 9 to 45 with a higher score indicates a more positive attitudes and beliefs.

Fear-Avoidance Beliefs Questionnaire

The fear-avoidance beliefs questionnaire (FABQ) is an instrument to measure participant's beliefs about how physical activity and work affect their low back pain (Waddell *et al.*, 1993). The physical component consists of four 7-level items and the work component consists of seven 7-level items. The individual item score ranges from 0 (completely disagree) to 6 (completely agree).

$$\text{FABQ}_{\text{physical}} = \text{sum}(\text{of items 2, 3, 4, and 5}).$$

$$\text{FABQ}_{\text{work}} = \text{sum}(\text{of items 6, 7, 9, 10, 11, 12 and 15}).$$

Thus, the total score for physical component ranges from 0 to 24 and for work component ranges from 0 to 42.

Tampa Scale for Kinesiophobia

The original Tampa Scale for Kinesiophobia (TSK) developed by Miller, Kopri and Todd was unpublished but was later published with permission in Vlaeyen *et al.* (1995). It consists of 17 items and aims to measure the fear of movement or (re)injury. Each item is scored from 1 (strongly

disagree) to 4 (strongly agree). For the computation of the total score, scores for items 4, 8, 12, and 16 are reversed.

TSK score = sum of all items.

The total score ranges from 17 to 68 with higher score indicates higher degree of kinesiophobia.

5.3.2 Catastrophizing

Coping Strategies Questionnaire

The Coping Strategy Questionnaire (CSQ) is a 48-item instrument that assesses the cognitive and behavioural pain coping strategies of participants with chronic LBP (Rosenstiel and Keefe, 1983). The 48 items summarize into six different cognitive coping strategies, namely, diverting attention (DA), reinterpreting pain sensations (RS), coping self-statements (CSS), ignoring pain sensations (IS), praying and hoping (PH) and catastrophizing (CAT), and two behavioural coping strategies, namely, increasing behavioural activity (IBA) and increasing pain behaviours (IPB). However, some subscales may have lower internal reliability and other shorter versions of the CSQ are sometimes used (see, for example, Harland and Georgieff, 2003).

Regardless of the version, each item in the CSQ is scored on a 7-point Likert scale from 0 (never do that) to 6 (always do that). Items that correspond to each of the subscale are summed up. Generally, six items from the CSQ sum up each subscale. Hence, the range of score for each subscale is from 0 to 36. The higher score means a more frequently used strategy in coping chronic pain.

Only the catastrophizing (CAT) dimension of the CSQ is used

Pain-Related Self Statement

The Pain-Related Self Statement (PRSS) scale assesses participant's cognitive coping with pain (Flor *et al.*, 1993). It consists of two subscales; "catastrophizing" and "coping". Each subscale is summarized by nine items. Participant is to rate on a 6-point Likert scale of how often the statement entered their mind when they experienced severe pain. The score ranges from 0 (almost never) to 5 (almost always).

PRSS-catastrophizing = sum of even numbered items.

PRSS-coping = sum of odd numbered items.

The total score for both subscales ranges from 0 to 45 with the higher score indicates more positive self-statements.

5.3.3 Coping

Coping Strategies Questionnaire

See section 5.3.2. Only the coping subscale of the CSQ (CSS) is used.

Pain-Related Self Statement

See section 5.3.2. Only the coping subscale of the PRSS (PRSS-coping) is used.

Pain Self-Efficacy Questionnaire

The Pain Self-Efficacy Questionnaire (PSEQ) is an instrument aims to measure the confidence of the participant in performing a particular behaviour or task despite of their pain (Nicholas, 2007). There are 10 items in the questionnaire and each item is made up of seven levels, ranging from 0 (not at all confident) to 6 (completely confident).

PSEQ score = sum of all items.

The total score ranges from 0 to 60 where the higher score reflects stronger self-efficacy beliefs.

5.3.4 Sensory and affective perception

McGill Pain Questionnaire

The long (Melzack, 1975) and short (Melzack, 1987) forms of the McGill Pain Questionnaire aim to quantify the sensory, affective and evaluative dimensions of pain experience and are commonly used in diagnosis. In the short form, there are 11 items associated with sensory dimension of pain experience and four items associated with affective dimension. Participant is to rate the intensity of each pain descriptor as “none” (score, 0), “mild” (score, 1), “moderate” (score, 2) or “severe” (score, 3).

Sensory index = sum of all 11 items associated with sensory perception.

Affective index = sum of all 4 items associated with affective perception.

The range of sensory index is from 0 to 33 and the range of affective index is from 0 to 12 where higher score indicates severe intensity.

Modified Somatic Perception Questionnaire

The Modified Somatic Perception Questionnaire (MSPQ) is an instrument that measures somatic and autonomic perception for chronic back pain patients (Main, 1983). It consists of 13 symptoms (items) and participant is to rate the extent of how they have felt over the past week for each item. The scoring ranges from 0 (not at all) to 3 (extremely).

MSPQ score = sum of all items.

The range of the score is from 0 to 39 where higher score indicates more marked general somatic symptoms.

Pain Experience Scale (Schmerzempfindungsskala)

The Pain Experience Scale (SES) is an instrument with 24 items that measures sensory and affective characterization of pain (Geissner, 1995). It is usually used as a diagnostic tool and has been proven to be suitable in different psychological pain management approaches, physio-therapeutic prevention

Effective: 9 December 2013

Version 1.0

and a multimodal treatment programme of a specialized pain clinic. Participant is asked to rate the appropriateness of each item, from fully appropriate (score, 4) to not appropriate (score, 1).

Affective score = sum of 14 items associate with affective characterization of pain.

Sensory score = sum of 10 items associate with sensory characterization of pain.

The range of affective score is from 14 to 56 and the range of sensory score is from 10 to 40. The higher score indicates severe pain experienced.

Table 6.1 Grouping of treatment arms.

