IDENTIFYING BACK PAIN SUBGROUPS; DEVELOPING AND APPLYING APPROACHES USING INDIVIDUAL PATIENT DATA COLLECTED WITHIN CLINICAL TRIALS

Patel S,[1]* Hee SW,[1] Mistry D,[1] Jordan J,[2] Brown S,[3] Dritsaki M,[1] Ellard D,[1] Friede T,[4] Lamb SE,[1,5] Lord J,[2] Madan J,[1] Morris T,[6] Stallard N,[1] Tysall C,[3] Willis A,[1] Underwood M[1]

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

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

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

[4] Department of Medical Statistics, University Medical Centre Göttingen, Göttingen, Germany

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

[6] Leicester Clinical Trials Unit, Diabetes Research Centre, University of Leicester, Leicester, UK

***Corresponding author:** Dr Shilpa Patel
Warwick Clinical Trials Unit, Division of Health Sciences, Warwick Medical School,
The University of Warwick, Coventry, CV4 7AL.
Tel: +44(0)24 7615 0405 e-mail: shilpa.patel@warwick.ac.uk

Declared competing interests of authors: Authors report grants from the National Institute for Health Research during the conduct of the study. SL is chair of the HTA CET Board.

Disclaimer - This report presents independent research commissioned 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 Programme Grants for Applied Research or the Department of Health.

Keywords - IPD, moderators, health economics; indirect network meta-analysis; subgroups; back pain; statistical methods

Total word count including appendix: 93,499

ABSTRACT

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

Overall aim was to improve the clinical 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 testing therapist-delivered interventions for LBP
- determine which participant characteristics, if any, predict clinical response to different treatments for LBP
- determine which participant characteristics, if any, predict the most cost-effective treatments for LBP.

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 main effect of interventions and no such interpretations should be made.

Methods

Firstly, 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 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 dataset we sought to identify which participant characteristics, if any, predict response to different treatments (moderators) for clinical and cost effectiveness outcomes.

We did an ANCOVA 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 to identify sub-groups defined by multiple parameters.

Results

We included data from 19 randomised controlled trials with 9,328 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 sub-group identification that we identified. Additionally we explored the application of Bayesian indirect network meta-analysis. This method produced varying probabilities that a particular treatment choice would be most likely to be effective for a specific patient profile.

Conclusion

These data lack clinical 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

Funding from the NIHR Programme Grants for Applied Research (RP-PG-0608-10076).

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF BOXES

GLOSSARY

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

Cross walking – this is a method of mapping multiple participant-reported outcome measures that measure the same domain, to a common scale.

Moderator – These are factors measured prior to randomisation and 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 – this is the score divided by the standard deviation of the baseline score of all participants.

LIST OF ABBREVIATIONS

AIP	active individual therapy
ALBPSQ	acute low back pain screening questionnaire
ANCOVA	analysis of covariance
ARDP	adaptive risk group refinement
ATP	active physical treatment
AUC	area under the curve
BBQ	back beliefs questionnaire
BDI	Beck depression inventory
BMI	body mass index
CBT	cognitive behavioural treatment
CENTRAL	Cochrane Central Register of Controlled Trials
CI	confidence interval
CES-D	Center for Epidemiologic Studies Depression
COST	European Cooperation in Science and Technology
CNSLBP	chronic non-specific low back pain
CPG	chronic pain grade scale
CRF	case report form
CSQ	coping strategy questionnaire
DASS	depression anxiety stress scales
DD	distressed-depressive
DRAM	distress and risk assessment method
DS	distressed-somatic
FABQ	fear-avoidance beliefs questionnaire
FFbHR	Hannover functional ability questionnaire for measuring back-pain related functional limitations (Funktionsbeeintrachtigung durch Ruckenschmerzen)
FRP	functional restoration program

GP	general practitioner
GPE	global perceived effect
HADS	hospital anxiety and depression scale
IASP	International Association of the Study of Pain
INMB	incremental net monetary benefit
IPD	individual patient data
IT	interaction tree
LBP	low back pain
MAR	Missing at random
MC	motivational counselling
MCS	mental component score
MI	multiple imputation
MNAR	missing not at random
MSPQ	modified somatic perception questionnaire
MVK	Modified Von Korff
MZDI	modified Zung depression index
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NMB	net monetary benefit
NSLBP	non-specific low back pain
ODI	Oswestry disability index
OPCO	operant behavioural treatment with cognitive coping skills training
PCS	physical component score
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

RCT	randomised control trial
RMDQ	Roland Morris disability questionnaire
SES	pain experience scale (Schmerzempfindungsskala)
SIDES	subgroup identification based on a differential effect search
SMD	standardised mean difference
SMT	spinal manipulative therapy
TENS	transcutaneous electrical nerve stimulation
TSK	Tampa scale for kinesiophobia
VAS	visual analogue scale
WLC	waiting list control

SCIENTIFIC SUMMARY

Background: Identifying subgroups of people living with low back pain who may do better, or worse, with different treatment choices is a high research priority internationally. Many 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 one to two million pounds, and taking at least six years from design to implementation); each will only address one small part of this complex problem.

Alternative methods can provide complementary information that could add value to our knowledge. Approaches, which 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 already invested in this area.

We present a programme of work, using systematic reviews, methodological development, and secondary analyses of existing datasets 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 datasets. The analysis plan for these data and modelling of clinical and cost-effectiveness are informed by our literature reviews.

Aims and objectives: The overall aim of this programme grant was to improve the clinical and cost-effectiveness of therapist delivered treatments for low back pain 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. Our objectives were:

1. To synthesise what is already known about the validity, reliability and predictive value of possible treatment moderators (participant factors that predict response to treatment).
2. To develop a repository of individual participant data from RCTs testing therapist-delivered interventions for LBP.

3. To determine which participant characteristics, if any, predict clinical response to different treatments for LBP
4. To 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 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, a review of the quality of subgroup analyses in low back pain 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 ($0.05<P\leq0.20$). 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 it provided a starting point for application in our analyses.

The second review looked at the quality and reporting of subgroup analyses in low back pain. 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. sub-group 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 RCTs testing therapist-delivered interventions for LBP.

To allow the identification of subgroups in appropriately powered datasets we developed a repository of data from completed trials. We used a systematic approach in identifying trials and approaching chief investigators for their data. Our pool of potential trials came from the search results generated in our review of moderators. As a starting point we were only interested in randomised control trials 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 of 19 trials, with 9,328 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 firstly 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 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 datasets 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 due to the lack of correlation and responsiveness in outcomes from two measures in the same individual, it would not appropriate to map any physical disability outcome measures to another.

To determine which participant characteristics, if any, predict clinical response to different treatments for LBP

We did ANCOVA analyses comparing all 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, Roland and Morris Disability questionnaire, SF12/36 physical and mental component scores, age, gender, pain, fear avoidance and coping as variable with a possible signal in one or more analysis.

In this programme grant we have explored in considerable details 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.

Our pre-specified analytical approaches; recursive partitioning and adaptive risk group refinement produced identifiable subgroups whose parameter definitions were grounded in the data. The differences in effect sizes, between groups, were however 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 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 deliver interventions tested. Targeting low intensity interventions at those higher levels of psychological distress for treatment might not be justified.

We did a post hoc exploratory individual participant data indirect network meta-analysis to identify sub-groups. 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 LBP.

We applied the directed peeling algorithm to the economic and resource use data. When exploring interventions vs 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 due to lack of power, or simply that there is no subgroup to be found.

Age, SF12/36 physical component score and Roland and Morris Disability score were the three potential moderators identified from the economic analysis. However the relationship of the Quality Adjusted Life Years (QALY) with the moderators differed in some cases to that of the clinical outcome measures. Subgroups were only identified in the comparison of treatment vs. control. Our interpretation is that those who are older, with worse RMDQ and SF12/36 physical component score are likely to gain a greater benefit on QALY outcomes from treatment. Doing this will not, however, improve overall QALY gain and is very unlikely to be seen as a cost-effective choice if the NICE 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 to 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 datasets, highlighted the it is not possible to map different outcome measures for a meta-analyses, and finally we have developed an important resource for back pain researchers wishing to do 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 that 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
- Re-analysis 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 we already have
- Making the dataset available to other researchers
- Adding additional trial datasets to the repository
- Developing and testing a web portal to help inform choice of treatments based on our network meta-analysis.

Overall, our results do not provide sufficient clinical or cost-effectiveness justification for the use of baseline characteristics in the development of subgroups for low back pain. We would however suggest such methods are applied in other clinical areas where subgroups may be important. The exploratory outputs from our Bayesian network meta-analysis provides some scope for deciding on optimal therapies. This however would need to be empirically testing before clinical recommendation.

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 therapist 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 save the National Health Service resources. To address this we pooled together data from 19 back pain trials from around the world. This provided us with a dataset of 9,328 patients. We developed novel statistical methods to identify sub-populations (groups of people with similar characteristics) likely to benefit from certain treatments. Of the three methods developed, two allowed us to identify sub-populations. The additional benefits for individuals in the sub-populations 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 sub-populations 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 sub-populations 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 .

CHAPTER 1 – OVERVIEW OF THE PROGRAMME

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

1.1 BACKGROUND

Chronic non-specific low back pain (CNSLBP) is a common problem affecting a large proportion of the population.[1-4] In the UK around 70 to 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 one year period.[5] The 2010 Global Burden of Disease study identified LBP as the leading cause of years lived with disability internationally.[8] Low back pain affects around a 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 that 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 to 1998, a 12% to 33% point prevalence; 22% to 65% 1-year prevalence and up to 84% life time 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 to 80 years. After adjusting for methodological variations the point prevalence of back pain lasting more than one day was 11.9% (95% confidence interval (CI), 7.98 to 15.82) and one month period prevalence was estimated at 23.2% (95% CI, 17.52 to 28.88).[12]

1.2 DEFINING LOW BACK PAIN

The International Association of the Study of Pain (IASP) define 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 review 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%.[16] Once specific causes for LBP have been excluded (malignancy, fracture, infection, inflammatory disorders such as ankylosing spondylitis) then a diagnosis of non-specific low back pain (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 2231/100 000 (95% CI 1555 to 3015) to 2313/100 000 (95% CI, 1574 to 3113) of the age standardised population.[17]

1.3 ECONOMIC BURDEN OF LOW BACK PAIN

Low back pain is a costly condition to society, healthcare and the individual. It is the leading cause of sickness absence and health care use.[18-21] In the UK the direct healthcare costs of back pain in 1998 was £1632 million. However the larger burden is that of the indirect costs related to lost production and informal care which were estimated to be at least £5,018 million.[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 eight and 12 million General Practitioner (GP) consultations per year.[22, 24] In 2013 the Office of National Statistics reported a 131 million lost working days due to sickness absences in that year in the UK; 30.6 million of these (23%) were lost due to musculoskeletal conditions including back and neck pain.[25]

1.4 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 its management was based largely on expert opinion, custom and practice. Since the mid-1990s, there has been a substantial investment in high quality randomised control 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, registered complementary practitioners such as osteopaths, or chiropractors or by sometimes unregistered professionals providing treatments such as acupuncture or 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 guidance advised that all people with persistent LBP should be given advice and encouraged to self-manage. As part of this advice they are encouraged to remain physically active and to engage in daily activity. Subsequently those affected should be offered a course of acupuncture needling, exercise, or manual therapy.[15] The decision on which treatment to select should be a collaborative decision taking into account the patients 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 they should be considered for an intense physical and psychological intervention. NICE is currently revising their LBP guidelines.

1.5 EFFECTIVENESS AND COST EFFECTIVENESS OF TREATMENTS FOR LOW BACK PAIN

Whilst 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),[28] the most commonly used outcome measure in back pain trials has been established as five points.[29, 30] Typical between group differences in high-quality randomised controlled trials are in the order of 1 to 2 points on the Roland Morris Disability Questionnaire; 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 to 10.[29, 31] These are of a similar to the numbers needed to treat found with antidepressant or anti-epileptic drugs used to treat chronic painful disorders.[32]

The cost per quality adjusted life year (QALY) for some of these treatments are well within cost-effectiveness thresholds usually used by NICE. In spite of this evidence access to such treatments within the National Health Service (NHS) remains patchy. The guideline endorsed treatments of interdisciplinary rehabilitation, exercise, acupuncture, spinal manipulation and cognitive-behavioural therapy for sub-acute 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.[33] The scarcity of economic evaluations for some guideline endorsed treatments means well-conducted economic evaluations are required to strengthen the evidence-base of treatments for LBP.

Table 1 Between group differences for the Roland Morris Disability Questionnaire

			Mean Difference in RMDQ[a] (95% CI),p-value	
Study	**Control**	**Intervention**	**3 month**	**12 month**
UK BEAM[b] [34]	GP care	Exercise	1.36 (0.63, 2.10); 0.34	0.39 (-0.41, 1.19); 0.10
		Manipulation	1.57 (0.82, 2.32); 0.39	1.01 (0.22, 1.81); 0.25
		Manipulation plus exercise	1.87 (1.15, 2.60); 0.47	1.30 (0.54, 2.07); 0.33
A-TEAM[c] [35]	Usual care	Massage	1.96 (0.74, 3.18); 0.39	0.58 (0.77, 1.94); 0.12
		Alexander technique (6 sessions)	1.71 (0.47, 2.95); 0.34	1.40 (0.03, 2.77); 0.28
		Alexander technique (12 sessions)	2.91 (1.66, 4.16); 0.58	3.40 (2.03, 4.76); 0.68
BeST[d] [31, 36]	Advice only	Cognitive behavioural therapy	1.10 (0.38, 1.71); 0.22	1.30 (0.56, 2.06); 0.27
York Yoga [37]	Usual care	Yoga	2.17 (1.03, 3.31); 0.50	1.57 (0.42, 2.71); 0.36

a RMDQ, Roland Morris disability questionnaire; b UK BEAM, United Kingdom Back pain Exercise And Manipulation; c A-TEAM, Alexander technique lessons, exercise, and massage; d BeST, Back Skills Training Trial.

1.6 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.

Since different treatment options are agued to work in very different ways it is a reasonable hypothesis that by matching people with LBP to those treatments more likely to be effective for their back pain will be a more efficient use of health care resources and improve patient outcomes. One might expect that people with high levels of psychological distress related to their back pain may gain greater benefit from a psychologically oriented 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 StartBack trial successfully demonstrated that a combination of using a stratification tool, and enhanced physiotherapy packages for selected participants, improves outcomes, and reduces costs, when compared to usual physiotherapy care.[38] This study does not, however, allow the performance of the stratification tool to identify subgroups to be assessed.

There are 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 only address 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 already invested in this area.

We present a programme of work, using systematic reviews, methodological development, and secondary analyses of existing datasets 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 datasets. The analysis plan for these data and modelling of clinical – cost effectiveness are informed by our literature reviews.

1.7 AIM AND OBJECTIVES

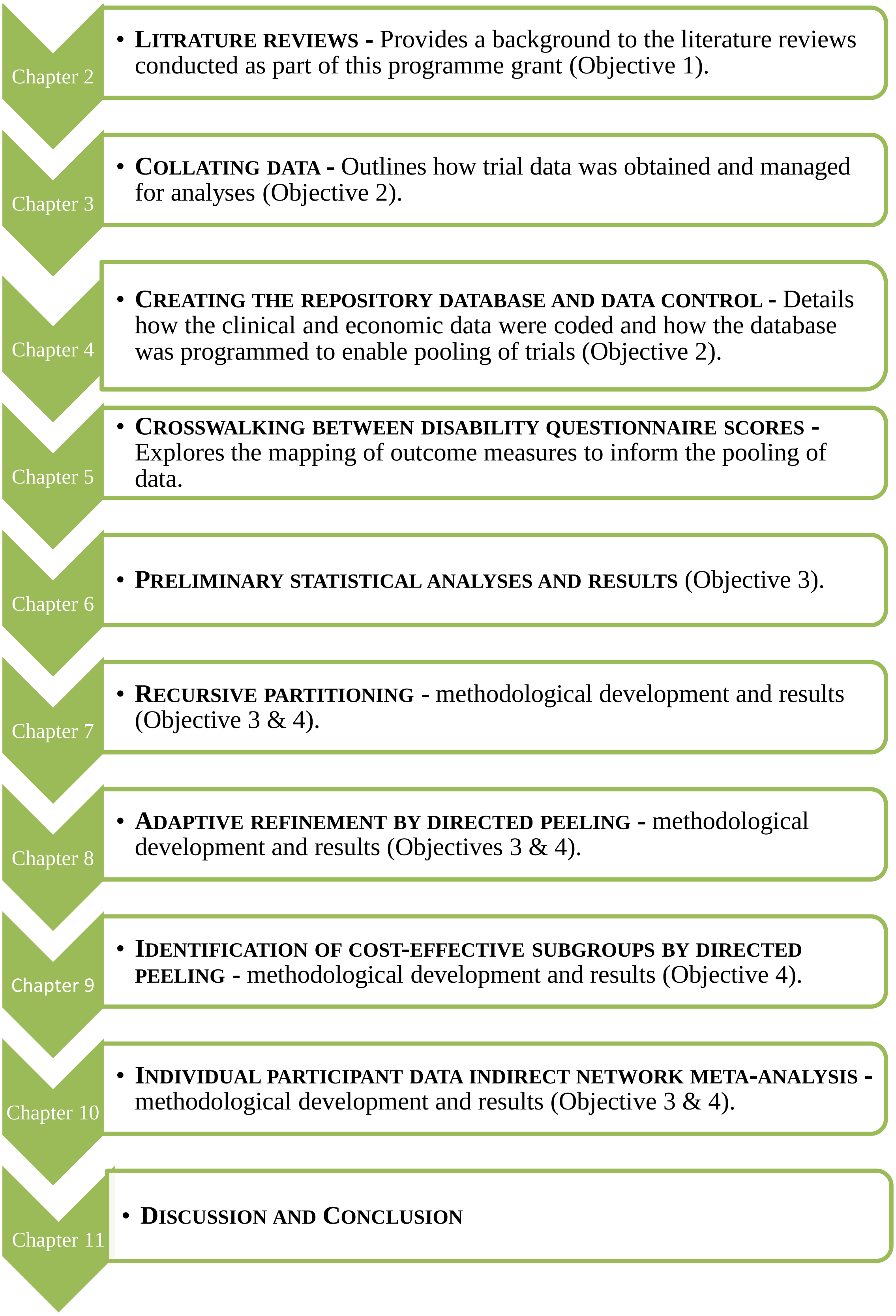
The overall aim was to improve the clinical and cost-effectiveness of LBP treatment by providing participants, their clinical advisors, and health service purchasers with better information about which participants are most likely to benefit from which treatment choices. To achieve this, our objectives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

1.8 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.



Chapter 2

- **Litrature reviews -** Provides a background to the literature reviews conducted as part of this programme grant (Objective 1).