Parent group	Subgroup	Subtype
Intervention	Active physical	Exercise
		Graded activity
	Passive physical	Acupuncture
		Manual therapy
		Individual physiotherapy
	Psychological	Advice/education
		Psychological (cognitive behavioural)
	Sham control	Sham acupuncture
		Sham electrotherapy
		Mock transcutaneous electrical nerve stimulation (TENS)
		Sham advice/education
Control	GP/usual care	General practitioner (GP)
		Waiting list

5.3.5 Selection of instrument

All of the instruments will be mapped into a single ordinal categorical variable. The scores will be categorized by the 33.33rd and 66.67th percentile or by the instrument's cut-off that discriminate the low and high risk from the moderate risk group.

6. Treatment arms

The therapist delivered interventions are broadly classified into intervention, sham control and control. The intervention grouping may be further classified into three broad categories, namely, active physical, passive physical and psychological (Table 6.1).

7. Follow-up time points

Due to the design of individual trial's protocol, the follow-up time points are inherently different between trials. The follow-up times are classified broadly into short-term, mid-term and long-term (Table 7.1).

Table 7.1 Follow-up time points.

Follow-up	Definition
Short-term	Between baseline and anytime from 8 weeks to 3 months from randomization or start of first day of treatment.
Mid-term	Between baseline and 6 months from randomization or start of first day of treatment.
Long-term	Between baseline and 12 months from randomization or start of first day of treatment.

8. Datasets

8.1 Complete case analysis

The main analysis is to confirm proof of concept and hence will be based on complete case analysis.

8.2 Missing data

Missing data may be due to non-responders/withdrawals or missing items. Missingness due to non-responders or withdrawals will not be imputed. Missing items (at each follow-up time point) may be imputed and the method for imputation is as described in Section 8.3.

8.3 Imputed dataset

Instruments that have a standardize method to impute missing items will be followed. For example, imputation for items in SF-12 and SF-36 will be according to the algorithm detailed in the manual (Ware *et al.*, 2000, 2002).

For other instruments that do not provide any recommendation, multiple imputation (MI) will be used. The standard implementations of MI assume that data are missing at random (MAR) but it can also be implemented under the assumption of missing not at random (MNAR). Thus, MI will be used to handle missing items. Imputation will only be performed if the fraction of missing items for an instrument is less than 30 per cent (White *et al.*, 2011) for that particular follow-up time point. The method(s) and model(s) used will be according to the recommendations given by Little and Rubin (2002) and White *et al.* (2011).

Imputation will not be performed on summary/composite-level for clinical outcomes as it is impossible to infer whether the participant was a non-responder or had withdrawn from the trial. However, for some of the economic variables used to estimate health utility and costs, it may be necessary to impute on a summary/composite-level.

Missing data for economic health outcomes will fall into 3 categories:

Effective: 9 December 2013

Version 1.0

1. Individual dimensions missing for an outcome at a specific time-point.
2. Entire response for a health outcome missing from one or more time-points.
3. Entire response missing from a specific time-point forward to the end of the trial, where it is unknown if this is non-response or censoring due to drop out or death.

Category 1 is unlikely to be present, however if found will be dealt with via MI for that time-point alone and performed at the level of the individual dimension. For category 2, MI will be used to estimate the missing data-point as a summary/composite index score. A suitable regression equation will be specified for each trial and MI will be performed for each trial separately. Each of the variables to be imputed will be left-hand side dependent variables, estimated simultaneously to preserve covariance between them. Baseline index score, demographics and all other relevant covariates with complete data will be right-hand side independent variables. The model specification will be adjusted to find the best predictors and a model that leads to a stable convergent MI process. Individuals with no baseline data are unlikely to occur, however if they occur those individuals may have to be excluded from the analysis.

For individuals that fall into category 3, the process will be the same as for 2, however if a censored individual is known to have died this will be controlled for using a categorical dummy variable and they will be given a health utility value of 0 beyond the time of death. If the reason for censoring is not known for a particular trial/individual, the data will still be imputed. However, we will need to be mindful of the potential bias in the result. Due to the nature of the conditions being explored in these trials death is unlikely to have occurred over and above the national average rate, so should not be a concern for this process.

Truncated regression techniques will be used to constrain imputation results between the accepted ranges, for example, EQ-5D index scores can only lie between -0.59 and 1.0.

Costs as described in Section 4.6.1 will be calculated from the underlying resource use. The imputation of missing data will be performed as part of the same process as the missing health outcomes, with resource use items/costs being estimated simultaneously with the missing health outcomes data to preserve the underlying relationship (assuming correlation between healthcare resource use and health outcomes is present).

Specifically for costs, if some resource use has been captured for an individual at a time-point, any blanks at that time-point will be considered 0 rather than missing. Only resource items explicitly coded as missing in the original trial data, or where there is no resource use information for an entire time-point will be treated as missing. Resource use will, therefore, be imputed at a composite/summary level for each time-point. In this case total costs may be used as the dependent variable to be imputed. As with health outcomes this will be conditional on being able to specify a

suitable model that leads to a robust and stable MI solution. Censoring will be dealt with in the same manner as for health outcomes.

Sensitivity analysis will be performed to check the validity of the assumptions.

9. Statistical Analysis

9.1 Descriptive summary

The baseline information for each RCT and treatment arm will be summarized. The continuous data will be summarized as mean, standard deviation, median and interquartile range. The categorical data will be summarized as the number of participants and percentage within each category.

9.2 Meta-analysis

A one step individual patient data meta-analysis will be performed to explore the efficacy between intervention against control (sham treatment and GP/usual care). Trials will be modelled as random effect (Riley *et al.*, 2010).

9.3 ANCOVA analysis

An individual patient data or summary/composite meta-analysis will be performed to identify any covariates that predict outcomes. Continuous covariate will be analysed with analysis of covariance (ANCOVA) method with trials as the random effect. Categorical covariate will be analysed with logistic regression. Variables are statistically significant at a two-sided 0.05 level.

9.4 Clinical and health economic prediction rule and identification of subpopulations

The construct of a clinical and health economic prediction rule and the identification of a subpopulation that may benefit from different treatment modalities will be as detailed below. Only two treatment arms will be compared at each construction. For example, intervention arm against control arm, active physical arm against control arm, and others (see Table 6.1 for the grouping of treatment arms). Results from each construction will be collated and report together.