Chapter 3

- **Collating data -** Outlines how trial data was obtained and managed for analyses (Objective 2).

Chapter 4

- **Creating the repository database and data control -** Details how the clinical and economic data were coded and how the database was programmed to enable pooling of trials (Objective 2).

Chapter 5

- **Crosswalking between disability questionnaire scores -** Explores the mapping of outcome measures to inform the pooling of data.

Chapter 6

- **Preliminary statistical analyses and results** (Objective 3).

Chapter 7

- **Recursive partitioning -** methodological development and results (Objective 3 & 4).

Chapter 8

- **Adaptive refinement by directed peeling -** methodological development and results (Objectives 3 & 4).

Chapter 9

- **Identification of cost-effective subgroups by directed peeling -** methodological development and results (Objective 4).

Chapter 10

- **Individual participant data indirect network meta-analysis -** methodological development and results (Objective 3 & 4).

Chapter 11

- **Discussion and Conclusion**

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.

2.1 SYSTEMATIC REVIEW 1 – IDENTIFICATION OF POTENTIAL MODERATORS

This review has been published in Physiotherapy. Here we present a summary of the paper.[39]

2.1.1 Abstract

Background: Within randomised controlled trials, moderators are baseline characteristics that predict whether an intervention will be more or less effective for an individual in the trial. For our final individual participant data meta-analyses we needed to select 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 dataset.

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 and 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 randomised controlled trials with ≥500 participants, and cohort studies of ≥1000 participants. We classified potential moderators into those with strong (p<0.05) or weaker evidence (p<0.20, ≥0.05).

Results: We identified 914 potential citations. We selected 64 papers for detailed evaluation. Four papers, all randomised controlled trials, 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 was weak and lacking in 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.

2.1.2 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 datasets are required to provide sufficient statistical power to detect any interaction between the moderator and treatment.[41]

2.1.3 AIMS

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

2.1.4 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$^{(R)}$ In-Process & Other Non-Indexed Citations
- EMBASE
- Web of Science
- Citation Index and Cochrane Central Register of Controlled Trials (CENTRAL)

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

2.1.4.1 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 only allowed identification of 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. 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.

2.1.4.2 Minimum sample size for included studies

To allow us to pick up 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) 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 1,000 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 (between group difference/baseline standard deviation) of 0.5 for the interaction at a 0.05 level of significance and 80% power. The standardised mean differences in high-quality RCTs of therapist delivered interventions for LBP are typically in range 0.1 to 0.7 (see *Chapter 1, Table 1*). Smaller trials would only be able to detect treatment moderation, at this level, 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 $0.05<p\leq0.20$ 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.

2.1.4.3 Inclusion and exclusion criteria

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

Box 1 Review 1 - Inclusion and exclusion criteria

Inclusion Criteria

- Aged 18 and over
- NSLBP of any duration
- Therapist delivered interventions
- RCTs with sample size of ≥500
- non-RCTs and observational studies with sample size ≥1,000
- 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.

2.1.4.4 Screening and data extraction

At all stages two researchers (TG & 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 onto 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 methodological considerations related to observational studies further.

2.1.4.5 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.

2.1.5 RESULTS

Our initial electronic searches generated 7,208 hits; 6,294 were removed based on title, abstract and duplicates. We obtained 64 papers for detailed review; of these 60 were excluded (see *Figure 2*). Four studies were included in this review (see *Table 2*). All four trials were RCTs comprising of a total sample of n = 5,514.

Once we had identified these paper 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 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 paper 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 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.

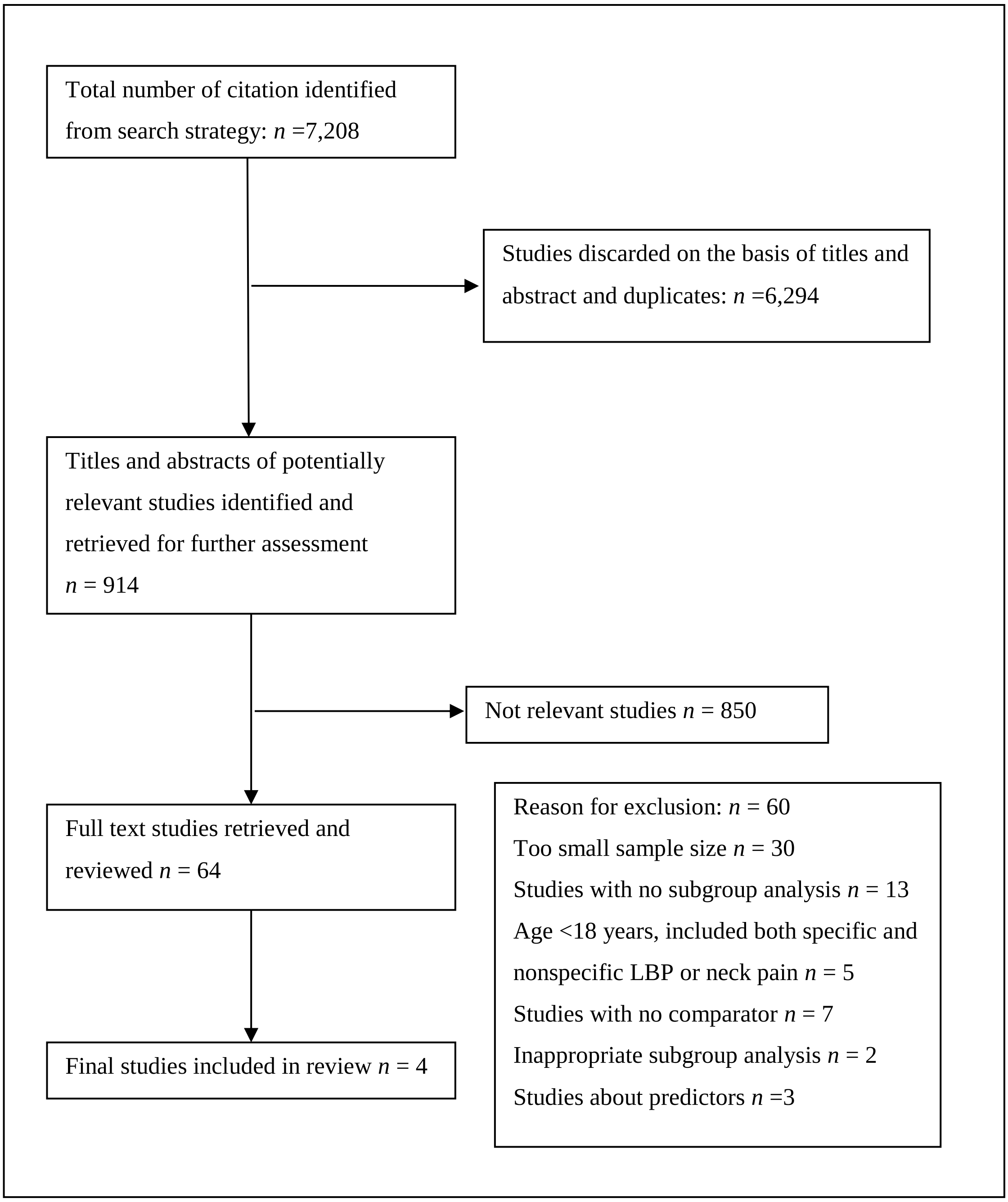

Total number of citation identified from search strategy: *n* =7,208
Studies discarded on the basis of titles and abstract and duplicates: *n* =6,294
Titles and abstracts of potentially relevant studies identified and retrieved for further assessment *n* = 914
Not relevant studies *n* = 850
Full text studies retrieved and reviewed *n* = 64
Reason for exclusion: *n* = 60
Too small sample size *n* = 30
Studies with no subgroup analysis *n* = 13
Age <18 years, included both specific and nonspecific LBP or neck pain *n* = 5
Studies with no comparator *n* = 7
Inappropriate subgroup analysis *n* = 2
Studies about predictors *n* =3
Final studies included in review *n* = 4

Figure 2 Review 1 - Quorum statement flow diagram

Table 2 Review 1 - Included studies

Study	Country	Sample	Interventions
UK BEAM[45]	UK	1,334	Group exercise, manual therapy and combination therapy
BeST[46]	UK	701	Group cognitive behavioural approach
Witt[47]	Germany	2,841	Acupuncture
Cherkin[48]	USA	638	Acupuncture

2.1.5.1 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 (see *Table 3* and *Table 4*).

Table 3 Review 1 - Results of the risk of bias assessment

Quality of the study based on main trial paper/s	UK BEAM[34]	BeST[31, 36]	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[a]	NIHR HTA[b]	Social Health Fund Providers	National Institutes of Health

L, Low risk of bias; H, High risk of bias; U, Unclear; a MRC, Medical Research Council; b, NIHR HTA,, National Institute for Health Research Health Technology Assessment

Table 4 Review 1 - Results of methodological quality assessments

Quality of the moderator analyses based on subgroup paper/s	UK BEAM	BeST	Witt	Cherkin
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 randomization?	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

Y, Yes N, No; U, Unclear; EE, exploratory evidence - fulfils three, four or five criteria for moderator studies; CE, confirmatory evidence- fulfils all five criteria for moderator studies, IE, insufficient evidence to judge quality.