Table 9.1 Moderators identified from literature review (Gurung *et al.* 2013).

Age
Sex
Employment status
Education
Use of narcotic
Back pain status (baseline RMDQ)
Treatment expectations
Quality of life
Psychosocial status (baseline anxiety and/or depression)

Effective: 9 December 2013

Version 1.0

Stage 1: Interaction with treatment

All covariates that are potential moderators will be tested for interaction treatment effects. Linear models will be used to test the moderator-by-treatment interaction effects. In the event that the assumed linear relationships between the covariate and outcome are not appropriate then an alternative non-linear functional forms will be explored, *e.g.* through fractional polynomials (Royston and Sauerbrei, 2008). As model selection can lead to overoptimistic results, shrinkage methods will be applied to correct for such bias (Tibshirani, 1996). Covariate is declared as statistically significant at the 20% level. This will ensure that covariates that approach statistical significance will not be missed and not to overwhelm the pool of potential moderators for Stage 2.

Stage 2: Construction of clinical/health economic prediction rule

2.1 Modelling

Treatment moderators identified in Stage 1 and those that have been identified in the systematic review (see Table 9.1; Gurung *et al.*, 2013) will make up the list of covariates to be considered for the clinical/health economic prediction rules analysis.

There is no standard method that can be readily applied to this IPD subgroup identification. As such, we will explore and adapt two methods that are commonly used in identifying subgroups of poor prognosis in cohort studies. The first method, the Adaptive Risk Group Refinement (LeBlanc *et al.*, 2005) that identifies subgroups by a greedy algorithm “peeling” of fractions of the total data in a series of steps. The second method is based on recursive partitioning that, as the name suggests, recursively partition the covariate space to identify subgroups of patients who most (or least) benefit from treatment (see, for example, Dusseldorp *et al.*, 2010; Lipkovich *et al.*, 2011; and Su *et al.*, 2009).

Issues such as the splitting of a continuous variable or grouping of a categorical variable into fewer levels/groups, multiplicity adjustment and internal validation (*e.g.* cross-validation) will be handled within each method.

2.2 Minimum subgroup size

In splitting the covariate into two or more parts, it may be possible that the sample size of a subpopulation for a treatment arm (Table 6.1) may be very small. Prediction rules based on a very small sample size may produce unreliable and very poor estimates. As there is no clear threshold as to what is considered as a reasonable size, two proportions, namely, 1/10 and 1/20, of the population will be explored. The reliability of the estimates for each minimum size will be reported.

2.3 Formulation of economic prediction rule

The primary objective function for the economic prediction rule will be maximizing the expected net monetary benefit (NMB) as NMB combines both cost and effects simultaneously. We will also run parallel streams of analysis to maximise the sum of QALYs and minimise the total costs independently.

Effective: 9 December 2013

Version 1.0

The NMB will be estimated for each patient and substituted for the clinical outcome indicator in the prediction rule algorithm. Within this algorithm, a regression approach will be used to estimate the mean difference in outcome between one intervention and some comparator, in a sequence of subgroups defined by specified moderators and of varying size. By substituting the NMB as the dependent variable within the prediction rule algorithm, we can estimate the Incremental Net Monetary Benefit (INMB) for the intervention (relative to the comparator), for each of the subgroups tested. The optimum subgroup will be that which maximises the sum of INMB for all of the individuals in the subgroup.

Alternative regression specifications may be more robust to potential bias from endogeneity between costs and effects, skew in the distribution of costs (Nixon and Thompson, 2005), and ultimately lead to more efficient estimates than this simple NMB approach. This will be explored within the analysis. We will also investigate the possibility of using a two-equation model (Willan, *et al.* 2004) to estimate the two related dependent variables of cost and QALYs, and to control for factors that might confound the treatment effects and potential heterogeneity between trials.

For a specific treatment j , the expected NMB per individual can be expressed as:

$$E(NMB_j|P_j) = [\lambda \times E(E_j|P_j) - E(C_j|P_j)]$$

Two comparators, treatment A vs. B

In the simple case, one treatment of interest (B) will be compared to a control of usual care (or best current practice) (A). Let P_j denote the proportion of the total population P treated with intervention j ($j = A, B$), ranging from 0 to 1. The treatment options are considered exhaustive and mutually exclusive. Therefore, the subsets of the population given each treatment can be defined in terms of one another; $P_B = P - P_A$. There will be a minimum sample size equal to 10% of P , denoted by $P_{10\%}$.

Let us consider the peeling algorithm to maximize expected NMB across the total population P . The starting case is that the maximum number patients receive treatment B. Based on the moderators of interest, the peeling algorithm will iteratively reduce the sample receiving treatment B provided a higher expected NMB across the whole population (P) can be achieved. This process will continue until the expected NMB can no longer increase, or the minimum sample size of $P_B = P_{10\%}$ is reached.

As the algorithm reduces the size of the subgroup (P_B) for treatment B by 10%, the subgroup (P_A) for treatment A will be increased in size by 10%. The 10% will be made up of patients with the same characteristics as those removed from B, defined by the treatment modifier criteria. By weighting the $E(NMB)$ by P_j for each treatment a representative total $E(NMB)$ across the total population is estimated.

Effective: 9 December 2013

Version 1.0

The objective function being maximized can therefore be expressed as

$$E(NMB|P) = (P_A) \times E(NMB_A|P_A) + (1 - P_A) \times E(NMB_B|P_B),$$

provided P_A and P_B satisfied these conditions; $P_A \geq P_{10\%}$, $P - P_A \geq P_{10\%}$ and $P_B = (1 - P_A)$. Note that both proportions, P_A and P_B change as a function of the moderators of interest.

Three comparators A vs. B vs. C

At the next level of complexity, three comparators are introduced; A (usual care), treatment B and treatment C. The same constraints of mutual exclusivity and exhaustiveness apply, thus each patient in the population P must receive one and only of treatments A, B or C. In this case the process can be considered as a network, or series of sequential optimizations.