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

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[a]	MVK[b] pain	MVK disability
BeST[36, 46]	Troublesomeness (Very/Extremely – Moderately)	*p* = 0.190 -1.01 (-2.52, 0.50)	*p* = 0.184 -5.04 (-12.47, 2.40)	NS[c]
	Age (≥54 years – <54 years)	*p* = 0.035 -1.58 (-3.05, -0.12)	NS	NS
	Female – Male	*p* = 0.102 -1.27 (-2.79, 0.25)	NS	NS
	Left FT Education (>16 years of age – ≤16 years of age)	*p* = 0.098 1.29 (-0.24, 2.82)	NS	NS
	Employed – Not Employed	*p* = 0.011 1.89 (0.43, 3.35)	*p* = 0.181 5.01 (-2.33, 12.34)	NS
	HADS – Anxiety (≥11 – <11)	*p* = 0.195 -1.12 (-2.83, 0.58)	NS	NS
	HADS – Depression (≥11 – <11)	*p* = 0.135 -2.07 (-4.79, 0.65)	NS	*p* = 0.051 -14.58 (-29.19, 0.03)

Study ID	Potential moderators	Significant interactions; outcome, Roland Morris Disability Questionnaire (RMDQ[a])					
		8 Weeks			52 Weeks		
		IA[d]	StA[e]	SiA[f]	IA	StA	SiA
Cherkin[48, 49]	Age	NS	$p = 0.08$ 0.08 (-0.02, 0.18)	NS	NS	$p = 0.15$ 0.07 (-0.03, 0.17)	NS
	Self-efficacy	$p = 0.04$ -6.17 (-12.01, -0.33)	NS	NS	NS	NS	NS
	RMDQ (B/L)[g]	$P < 0.0001$ -0.48 (-0.72, -0.24)	$p = 0.004$ -0.37 (-0.62, -0.12)	$p = 0.001$ -0.41 (-0.66, -0.16)	$p = 0.07$ -0.23 (-0.48, 0.02)	$p = 0.07$ -0.24 (-0.49, 0.01)	NS
	Bothersomeness score (B/L)	NS	$p = 0.10$ 0.47(-0.10-1.04)	NS	NS	NS	NS
	Heavy lifting	$p = 0.03$ 4.29 (0.43, 8.15)	$p = 0.13$ 3.00 (-0.86, 6.86)	$p = 0.18$ 2.73 (-1.27, 6.73)	$p = 0.01$ 5.19 (1.17, 9.21)	$p = 0.15$ 3.03 (-1.05, 7.11)	$p = 0.04$ 4.45 (0.28, 8.62)
	Sedentary	NS	NS	NS	$p = 0.12$ 2.73 (-0.72, 6.18)	$p = 0.15$ 2.47 (-0.90, 5.84)	NS
	Use of narcotic medication	$p = 0.08$ 3.52 (-0.38, 7.42)	NS	$p = 0.01$ 4.81 (0.97, 8.65)	NS	$p = 0.04$ 4.06 (0.18, 7.94)	$p = 0.19$ 2.71 (-1.31, 6.73)
	Acupuncture expectation (top tertile)	$p = 0.05$ -2.65 (-5.28, -0.02)	NS	NS	NS	$p = 0.17$ -1.9 (-4.60, 0.80)	$p = 0.03$ -2.91 (-5.56, -0.26)

Study ID	Potential moderators	Significant interactions; outcome, bothersomeness score					
		8 Weeks			52 Weeks		
Cherkin[48, 49] (cont)	Age	NS	*p* = 0.09 0.04 (0.001, 0.08)	*p* = 0.07 0.04 (0.001, 0.08)	NS	*p* = 0.15 0.04 (-0.02, 0.10)	*p* = 0.08 0.05 (-0.01, 0.11)
	Self-efficacy	*p* = 0.14 -2.21 (-5.13, 0.71)	NS	NS	NS	NS	NS
	Baseline RMDQ score	*p* = 0.01 -0.15 (-0.27, -0.03)	NS	*p* = 0.0005 -0.22 (-0.34, -0.10)	*p* = 0.16 -0.09 (-0.23, 0.05)	NS	NS
	Heavy lifting	*p* = 0.05 1.97 (0.03, 3.91)	NS	*p* = 0.04 2.10 (0.10, 4.10)	*p* = 0.02 2.51 (0.43, 4.59)	NS	NS
	Light/medium lifting	NS	*p* = 0.12 -1.28 (-2.87, 0.31)	NS	*p* = 0.12 1.35 (-0.36, 3.06)	NS	NS
	Sedentary	NS	NS	NS	*p* = 0.19 1.20 (-0.58, 2.98)	NS	NS
	Acupuncture expectation (top tertile)	*p* = 0.10 -1.10 (-2.41, 0.21)	NS	NS	*p* = 0.051 -1.44 (-2.87, -0.01)	NS	*p* = 0.06 -1.29 (-2.64, 0.06)

Study ID	Potential moderators	3 months for RMDQ outcome	12 months for RMDQ outcome
		Combined treatment	Combined treatment
UK BEAM [34, 45]	Quality of life	p = 0.174 -0.1 (-0.26, 1.43)	NS
	Treatment expectation (helpful)	p = 0.073 -3.2 (-6.74, 0.30)	p = 0.038 -3.8 (-7.39, -0.20)
	Treatment expectation (very helpful)	p = 0.192 -2.2 (-5.49, 1.11)	p = 0.019 -4.0 (-7.38, -0.67)
		Manipulation	Manipulation
	Beliefs	p = 0.07 -0.8 (-1.62, 0.06)	NS
	Quality of life	p = 0.118 1.4 (-0.35, 3.07)	NS
	Pain/Disability	p = 0.176 -1.9 (-4.61, 0.85)	p = 0.143 -2.2 (-5.16, 0.75)
	Treatment expectation (helpful)	NS	p = 0.083 -0.1 (-0.16,0.01)
	Treatment expectation (very helpful)	p = 0.113 1.6 (-0.38, 3.60)	NS

Study ID	**Potential moderators**	Outcome, Hannover functional ability questionnaire for measuring back-pain related functional limitations (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

a RMDQ, Roland Morris disability questionnaire; b MVK, Modified Von Korff; c NS, no significant interaction found; d IA, individualised acupuncture; e StA: standardised acupuncture; f SiA, simulation acupuncture; g B/L, baseline.

2.1.5.1.1 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 greater than 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, Cherkin and Witt trials found an interaction with age.[46, 48, 50] In BeST, younger participants gained more benefit from cognitive behavioural therapy than older participants on the RMDQ score.[46] 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 found a statistically significant additional benefit from acupuncture treatment in younger participant (p <0.001).[50]

Gender

BeST found that gender had a moderating effect on treatment.[46] In this trial, females had comparatively greater improvement following group cognitive behavioural therapy compared to 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 BeST, the authors found that employed participants gained additional benefit from a cognitive behavioural approach when compared to those who were not employed. The treatment difference between employed and not employed 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 MVK pain score respectively. The interaction effect in the analysis of the MVK pain score was weak.[46]

The Cherkin trial found some moderating effect according to types of employment status. The participants in this trial received acupuncture therapy.[48] Those participants whose job involved lifting heavy materials showed positive moderating effect against back related dysfunction score at eight 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 eight 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

BeST found that participants who had left full-time education after the age of 16 had better improvement from cognitive behavioural therapy compared to participants who left full time education aged 16 years or less.[46] The treatment difference was 1.29 (p = 0.098; 95% CI -0.24 to 2.82) for the RMDQ score. The interaction effect was greater than 0.05 therefore this provides weak evidence. Witt found that those participants who have had more than 10 years of schooling gained a greater benefit from acupuncture (p = 0.01).[50]

Back pain status

In the Cherkin and Witt trials participants with a worse initial back pain status (baseline RMDQ) gained an increased benefit from acupuncture when compared to those with a better back pain status at baseline (p-values ranged from <0.001 to 0.16).[48, 50] 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 greater than 0.05. In the Cherkin trial, the *p-value* was 0.10 while in the BeST trial, 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.[46, 48]

Pain/disability

Similarly, those participants with greater pain/disability at baseline seemed to benefit more at three months (p = 0.176) and 12 months (p = 0.143) for the RMDQ score with manipulation treatment (UK BEAM) (see *Table 5*). The p-values are greater than 0.05 and less than 0.2 therefore providing weak evidence.[45]

Narcotic

Cherkin 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.[48]

Treatment expectations

Having better expectations about the treatment was found to be a moderating factor in two trials.[45, 48] The p-values ranged between p = 0.03 and p = 0.192 demonstrating a spectrum of strong to weak evidence for the interactions.[48]

Cherkin found that participants with higher expectation of acupuncture treatment helpfulness gained more benefit in the back related dysfunction score (p = 0.03 to 0.17) and bothersomeness score (p = 0.05 to 0.10).[48] In the UK BEAM trial, manipulation at three 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 three and 12 months showed positive moderating effect as was demonstrated by the RMDQ score.[45] 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 BeST, psychosocial status moderated treatment effect. The trial 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 score. 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 score respectively.[46]

2.1.6 DISCUSSION & 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 identified as moderators of treatment effect at $p<0.05$ as potential moderators with strong evidence and those at $p<0.20$, ≥ 0.05 as potential moderators with weak evidence. Only for two comparisons, in one study, 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 where there is some weak evidence of moderation that is worth exploring further.

2.2 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.

2.2.1 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 randomized controlled trials of therapist delivered interventions for NSLBP.

Method: Electronic databases were searched for randomized controlled trials of therapist delivered interventions for NSLBP. We only included papers reporting 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 ten reported results from interaction tests, four incorrectly reported results within individual subgroups and the remaining papers either reported *p-values* or nothing at all.

Conclusion: Subgroup analyses performed in NSLBP trials have been severely underpowered, are only able to provide 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.

2.2.2 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]

2.2.3 AIMS

The objective of this literature review is to firstly identify randomized controlled trials of therapist delivered interventions for NSLBP that 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 was also assessed.

2.2.4 METHOD

This literature review work was done 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$^{(R)}$ In-Process & Other Non-Indexed Citations
- EMBASE
- Web of Science
- Citation Index and Cochrane Controlled Trial Registered (CENTRAL)

2.2.4.1 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 re-ran searches to identify all ‘low back pain’ and ‘RCTs’ which we filtered for therapist delivered interventions.

2.2.4.2 Inclusion and exclusion criteria

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

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 x moderator x time
- Pooled datasets of similar trials.

Reproduced with permission from Lippincott, Williams & Wilkins publishers

2.2.4.3 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 onto a standardised extraction form and any discrepancies were resolved using a second reviewer.

2.2.4.4 Quality assessment of subgroup analysis

We used the same Pincus et al criteria described in the previous review (see *2.1.4.5 Risk of bias and quality assessment*) the review above to assess the quality of subgroups.[44]

Three independent reviewers (DM, SP, SWH) assessed 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.

2.2.4.5 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 it if they conformed to four key recommendations in the area of subgroup analyses (see *Box 3*).

Box 3 Key recommendations in the area of subgroup analyses.

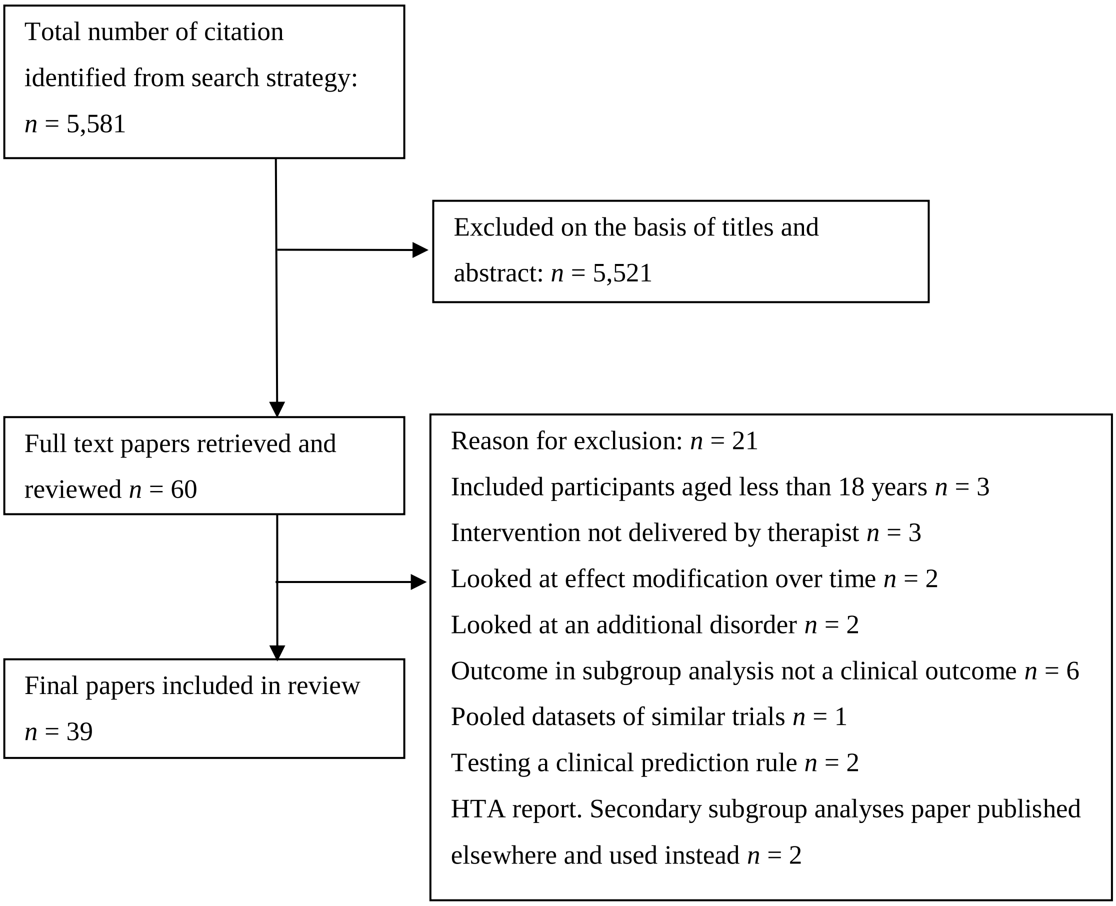
Key recommendations:

- 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 with permission from Lippincott, Williams & Wilkins publishers

2.2.5 Results

Our initial search identified 5,581 papers. All titles and abstracts were screened to identify potential papers reporting results of RCTs of therapist delivered interventions for LBP. We excluded 5,521 papers during the screening process. The full text for the remaining 60 papers were 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 (see *Figure 3*).


Total number of citation identified from search strategy: *n* = 5,581
Excluded on the basis of titles and abstract: *n* = 5,521
Full text papers retrieved and reviewed *n* = 60
Reason for exclusion: *n* = 21
Included participants aged less than 18 years *n* = 3
Intervention not delivered by therapist *n* = 3
Looked at effect modification over time *n* = 2
Looked at an additional disorder *n* = 2
Outcome in subgroup analysis not a clinical outcome *n* = 6
Pooled datasets of similar trials *n* = 1
Testing a clinical prediction rule *n* = 2
HTA report. Secondary subgroup analyses paper published elsewhere and used instead *n* = 2
Final papers included in review *n* = 39

Reproduced with permission from Lippincott, Williams & Wilkins publishers

Figure 3 Review 2 - Quorum statement flow diagram.

A summary of the included studies is given in *Table 6* and excludes studies in *Appendix 1*. A total of 63% of the included papers were from the Netherlands, UK or USA. The median study size was 223 and ranged from 100 to 3,093.

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

Subgroup Quality Assessment	Author	Published	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. Global perceived effect at 3 weeks.	None
	Smeets[59]	2009	Australia & New Zealand	259	Exercise and advice vs Exercise and sham advice vs Sham exercise and advice vs Sham exercise and sham advice	Pain intensity (11 point scale) and patient specific function scale (0-10 scale) measured at baseline 6 weeks and 52 weeks	None
	Underwood[46]	2011	UK	701	Advice plus Cognitive behavioural intervention vs Advice only	RMDQ[a] and MVK[b] measured at baseline, 3 months, 6 months and 12 months	Age & Employment

Subgroup Quality Assessment	**Author**	**Published**	**Country**	**Study Size**	**Interventions compared**	**Outcome measure and follow-up***	**Subgroups Identified (Interaction test only)**
Exploratory Findings	Becker[60]	2008	Germany	1,378	Multifaceted guideline implementation (GI) vs GI plus motivational counselling (MC) vs Postal dissemination of guideline (Control)	Hannover Functional Ability Questionnaire measured at baseline and 6 months	None
Exploratory Findings	Cecchi[61]	2012	Italy	210	Back school vs Individual physiotherapy vs Spinal manipulation	RMDQ measured at baseline, 3 months, 6 months and 12 months	None
Exploratory Findings	Cherkin[62]	1998	USA	321	Physical therapy vs Chiropractic manipulation vs Educational booklet	Bothersomeness of symptoms and RMDQ measured at baseline, 4 weeks and 12 weeks	Mental Health

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Exploratory Findings	Cherkin[63]	2001	USA	262	Chinese acupuncture vs Therapeutic Massage vs Self-care education	Bothersomeness of symptoms and RMDQ measured at baseline, 4 weeks, 10 weeks and 1 year	None
Exploratory Findings	Cherkin[49]	2009	USA	638	Individualised acupuncture vs Standardized acupuncture vs Simulated acupuncture vs Usual care	Bothersomeness of symptoms and RMDQ measured at baseline, 8 weeks, 26 weeks and 1 year	None

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Exploratory Findings	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
Exploratory Findings	Hay[65]	2005	UK	402	Brief pain management vs Manual physiotherapy	RMDQ measured at baseline, 3 months and 12 months	None
Exploratory Findings	Juni[66]	2009	Switzerland	104	Standard care alone vs Standard care plus Spinal Manipulative Therapy (SMT)	Pain intensity (11 point scale) and analgesic use measured at baseline, days 1 to 14 and 6 months	None

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Exploratory Findings	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 & type of occupation (comparing mini-intervention vs usual care and worksite visit vs usual care)

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Exploratory Findings	Kole-Snijders[68]	1999	Netherlands	159	Operant behavioural treatment with cognitive coping skills training (OPCO) vs Operant behavioural treatment with group discussion (OPDI) vs Waiting list control (WLC)	Main outcome unclear. Outcomes measured at post-treatment, 6 months and 1 year	None
Exploratory Findings	Roche[69]	2007	France	132	Active individual therapy (AIP) vs Functional restoration program (FRP)	Main outcome unclear. Outcomes measured at baseline and 5 weeks	Sorenson score

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Exploratory Findings	Sherman[48]	2009	USA	638	Individualised acupuncture vs Standardized acupuncture vs Simulated acupuncture vs Usual care	Bothersomeness of symptoms and RMDQ measured at baseline, 8 weeks, 26 weeks and 1 year	Baseline RMQ
Exploratory Findings	Smeets[70]	2006	Netherlands	223	Active physical treatment (ATP) vs Cognitive behavioural treatment (CBT) vs Combined APT and CBT (CT) vs Waiting list (WL)	RMDQ measured at baseline, 10 weeks, 6 months and 12 months	Baseline RMQ