Firstly, the optimal allocation of patients between treatment B and treatment A is assessed exactly as before. We are left with two subgroups of size P_A and $P_B = (P - P_A)$. In the second phase we must identify if anyone in the two subgroups P_A and P_B would yield a better result if they were moved to treatment C. Here we define a new subgroup P_C where

$$P_A + P_B + P_C = P = 1.$$

We now have a series of three optimization problems.

Optimization 1

The first being identical to our two-treatment scenario but with treatment C included and explicitly constrained to a sample set of 0. Thus, the expected NMB is expressed as

$$E(NMB|P) = [(P_A) \times E(NMB_A|P_A)] + [(P_B) \times E(NMB_B|P_B)] + [(P_C) \times E(NMB_C|P_C)], \quad (1)$$

where P_A and P_B satisfied these conditions; $P_A \geq P_{10\%}$, $P_B \geq P_{10\%}$, $P_C = 0$, and $P_A + P_B + P_C = 1$.

At this point the optimal subgroup between P_A and P_B has been determined excluding treatment C. This has determined the starting subgroups for the next round of optimization.

$$\text{Starting sample set of treatment A} = P_A^1,$$

$$\text{Starting sample set of treatment B} = P_B^1.$$

Optimization 2

Now we will identify if anyone from subgroup P_B should be moved to treatment C. In this case subgroup P_A will be held constant at P_A^1 . The expected NMB is as expressed as equation (1) but P_A is fixed at P_A^1 whilst P_B and P_C satisfied these conditions; $P_{10\%} \leq P_B \leq P_B^1$ and $P_C \geq P_{10\%}$.

The output of this optimization will determine the final optimal solution for treatment B, designated as the subset P_B^* where treatment B is preferred over treatment A and C. There will also be those

allocated to treatment C where we know treatment C is preferred to A and B, these will be designated as P_C^1 .

Optimization 3

We will now conduct the same process for subgroup P_A^1 , as identified in Optimization 1. However, for treatment B subgroup P_B will be held constant at P_B^* and subgroup P_C will start at P_C^1 . The expected NMB is as expressed as equation (1) but P_B is fixed at P_B^* whilst P_A and P_C satisfied these conditions; $P_{10\%} \leq P_A \leq P_A^1$ and $P_C \geq P_C^1$.

Table 10.1 Items to be included in the statistical and health economic reports.

Section and topic	Description
Methods	
Statistical method	The statistical methods used for analyses as described in Sections 9.1 to 9.3. The statistical models used for analyses as described in Section 9.4 with references and a detailed description of changes made on the cited models so that they can be used in this project specifically. The validation methodology
Results (for each clinical and health economic outcomes described in Section 4)	
Trials (participants)	The trials involved.
Interventions	The interventions involved.
Outcomes	The specific instruments that have been selected for analysis.
Discussion	
Interpretation	Interpretation of the results.
Generalizability/overall evidence	General interpretation and recommendation to the community based on the current evidence.

The output of this final optimization will yield subgroups P_A^* and P_C^* . From Optimization 2 we know P_B^* . By construction, $P_A^* + P_B^* + P_C^* = P = 1$ always.

As can be seen, as this process expands beyond three comparators, the number of optimization problems will increase as a function of the number of treatment options. However the approach will be the same. The order in which the alternative treatments are compared should not influence the result of the peeling algorithm. However, for completeness the algorithm will be run on treatment comparisons in different orders to verify the result.

The same process will be followed for the purpose of maximizing total QALYs and for costs, simply substituting these measures for NMB.

10. Reporting of the Results

The statistical and health economics reports will consist of the features shown in Table 10.1. The reports will also be supported by figures and tables as appropriate.

References

- Beck, A. T., Rush, A. J., Shaw, B. F., and Emery, G. (1979). *Cognitive Therapy of Depression*. New York: Guilford Press.
- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, 4:561-571.
- Beck, A. T., Steer, R. A., and Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review* 8, 77-100.
- Brazier, J. E., and Roberts, J. (2004). The estimation of a preference-based measure of health from the SF-12. *Medical Care*, 42:851-859.
- Brooks, R. (1996). EuroQol: the current state of play. *Health Policy*, 37:53-72.
- Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical Care*, 35:1095-1108.
- EuroQol Group (1990). EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy*, 16:199-208.
- Fairbank, J., Couper, J., Davies, J., and O'Brien, J. (1980). The Oswestry low back pain disability questionnaire. *Physiotherapy*, 66:271-273.
- Flor, H., Behle, D. J., and Birbaumer, N. (1993). Assessment of pain-related cognitions in chronic pain patients. *Behaviour research and therapy*, 31:63-73.
- Geissner, E. (1995). The Pain Perception Scale--a differentiated and change-sensitive scale for assessing chronic and acute pain. *Die Rehabilitation*, 34:XXXV-XLIII.
- Gurung, T., Ellard, D. R., Mistry, D., Patel, S., and Underwood, M. Identifying potential moderators for response to treatment in low back pain: a systematic review. Submitted revisions to *BMC Musculoskeletal Disorders* in August 2013.
- Harland, N. J., and Georgieff, K. (2003). Development of the Coping Strategies Questionnaire 24, a Clinically Utilitarian Version of the Coping Strategies Questionnaire. *Rehabilitation Psychology*, 48:296-300.
- Henry, J. D., and Crawford, J. R. (2005). The short-form version of the depression anxiety stress scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 44:227-239.
- Justice, A. C., Covinsky, K. E., and Berlin, J. A. (1999). Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, 130:515-524.
- Kohlmann, T., and Raspe, H. (1996). Hannover functional questionnaire in ambulatory diagnosis of functional disability caused by backache. *Die Rehabilitation*, 35:I-VIII.
- Melzack, R. (1975). The McGill pain questionnaire: major properties and scoring methods. *Pain*, 1:277-299.

Effective: 9 December 2013

Version 1.0

LeBlanc, M., Moon, J., and Crowley, J. (2005). Adaptive risk group refinement. *Biometrics*, 61:370-378.

Linton, S. J., and Halldén, K. (1998). Can we screen for problematic back pain? A screening questionnaire for predicting outcome in acute and subacute back pain. *The Clinical Journal of Pain*, 14:209-215.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search - A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30:2601-2621.

Little, R. J. A., and Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, N.J.: Wiley.

Longworth, L., and Rowen, D. (2013). Mapping to obtain EQ-5D utility values for use in NICE Health Technology Assessments. *Value in Health*, 16:202-210.

Lovibond, P. F., and Lovibond, S. H. (1995). The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with the Beck depression and anxiety inventories. *Behaviour Research and Therapy*, 33:335-343.

Main, C. J. (1983). The modified somatic perception questionnaire (MSPQ). *Journal of Psychosomatic Research*, 27:503-514.

Main, C. J., Wood, P. L. R., Hollis, S., Spanswick, C. C., and Waddell, G. (1992). The distress and risk assessment method: A simple patient classification to identify distress and evaluate the risk of poor outcome. *Spine*, 17:42-52.

Melzack, R. (1975). The McGill pain questionnaire: Major properties and scoring methods. *Pain*, 1:277-299.

Melzack, R. (1987). The short-form McGill pain questionnaire. *Pain*, 30:191-197.

Nicholas, M. K. (2007). The pain self-efficacy questionnaire: Taking pain into account. *European Journal of Pain*, 11:153-163.

Nixon, R. M., and Thompson, S. G. (2005). Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Economics* 14, 1217-1229.

Parsons, S., Carnes, D., Pincus, T., Foster, N., Breen, A., Vogel, S., and Underwood, M. (2006). Measuring troublesomeness of chronic pain by location. *BMC Musculoskeletal Disorders*, 7:34.

Radloff, L. S. (1977). The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement* 1, 385-401.

Riley, R. D., Lambert, P. C., and Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 340:c221.

Roland, M., and Morris, R. (1983). 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*, 8:141-144.

Rosenstiel, A. K., and Keefe, F. J. (1983). The use of coping strategies in chronic low back pain patients: Relationship to patient characteristics and current adjustment. *Pain*, 17:33-44.

Rowen, D., Brazier, J., and Roberts, J. (2009). Mapping SF-36 onto the EQ-5D index: How reliable is the relationship? *Health and Quality of Life Outcomes*, 7:27.

Royston, P., Parmar, M. K. B., and Sylvester, R. (2004). Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Statistics in Medicine*, 23:907-926.

Royston, P., and Sauerbrei, W. (2008). *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley.

Sheffield. SF-6D preference based algorithm. Available at: <http://www.shef.ac.uk/scharr/sections/heds/mvh/sf-6d>. Accessed 30 Sep 2013.

Snaith, R. P. (2003). The hospital anxiety and depression scale. *Health and Quality of Life Outcomes*, 1:29.

Stratford, P., Gill, C., Westaway, M., and Binkley, J. (1995). Assessing disability and change on individual patients: A report of a patient specific measure. *Physiotherapy Canada*, 47:258-263.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10:141-158.

Symonds, T. L., Burton, A. K., Tillotson, K. M., and Main, C. J. (1996). Do attitudes and beliefs influence work loss due to low back trouble? *Occupational Medicine*, 46:25-32.

Tait, R. C., Chibnall, J. T., and Krause, S. (1990). The pain disability index: Psychometric properties. *Pain*, 40:171-182.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267-288.

Underwood, M. R., Barnett, A. G., and Vickers, M. R. (1999). Evaluation of two time-specific back pain outcome measures. *Spine*, 24:1104.

Vlaeyen, J. W. S., Kole-Snijders, A. M. J., Boeren, R. G. B., and van Eek, H. (1995). Fear of movement/(re)injury in chronic low back pain and its relation to behavioral performance. *Pain*, 62:363-372.

Von Korff, M., Ormel, J., Keefe, F. J., and Dworkin, S. F. (1992). Grading the severity of chronic pain. *Pain*, 50:133-149.

Waddell, G., Newton, M., Henderson, I., Somerville, D., and Main, C. J. (1993). A fear-avoidance beliefs questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. *Pain*, 52:157-168.

Ware, J. E., Jr., Kosinski, M., and Dewey, J. E. (2000). *How to score version 2 of the SF-36 health survey*. Lincoln, RI: QualityMetric Incorporated.

Ware, J. E., Jr., Kosinski, M., Turner-Bowker, D. M., and Gandek, B. (2002). *How to score version 2 of the SF-12 health survey (with a supplement documenting version 1)*. Lincoln, RI: QualityMetric Incorporated.

Westfall, P. H., and Young, S. S. (1993). *Resampling-based multiple testing: examples and methods for p-value adjustment*. Wiley.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30:377-399.

Willan, A. R., Briggs, A. H., and Hoch, J. S. (2004). Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics* 13, 461-475.

World Health Organization (2011). World Report on Disability. Available at: http://www.who.int/disabilities/world_report/2011/report.pdf. Accessed 30 Sep 2013

Appendix A

Project Specific Guide: Transfer, Query, Map, Report and Upload Data to the Repository

Project Specific Guide for the Low Back Pain Repository

Transfer, Query, Map, Report and Upload Data to the Repository

Version:	1.0
Effective date:	24 June 2013
Prepared by:	Siew Wan Hee Melina Dritsaki
Approved by:	Martin Underwood

Revision chronology	Effective date	Reason for change
Version 1.0	24 June 2013	

Effective: 9 December 2013

Version 1.0

Contents

- 1. Introduction**
- 2. Create Trial Folders**
- 3. Transferring Data from Shared Space to Encrypted Drive**
- 4. Querying and Reporting Data**
- 5. XML Mapping**
- 6. Uploading Data to Repository**
- 7. Verification of Uploaded Data**
- 8. Adding or Editing Classes and Attributes**
- 9. Data Analysis**
- A. Screenshot of the ETL Program**
- B. Screenshots of SPSS**
- C. Screenshots of STATA**

1. Introduction

- 1.1. These guides are intended as a detailed procedure to the individuals working to transfer, query, map, report and/or upload the trial data submitted to the Low Back Pain Trial Repository.