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Exploratory Findings	Smeets[71]	2008	Netherlands	223	Active physical treatment (ATP) vs Graded activity with problem solving training (GAP) vs Combination treatment (CT) vs Waiting list (WL)	RMDQ measured at baseline, 10 weeks, 6 months and 12 months	None
Exploratory Findings	Tilbrook[37]	2011	UK	313	Yoga vs Usual care	RMDQ measured at baseline, 3 months, 6 months and 12 months	None

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Exploratory Findings	Underwood[45]	2007	UK	1,334	Control (Best care in General Practice) vs Exercise programme vs Spinal manipulation vs Combined treatment (manipulation and exercise)	RMDQ measured at baseline, 3 months and 1 year	Expectation
Exploratory Findings	van der Hulst[72]	2008	Netherlands	163	Roessingh Back Rehabilitation (RRP) vs Usual care	RMDQ measured at baseline, 1 week after treatment and 4 months after treatment	Pain intensity & Depression

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Exploratory Findings	Witt[50]	2006	Germany	3,093	Acupuncture vs Control (delayed acupuncture treatment 3 months later)	Hannover Functional Ability Questionnaire (0-100 scale) measured at baseline, 3 months and 6 months	Initial back pain, age & years of schooling
Insufficient Findings	Bendix[73]	1998	Denmark	816	Functional restoration (FR) program vs Outpatients program (Control)	Main outcome unclear. Outcomes measured at baseline and 1 year	
Insufficient Findings	Beurskens[74]	1995	Netherlands	151	Traction vs Sham traction	GPE and severity measured on visual analogue scale (VAS) at baseline and 5 weeks	

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Insufficient Findings	Bishop[75]	2011	USA	112	Supine thrust technique vs Side-lying thrust vs Non-thrust technique	ODI[c] measured at 1 week, 4 weeks and 6 months	None
Insufficient Findings	Carr[76]	2005	UK	237	Group exercise programme vs Individual physiotherapy	RMDQ measured at baseline, 3 months and 6 months	
Insufficient Findings	Ferreira[77]	2009	Australia	191	General exercise vs Motor control exercise vs Spinal manipulative therapy	GPE[d] (11 point scale), Patient specific functional status, RMDQ, Pain intensity (10 point scale) and spinal stiffness measured at baseline and 8 weeks	None

Subgroup Quality Assessment	**Author**	**Published**	**Country**	**Study Size**	**Interventions compared**	**Outcome measure and follow-up***	**Subgroups Identified (Interaction test only)**
Insufficient Findings	Glasov[78]	2009	Australia	100	Laser acupuncture vs Sham acupuncture (control)	Pain (VAS) measured at baseline, immediately after treatment, 6 weeks and 6 months	
Insufficient Findings	Gudavalli[79]	2006	USA	235	Flexion distraction (FD) vs Active trunk exercise protocol (ATEP)	Perceived pain (VAS), RMDQ and SF-36 measured at baseline, 4 weeks, 3 months, 6 months and 1 year	
Insufficient Findings	Hsieh[80]	2004	China	146	Acupressure vs Physical therapy	Short-form pain questionnaire measured at baseline, 4 weeks and 6 months	

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Insufficient Findings	Jellema[81]	2005	Netherlands	314	Minimal intervention strategy (MIS) vs Usual care	RMDQ, perceived recovery (7 point scale) and sick leave measured at baseline, 6 weeks, 13 weeks, 26 weeks and 1 year	
Insufficient Findings	Johnson[82]	2007	UK	234	Group exercise and education using a cognitive behavioural approach vs Usual care	Pain (VAS) and RMDQ measured at baseline, 3 month, 9 month and 15 months	Patient preference
Insufficient Findings	Kalauokalani[83]	2001	USA	166	Acupuncture vs Massage (Subanalysis of Cherkin 2001 paper)	RMDQ measured at baseline, 4 weeks, 10 weeks and 1 year	Patient expectations

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Insufficient Findings	Mellin[84]	1989	Finland	456	Inpatient treatment vs Outpatient treatment vs Control (Advice)	Low back pain disability index (scale 0-45) measured at baseline and 3 months	
Insufficient Findings	Klaber Moffett[85]	2004	UK	187	Exercise vs Usual care	RMDQ measured at baseline, 6 weeks, 6 months and 1 year	
Insufficient Findings	Myers[86]	2008	USA	444	Usual care vs Usual care plus patient choice of acupuncture, chiropractic or massage	RMDQ measured at baseline, 5 weeks and 12 weeks	None

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Insufficient Findings	Seferlis[87]	1998	Sweden	180	Manual therapy program (MTP) vs Intensive training program (ITP) vs General practitioner program (GPP)	Main outcome unclear. Outcomes measured at baseline, 1 month, 3 months and 12 months	
Insufficient Findings	Thomas[88]	2006	UK	241	Traditional acupuncture vs Usual care	Bodily pain dimension of the SF-36 (0-100 scale) measured at baseline, 3 months, 12 months and 24 months	Expectation
Insufficient Findings	van der Roer[89]	2008	Netherlands	114	Intensive group training protocol vs Guideline group	RMDQ measured at baseline, 6 weeks, 13 weeks, 26 weeks and 52 weeks	

Subgroup Quality Assessment	Author	Published	Country	Study Size	Interventions compared	Outcome measure and follow-up*	Subgroups Identified (Interaction test only)
Insufficient Findings	Vollenbroek-Hutten[90]	2004	Netherlands	163	Roessing Back Rehabilitation (RRP) vs Usual care	RMDQ measured at baseline, 1 week after treatment and 4 months after treatment	

a RMDQ, Rolland and Morris Disability Questionnaire; b MVK, Modified Von Korff (pain and disability); c ODI, Oswestry disability index; d GPE, Global perceived effect; Reproduced with permission from Lippincott, Williams & Wilkins publishers

2.2.5.1 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:

- Three (8%) papers met all five criteria and therefore provided confirmatory evidence; Sheets[53] , Smeets[54], & Underwood[46, 58, 59]. Two of these were too small to anticipate finding any important interaction if it were present (*n* = 148 & 259)
- Eighteen (46%) papers provided exploratory evidence i.e. they met criteria three, four and five (see *Table 6*)
- Eighteen (46%) papers provided insufficient evidence (see *Table 6*).

2.2.5.2 Assessment of conduct and reporting of subgroups

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

- One study had sufficient power to detect an interaction however subgroups of interest were not pre-specified *a priori*[50]
- Thirty-one (79%) studies did not pre-specify subgroups of interest
- Eight studies reported pre-specified subgroups for confirmatory analyses[46, 58, 59, 64, 65, 75, 79, 82]; six of these also carried out exploratory analyses without clear distinction between analysis types.
- Sometimes it was not clear from methods that subgroups 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 paper reported the expected size and direction of the subgroup effect.[58] A further three papers predicted the direction of the subgroup effect
- A third (13/39) of the papers provided some justification regarding the choice of subgroups to be analysed.

- In two papers around sixty interaction tests were conducted substantially increasing the chances of detecting false positive findings.[45, 59] Of the three papers that provided confirmatory findings, only one of them adjusted for multiplicity. The authors applied a Bonferroni correction to their confirmatory subgroup analyses.[46]

- Twelve (31%) of the papers did not use a statistical test for interaction to assess for treatment effect modification. Of these, two of the papers did not give any indication as to what statistical method they used.[74, 87] Two papers looked at correlations between individual subgroups and outcomes within each treatment arm separately.[73, 84] Two papers used t-tests between treatment groups within individual subgroups.[79, 80] Five papers used either multiple linear regression or multiple logistic regression for each individual subgroup.[76, 81, 85, 89, 90] One paper compared the medians across three trial arms within individual subgroups using Kruskal-Wallis tests.[64]

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

- A statistical test for interaction was reported to have been used in 27 (69%) of papers

- Six studies reported both the interaction effect sizes with confidence intervals and the corresponding *p*-values[45, 48, 61, 72, 75, 77]

- Four studies reported only the interaction effect sizes with confidence intervals[46, 58, 59, 82]

- Eight studies reported the *p*-values only[37, 50, 66, 67, 69, 83, 86, 88]

- Nine papers did not report either the interaction effect sizes or confidence intervals or *p*-values.[49, 60, 62-65, 68, 70, 71]

- Four studies reported subgroup analyses within individual subgroups rather than between group interaction.[60, 66, 70, 88]

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

- Four 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 a third of the papers.
- The limitations of subgroup analyses were reported in 12 papers.

2.2.6 DISCUSSION & 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 only providing 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 trial was appropriately powered for the analysis, however the authors failed to specify the subgroups *a priori*.[50] 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 methodology concerns highlighted.

2.3 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. To design 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 where care would also need to be taken to select moderators that could be easily applied clinically.

In addition to these reviews we have previously published a systematic review which summarised findings from randomised controlled trials testing the effects of a clinical prediction rule for NSLBP.[93] Clinical prediction rules have been developed and are being used in clinical practice to help clinicians make decisions on treatment however the overall effect of such tools is unclear. Multi-component clinical prediction rules have the potential to be much more powerful tools for targeting treatments than single component measures. We identified 1,821 potential citations after all duplications had been removed. Two people independently screened the titles and abstracts, consensus was reached on obtaining 35 papers for full detailed evaluation. Of these only three papers 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 the existing RCTs looking to validate clinical prediction rules in LBP to be 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. 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.[94-96]
2. To apply individual participant data meta-analysis to homogenous pooled datasets 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 one and two, leaving point three for others in 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.