2. Create Trial Folders

- 2.1. Create a physical folder for each trial.
- 2.2. Create a folder in the encrypted drive for storage of dataset (e.g. "O:\Original", where O: drive is the encrypted drive) and one in the shared drive for storage of all other trial related electronic files in "M:\WMS\CTU\Rehabilitation Trials\Repository".
- 2.3. For ease of identification, the name of folders in the encrypted and shared drives should be the same.

3. Transferring Data from Shared Space to Encrypted Drive

- 3.1. Follow the instructions detailed in "Instructions for moving data from shared space to Repository.docx" in "M:\WMS\CTU\Rehabilitation Trials\Repository\3. DOCUMENTS TO SEND\File Transfer – Researchers".

4. Querying and Reporting Data

- 4.1. Open the encrypted drive. The original data is found in the folder "Original". In order for not editing and changing the original data accidentally during data query, create and copy a duplicate of the data and saved it in the folder "Temporary" which is located in the same drive.
- 4.2. All querying will be performed on this duplicate data set.
- 4.3. The data query can be performed with the following statistical programs:
 - a. Stata
 - b. SPSS
 - c. SAS
- 4.4. Each and every syntax use for the query should be recorded and saved in a folder named "Syntax" in the trial's folder (see Section 2.2), *e.g.* the query of data set from the trial BeST is saved in "M:\WMS\CTU\Rehabilitation Trials\Repository\Statistics and Health Economics\BeST\Syntax". The output from the query should also be saved in the same "Syntax" folder.

-
- 4.5. Any inconsistency, *e.g.* out-of-range values, inconsistent dates, *etc.* has to be recorded and dated. The actions taken to resolve these inconsistencies have to be recorded and dated, too. A query file template ("Data query.xlsx") is in "M:\WMS\CTU\Rehabilitation Trials\Repository\Statistics and Health Economics\Templates".
 - 4.6. Any email communication regarding the data set should be printed and kept in the trial's physical folder.
 - 4.7. The demographic and clinical outcomes at each time point have to be summarized. Any issues arising from the data query should be included in the appendix of that summary report. This summary will be sent to the trial custodian (template "Template - Data Quality Report.docx" in "M:\WMS\CTU\Rehabilitation Trials\Repository\Statistics and Health Economics\Templates").
 - 4.8. The summary will be sent off with a cover letter. The template of the cover letter is in the same folder and the name of the file is "Template - Letter for Data Quality Report.docx".
 - 4.9. The cover letter requires wet-ink signature from the Repository Principal Investigator (Professor Martin Underwood). A copy of the summary report and cover letter has to be saved in the individual trial's folder (physical and electronic versions).

5. XML Mapping

- 5.1. The mapping instructions are written in the XML language and the program for it is <oXygen/>.
- 5.2. The XML file should be saved in "M:\WMS\CTU\Rehabilitation Trials\Repository\Statistics and Health Economics\XML mapping" and the name of the file should be clear and informative on which trial it is for.

6. Uploading Data to Repository

- 6.1. Before the original data is uploaded to the Repository, it has to be saved as a comma separated value (CSV) file. The CSV file is to be saved in the folder "Processed" in the encrypted drive.
- 6.2. In some instances the original data set have to be manipulated before saving it in the CSV format. Some examples of the possibility and circumstances:

-
- a. A few data files were submitted to the Repository and so they need to be merged into a single file as the uploader requires one single data file for each trial.
 - b. Two or more variables have to be merged into one variable.
 - c. One variable has to be split into two or more variables.
- 6.3. The syntax used in the manipulation have to be recorded and saved as detailed in Section 4 before saving the modified file into a CSV file for uploading.
- 6.4. The syntax to merge data files:

SPSS syntax (example):

```
GET file="O:\Temporary\Trial01\Example01.sav" .
SORT CASES by ID .
DATASET NAME Base1 .
GET file=" O:\Temporary\Trial01\Example02.sav" .
SORT CASES by ID .
DATASET NAME Month3 .
GET file=" O:\Temporary\Trial01\Example03.sav" .
SORT CASES by ID .
DATASET NAME Month12 .
MATCH FILES
    / FILE = "Base1"
    / FILE = "Month3"
    / FILE = "Month12"
    / BY ID .
EXECUTE .
```

- 6.5. The syntax to merge two or more variables into one variable:

SPSS syntax (example):

See section 6.6

Stata syntax (example):

```
* There are two dates of interview: "var1" and "var2" and they are mutually exclusive
* Combine these two into one variable "interview"
GENERATE interview = .
REPLACE interview = var1
REPLACE interview = var2 if var1 == .
FORMAT interview %td
```

6.6. The syntax to split one variable into two or more variables:

SPSS syntax (example):

```
* The original date of assessment was in a string format thus,
* need to extract the dates, months and years (that is, split
* the original variable into three variables before merging them
* into one .
* Define the variables .
STRING assess_dd assess_mm assess_yy (A2) .
* Extract the first two characters and assign it as date .
COMPUTE assess_dd = CHAR.SUBSTR(string_assess,1,2) .
* Extract the 3rd and 4th characters and assign them as month .
COMPUTE assess_mm = CHAR.SUBSTR(string_assess,3,2) .
* Extract the last two characters and assign them as year .
COMPUTE assess_yy = CHAR.SUBSTR(string_assess,5,2) .
EXECUTE .
STRING assess_dttemp (A8) .
COMPUTE assess_dttemp = CONCAT(rtrim(assess_dd),"-",
                                rtrim(assess_mm),"-",
                                rtrim(assess_yy)) .
EXECUTE .
COMPUTE assess_date = number(assess_dttemp, date) .
FORMATS assess_date (date11) .
```

6.7. Note that there may be some string variables in the original data set and they may contain commas. In order for the Repository uploader not to confuse that the comma in

a string variable is not meant to separate the data, these commas have to be replaced with semi-colons before saving it as a CSV file.

6.8. The syntax for replacing commas:

SPSS syntax (example):

```
DO REPEAT var = var1 var2 var3 .
      IF (char.index(var, ",") GE 1)  var = REPLACE(var, ",", ";") .
END REPEAT .
EXECUTE .
```

where var1 var2 and var3 are the short names of the string variables.