3.1 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 only interested in:

- randomised control trials
- trials of therapist delivered interventions
- those 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 trials not on our original list by researchers aware of the project. 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.

3.2 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 standardised mean difference 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 under 200; typically these were studies aiming for around 200 participants that fell short of the final target. We therefore revised our inclusion to >179. From a practical perspective of approaching trial investigators, this yielded a manageable number of trials to approach; large trials (those of thousands of participants) and small trials (less than 100 participants) each create a similar amount of work to collate.

3.3 PROCESS FOR APPROACHING INVESTIGATORS

We identified 42 trials which fitted our inclusion criteria. For these trials we identified the Chief Investigator and the best e-mail contact for them. Between 2011 and 2012 each investigator was sent an email to invite them to participate in the Repository. Each email included the following attachments:

- a 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 six to eight 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.

3.4 SECURE DATA TRANSFER

We requested all data for a trial. Investigators were advised that any datasets being sent to us needed to be anonymised and encrypted using an open-source compression software programme such as 7Zip (http://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 set up for the project available at https://files.warwick.ac.uk/repositorylbpdata/sendto

Once this 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 help aid this process we requested all trial specific information including the protocol and questionnaires if they were available.

3.5 FINAL DATA SET OBTAINED

We obtained 14 (33%) trial datasets from the original 42 trials we approached. A further five trials were obtained through snowballing, resulting in a total of 19 datasets (see

Figure 4). We were unsuccessful in getting a response from 15(36%) investigators and a further six (14%) datasets were not available for data sharing. We still have seven (17%) datasets in negotiation where we were unable to agree on the data sharing before starting our formal analysis, therefore these trials have not been included in this report.

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

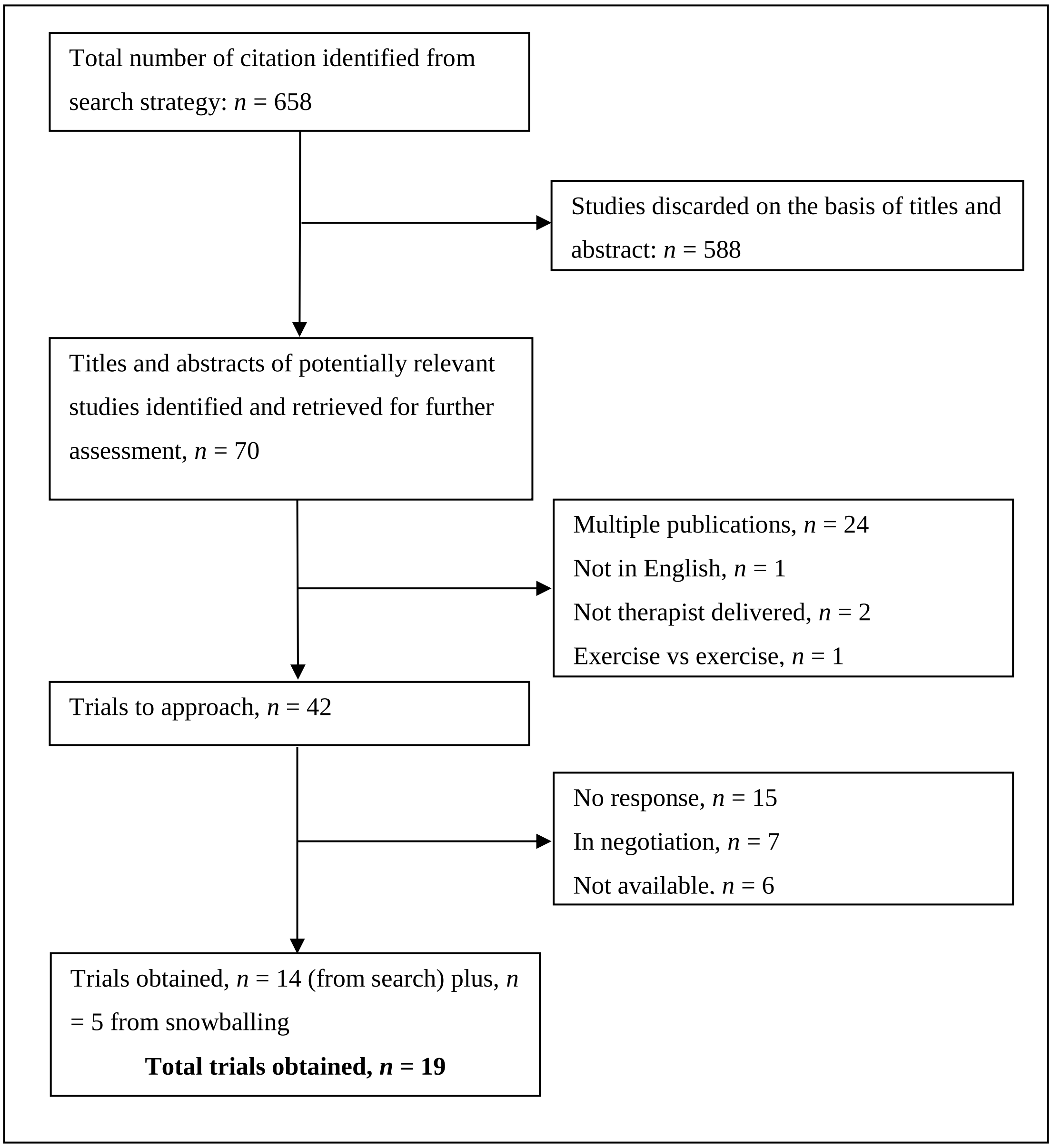

Total number of citation identified from search strategy: *n* = 658
Studies discarded on the basis of titles and abstract: *n* = 588
Titles and abstracts of potentially relevant studies identified and retrieved for further assessment, *n* = 70
Multiple publications, *n* = 24
Not in English, *n* = 1
Not therapist delivered, *n* = 2
Exercise vs exercise, *n* = 1
Trials to approach, *n* = 42
No response, *n* = 15
In negotiation, *n* = 7
Not available. *n* = 6
Trials obtained, *n* = 14 (from search) plus, *n* = 5 from snowballing
Total trials obtained, *n* = 19

Figure 4 Quorum statement flow diagram for database identification

Table 7 shows the trials that were excluded and the reason for the exclusion. Details of papers excluded due to multiple publications can be found in *Appendix 6*. A list of trials that were unavailable due to a lack of response for the investigator, datasets 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 P et al[97]	314	Not therapist delivered
Kainz B et al[98]	1,274	Paper not in English
Long A et al[99]	312	Trial of exercise vs exercise
Von Korff M et al[100]	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 et al. Pragmatic randomized trial evaluating the clinical and economic effectiveness of acupuncture for chronic low back pain. *Am J Epidemiol* 2006;164(5):487-96.	3,093
UK BEAM	Underwood	Underwood MR, Harding G, Klaber Moffett J. Patient perceptions of physical therapy within a trial for back pain treatments (UK BEAM) [ISRCTN32683578]. *Rheumatology* (Oxford) 2006;45(6):751-6. Underwood MR, Morton V, Farrin A. Do baseline characteristics predict response to treatment for low back pain? Secondary analysis of the UK BEAM dataset [ISRCTN32683578]. *Rheumatology* (Oxford) 2007;46(8):1297-302.	1,334
Haake	Haake	Haake M, Müller HH, Schade-Brittinger C, Basler HD, Schäfer H, Maier C et al. Acupuncture Trials (GERAC) for chronic low back pain: randomized, multicenter, blinded, parallel-group trial with 3 groups. *Arch Intern Med* 2007;167(17):1892-8.	1,163

Name of / given name of trial	Corresponding author/ Chief Investigator	Relevant publications related to the trial of interest	Number of participants
BeST	Lamb	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(9718):916-23. 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):1-253, iii-iv.	701
Keele	Hay	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(9476):2024-30. 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 randomized clinical trial. *Arthritis Rheum* 2007;57(3):466-73.	402
Brinkhaus	Brinkhaus	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 Intern Med* 2006;166(4):450-7.	298

Name of / given name of trial	Corresponding author/ Chief Investigator	Relevant publications related to the trial of interest	Number of participants
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. *Spine* (Phila Pa 1976) 2010;35(5):469-76.	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. *Ann Intern Med* 2007;146(11):787-96. 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(9):1202-9.	260

Name of / given name of trial	**Corresponding author/ Chief Investigator**	**Relevant publications related to the trial of interest**	**Number of participants**
YACBAC	Thomas	Thomas KJ, MacPherson H, Thorpe L, Brazier J, Fitter M, Campbell MJ, Roman M, Walters SJ, Nicholl J. Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain. *BMJ* 2006;333(7569):623. 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(7569):626. 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):iii-iv, ix-x, 1-109.	241

Name of / given name of trial	Corresponding author/ Chief Investigator	Relevant publications related to the trial of interest	Number of participants
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. *Eur Spine J* 2008;17(7):936-43. 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(1):51-5. 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(9599):1638-43.	240
Von Korff BIA	Von Korff	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(3):323-30.	240

Name of / given name of trial	Corresponding author/ Chief Investigator	Relevant publications related to the trial of interest	Number of participants
HullExPro	Carr	Carr JL, Klaber MJA, 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. *Disability and Rehabilitation* 2005;27(16):929-37.	237
Von Korff SC2	Moore	Moore JE, Von Korff M, Cherkin D, Saunders K, Lorig K. A randomized trial of a cognitive-behavioral program for enhancing back pain self care in a primary care setting. *Pain* 2000;88(2):145-53.	226
Smeets	Smeets	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 [ISRCTN22714229]. *BMC Musculoskelet Disord* 2006;7:5.	223
Cecchi	Cecchi	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(1):26-36.	210
York BP	Torgerson	Moffett JK, Torgerson D, Bell-Syer S, Jackson D, Llewlyn-Phillips H, Farrin A et al. Randomised controlled trial of exercise for low back pain: clinical outcomes, costs, and preferences. *BMJ* 1999 31;319(7205):279-83.	187