Stata syntax (example):

```
FOREACH CHVAR OF var1 var2 var3 {
      REPLACE `CHVAR' = SUBINSTR(`CHVAR', " , ", ";", .)
}
```

where the notation (`) before CHVAR is the grave accent and not a single quotation (').

6.9. There may be in some occasions where the carriage return, vertical tab, new line or new page/form has been accidentally entered in these string variables. As such, these extra spaces have to be replaced as well. The syntax:

Stata syntax (example):

```
* "new line" (ASCII dec 10)
FOREACH CHVAR OF var1 var2 var3 {
      REPLACE `CHVAR' = SUBINSTR(`CHVAR', "`=char(10)'" , ";", .)
}

* "vertical tab" (ASCII dec 11)
FOREACH CHVAR OF var1 var2 var3 {
      REPLACE `CHVAR' = SUBINSTR(`CHVAR', "`=char(11)'" , ";", .)
}

* "form feed/new page" (ASCII dec 13)
FOREACH CHVAR OF var1 var2 var3 {
      REPLACE `CHVAR' = SUBINSTR(`CHVAR', "`=char(12)'" , ";", .)
}

* "carriage return" (ASCII dec 13)
FOREACH CHVAR OF var1 var2 var3 {
```

```
REPLACE `CHVAR' = SUBINSTR(`CHVAR', " "=char(13)'", ";", .)
}
```

- 6.10. The Repository uploader requires that the patient's identification number to be named as "ID" (non-case sensitive) so the variable has to be renamed if it is not already defined as "ID". The syntax for renaming and saving the original file as a CSV file:

SPSS syntax (example):

```
SAVE TRANSLATE outfile = 'O:\Processed\LisetPengel\FullDat.csv'
/ TYPE = CSV
/ FIELDNAMES
/ MISSING = RECODE
/ CELLS = values
/ RENAME = (Envelope_number=ID) .
```

Stata syntax (example):

```
RENAME PTID ID
OUTSHEET USING "O:\Processed\BeST\BeST.csv", COMMA NOLABEL QUOTE
REPLACE
```

- 6.11. Finally, to upload the trial data to the Repository:
- Open the "LBP Repository ETL" program.
 - Select the CSV file and the corresponding XML file.
 - Click "Connect".
 - Select server "Palmer", and enter the username and password assigned by the programming team (Mr Ade Willis).
 - Under the field "LBP trial selection", select the name of the trial.
 - Choose either a specific "Class" of data to be uploaded or check "Select all Classes".
 - Click "Start".

A screenshot of the ETL program is in Appendix A.

7. Verification of Uploaded Data

-
- 7.1. Once the original data have been uploaded, it is crucial to verify that the data transformation and mapping (see Section 5) are done as requested and the process of uploading does not compromise the data integrity.
 - 7.2. To set up the ODBC connection for the first time, follow the instructions provided by the programming team.
 - 7.3. To access the uploaded data with SAS, an example of the macro syntax is in a file named "MacroConnectOLEDB.sas" which is in "M:\WMS\CTU\Rehabilitation Trials\Repository\Statistics and Health Economics\Data query".
 - 7.4. To access the Repository data with SPSS:
 - a. Open the SPSS program.
 - b. Click "File", "Open Database" and select "New Query..."
 - c. Select "lbpRepository" or "lbpRepository2" from the ODBC Data Sources panel.
 - d. Click "Next".
 - e. Enter the "Login ID" and "Password" assigned by the programming team (Mr Ade Willis).
 - f. Click "OK".
 - g. De-select "Tables" and select "Views".
 - h. Double-click the class that you wish to view, for example, to view TREATMENTS double-click "stats.TREATMENTS" and then "Next".
 - i. To restrict the data that is retrieved, select the variable to be restricted in the "Expression 1" box, select the relation in the "Relation" box, and enter the value to be restricted to in the "Expression 2" box. Then click "Finish".
-

Example 1:

To select only subjects from the Kennedy trial, the values to be entered in “Expression 1”, “Relation” and “Expression 2” are:

EXPRESSION 1	RELATION	EXPRESSION 2
prms_TrialName	=	'Kennedy'

Note that the string value (e.g. Kennedy) is enclosed in single quote.

Example 2:

To select only subjects over 50 years old, the values to be entered in “Expression 1”, “Relation” and “Expression 2” are:

EXPRESSION 1	RELATION	EXPRESSION 2
Age	>	50

- Step-by-step screenshots are shown in Appendix B.

7.5. To access the Repository data with STATA:

- Open the STATA program
- Increase memory size by typing in “set memory 1000m” in the command box
- Click “Enter”
- To get the data from the ODBC Data sources panel type “odbc lo, exec(“SELECT * FROM stats.HEALL;”) dsn(“lbpRepository2” or “lbpRepository”) p(password) u(username) low clear” in the command box
- Click “Enter”

Step-by-step screenshots are shown in Appendix C

7.6. Data from a few participants for each Class and time points (baseline and any follow-up) should be chosen for the data verification.

-
- 7.7. Syntax used to verify data should be saved in the individual trial's folder called "Mapping" and saved as "Verification Syntax".
 - 7.8. Any inconsistency should be dealt with immediately to ensure data are mapped correctly.
 - 7.9. Once all the checks have been done and the mappings are correct, the data can be transferred from the server "Palmer" to the "live" server, that is, "Bauer". Email the programming team (Mr Ade Willis) to transfer the data from "Palmer" to "Bauer".

8. Adding or Editing Classes and Attributes

- 8.1. It is possible to add new classes, and both ETL program and the XML schema rules have to be updated with the new classes.
- 8.2. The XML schema rules file is "ImportRules.xsd" and this is in "M:\WMS\CTU\Rehabilitation Trials\Repository\Statistics and Health Economics". The new class(es) is(are) inserted under the heading `<xs:restriction base="xs:string">` which is under `<xs:simpleType name="typeClass">`
- 8.3. In order to update the ETL program, open the "LBP Repository ETL" program, select a dummy CSV file and a dummy XML file (available in the folder "M:\WMS\CTU\Rehabilitation Trials\Repository\Statistics and Health Economics\Examples and Dummy"). Follow steps (c) and (d) in Section 6.8 then select "Class Manager".
- 8.4. To add a new class, point to "Classes", right click, select "Add Class" and proceed.
- 8.5. To delete an existing class, point to the class, right click and select "Delete Class".
- 8.6. To add a new attribute (variable) into an existing class, point to that class, right click, select "Add Attribute" and proceed.
- 8.7. To edit an existing attribute, select that attribute and proceed.
- 8.8. To delete an existing attribute, point to the attribute, right click and select "Delete Attribute".
- 8.9. After all changes have been made, click "Refresh Stat Views". Email the programming team (Mr Ade Willis) of all the changes that have been made so that they can subsequently update the "Bauer" database.

9. Data Analysis

-
- 9.1. As the process of acquiring dataset is a fluid and continuing process, any statistical and health economic analyses to be done will be on data that have been acquired up to a cut-off time. Therefore, the statistician needs to inform the programming team (by email) to replicate the “live” database which is then saved in a server called “Buchanan”.
- 9.2. Analyses are then based on the replicated dataset.

A. Screenshot of the ETL Program

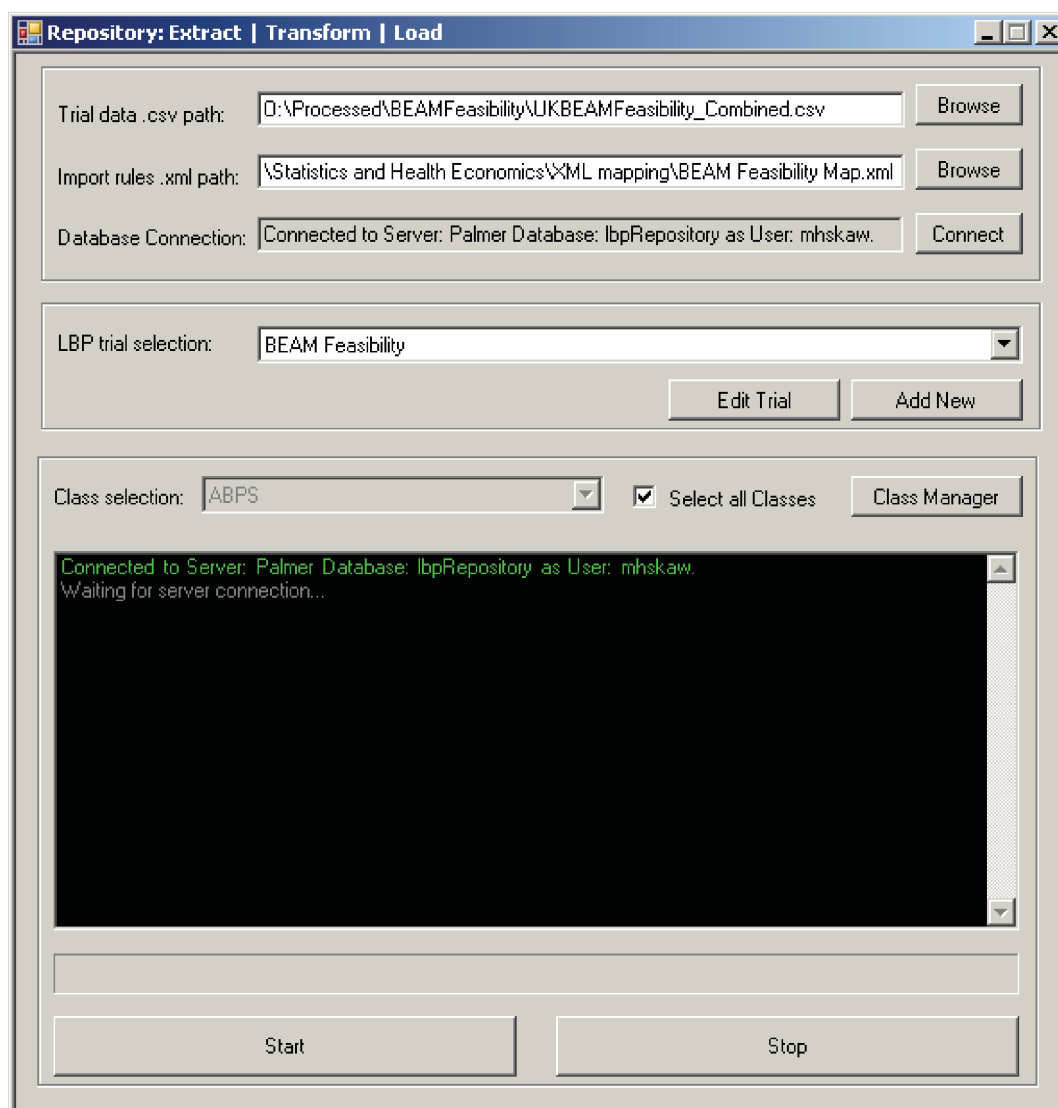


Figure A.1 The screenshot of the ETL program.

B. Screenshots of SPSS

Figure B.1 Screenshot of steps (a) – (b) to access Repository data with SPSS as given in Section 7.4. This page has been left intentionally blank. For a copy of the screenshots, please contact the corresponding author.

Project Specific Guide for the Low Back Pain Repository Analysis
Plan for the Low Back Pain Repository Transfer, Query, Map,

Effective: 9 December 2013

Version 1.0

Project Specific Guide for the Low Back Pain Repository Analysis
Plan for the Low Back Pain Repository Transfer, Query, Map,

Effective: 9 December 2013

Version 1.0

C. Screenshots of STATA

Figure C.1 Screenshots of step (a) – (c) to access Repository data with STATA given in Section 7.5. This page has been left intentionally blank. For a copy of the screenshots, please contact the corresponding author.

Project Specific Guide for the Low Back Pain Repository Analysis
Plan for the Low Back Pain Repository Transfer, Query, Map,

Effective: 9 December 2013

Version 1.0

A decorative graphic consisting of numerous thin, parallel green lines that curve from the left side of the page towards the right, creating a sense of movement and depth.

EME
HS&DR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

This report presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health

Published by the NIHR Journals Library