Name of / given name of trial	Corresponding author/ Chief Investigator	Relevant publications related to the trial of interest	Number of participants
Macedo	Macedo	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(3):363-77.	172
Carlsson	Carlsson	Carlsson CP, Sjölund BH. Acupuncture for chronic low back pain: a randomized placebo-controlled study with long-term follow-up. *Clin J Pain* 2001;17(4):296-305.	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. *Complement Ther Med* 2008;16(3):139-46.	48

3.6 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, n = 3,093[50]
Country	Germany
Interventions	In the RCT part of study there were two arms: • 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 where 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 neurologic symptoms; previous back surgery; infectious spondylopathy; low back pain 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, n = 1,334[45, 101]
Country	United Kingdom
Interventions	- Exercise programme – group exercise including cognitive behavioural principles delivered over up to eight 60 minute sessions over four to eight 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 provided over 12 weeks. - Combined treatment – provision of eight session of manipulation over six weeks plus eight sessions of exercise over the next six 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 The Back Book.
Recruitment	Recruited from GP practices after searching computerise records for potential eligible participants.
Inclusion criteria	Aged between 18 and 65 years; consulted with LBP; score of four or more 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 or over, potential spinal disorder, including malignancy, osteoporosis, ankylosing spondylitis, cauda equine 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 100m when free of back pain, inability to get up off the floor unaided, receipt of physical therapy in the preceding three months, RMDQ score of ≤3 on the day of randomisation, inability to read and write English fluently.

Name of/given name of trial	**Haake, *n* = 1,163[102]**
Country	Germany
Interventions	All groups received ten 30-minute sessions (2 per week). Five additional sessions were offered if after the tenth session patients experienced a 10% to 50% reduction in pain intensity (Von Korff Chronic Pain Grade Scale) • Verum acupuncture – Sterile disposable needles used to needle fixed points plus additional points from a pre-specified list. 14-20 needles used and manual stimulation to elicit *de Qi* • Sham acupuncture – Number of and type of needles were the same as verum acupuncture. Needling verum points or meridians avoided and needles were inserted superficially and without stimulation. • Conventional therapy – this was a multimodal treatment programme where patients received ten 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	Aged ≥18 years with a clinical diagnosis of CLBP of six months or longer, no previous experience of acupuncture for LBP. Grade one or higher for mean Von Korff Chronic Pain and a Hanover Functional Ability Questionnaire score of less than 70%.
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, *n* = 701[31, 36]**
Country	United Kingdom
Interventions	- Intervention arm received an initial 15 minute advice session and were provided with The Back Book. Subsequently they attended six 1.5 hour group sessions which covered cognitive behavioural topics - Control arm - 15 minute advice session and provided with The Back Book.
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 sub-acute or chronic low back pain, with a minimum of six weeks' duration, consultation with the GP for low-back pain within the preceding six months.
Exclusion criteria	Low back pain 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 low-back pain.

Name of/given name of trial	Keele, *n* = 402[65, 103]
Country	United Kingdom
Interventions	- Brief pain management program – 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 done 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 for the first or second time of less than 12 weeks' duration, able to give informed consent.
Exclusion criteria	Those with signs of red flags, sick leave or >12 weeks, diagnosed with osteoporosis or inflammatory arthritis, taking systemic steroids for longer than 12 weeks, pregnant, previous fracture or hip/back surgery, any abdominal surgery in the preceding three months, receipt of treatment by any other professional for the current episode of back pain.

Name of/given name of trial	**Brinkhaus, *n* = 298**[104]
Country	Germany
Interventions	The acupuncture and minimal acupuncture treatments consisted of 12, 30 min sessions delivered over 8 weeks. • Acupuncture treatment – this was semi-standardised. Single use sterile disposable needles were used. Physicians were instructed to achieve *de qi* (an irradiating feeling), if possible. Manual stimulation of needles at least once during each session. • Minimal acupuncture – therapist were advised to needle at least six of ten predefined non-acupuncture points using a superficial insertion with fine needles. None of the points were in the area of the lower back. *De qi* and manual stimulation of the needles were avoided. • Waiting list 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 snowballing from that.
Inclusion criteria	Aged between 40-75 years with a clinical diagnosis of chronic low back pain present for more than six months, a VAS of ≥40 for average pain intensity over the previous seven days and the use of only oral NSAIDs in the preceding four weeks before treatment.
Exclusion criteria	Disc prolapse/protrusion of 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, n = 286[105]
Country	Denmark
Interventions	- Multidisciplinary biopsychooscial rehabilitation – 12 week programme split into three periods of four weeks. Period 1 - exercise was performed 3 times a week in 2-hour sessions. Exercise comprised of 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 were provide 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 a week at home or a fitness centre. Period 3 – three times a 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 one hour twice a week for 12 weeks. The therapist initially supported the patient then over time reduced the amount of assistance.
Recruitment	Rheumatologists and GPs referred patients.
Inclusion criteria	Patients aged 18-60 with LBP of more than 12 weeks with or without pain radiating into the leg(s). The lumber 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.

Name of/given name of trial	Pengel, *n* = 260[59, 106]
Country	Australia
Interventions	- 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 they 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 their 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 waiting list for physiotherapy and advert in newspaper.
Inclusion criteria	Those aged 18-80 years, NSLBP lasting for at least six 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, *n* = 241**[88, 107]
Country	United Kingdom
Interventions	• Traditional acupuncture – up to ten session over three months • Usual care – this group received treatment as usual determined by the GP
Recruitment	Recruited from GP practices.
Inclusion criteria	18-65 with non-specific low back pain 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, *n* = 240**[108-110]
Country	Australia
Interventions	• Spinal manipulation - Patients in this arm received two to three session 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 initial session followed by 20 minute follow-up sessions). Four arms in the trial: • spinal manipulative therapy group (placebo drug and active spinal manipulative therapy); • spinal manipulative therapy and NSAIDs group (diclofenac and active spinal manipulative therapy). • NSAIDs group (diclofenac and placebo spinal manipulation); • Control group (placebo drug and placebo spinal manipulative therapy)
Recruitment	Recruited from GP practices
Inclusion criteria	Pain present in the region between the 12th 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 spinal manipulative therapy.

Name of/given name of trial	**Von Korff BIA, *n* = 240**[111]
Country	United States of America
Interventions	• 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 physiotherapists, the third 30 minutes with a physiotherapist and the final visit 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 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 that had consulted in primary care for their back pain who were enrolled in the group Health Cooperative.
Inclusion criteria	Patients with back pain aged 25-65 years, those with an RMDQ of seven or more on a 23 item scale
Exclusion criteria	Those waiting for back surgery, seeing a physical therapist or psychologist, patients planning to unenrolled from the Group Health Cooperative.

Name of/given name of trial	**HullExPro, *n* = 237**[76]
Country	United Kingdom
Interventions	- Back to fitness exercise programme – patients were invited to attend eight one hour sessions aimed at increasing activity over a four 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 low back pain lasting at least six 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 six weeks prior.

Name of/given name of trial	**Von Korff SC2, *n* = 226**[112]
Country	United States of America
Interventions	• Self-care arm – this was a group intervention of between 12-16 patients delivered over two, two 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-min videotape on back pain self-care and a 25-min videotape demonstrating exercises. • Usual care group - received usual care plus a book on back pain.
Recruitment	Patients were recruited from primary care by mail six to eight weeks after a back pain visit to a Group Health primary care physician.
Inclusion criteria	Patients with back pain, aged 25-70 years, patients that had been enrolled into Group Health for at least one year.
Exclusion criteria	Those being considered for surgery.

Name of/given name of trial	**Smeets, *n* = 223[70]**
Country	The Netherlands
Interventions	• Active physical treatment – this consisted of aerobic and strength training. This was delivered by two physiotherapists in a maximum group of four. Sessions were delivered three times a week lasting one hour and 45 minutes. • Cognitive behavioural treatment – this aimed to help patients reach their goals, manage beliefs and increase activity levels. Therapists used graded activity and problem solving training. • Active Physical Therapy (APT) – aimed at increasing aerobic capacity and muscle conditioning. • Cognitive behavioural treatment (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 to one session a week. In total 11 1/2 hours of treatment • Combined treatment (CT) – aim was to improve functioning by increasing fitness, behaviour change and management of beliefs. CT consisted of APT together with problem solving training. • Waiting list – Patients needed to wait 10 weeks before they were offered individual rehabilitation treatment. Whilst on the waiting list 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 more than three months with or without radiation to leg, an RMDQ score of > 3 and ability to walk at least 100 meters without interruption.
Exclusion criteria	Vertebral fracture, spinal inflammatory disease, spinal infections or malignancy, current nerve root pathology, spondylolysis or spondylolisthesis, lumbar spondylodesis. A co-morbidity 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, pregnancy and substance abuse.

Name of/given name of trial	**Cecchi, *n* = 210**[113]
Country	Italy
Interventions	All patients got an educational booklet on the back • Back school – 15 one hour sessions delivered over 15 days. The first five sessions focused on back physiology and pathology. Remaining ten 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 four to six weeks.
Recruitment	Rehabilitation out patients department by psychiatrists.
Inclusion criteria	NSLBP over at least the last six month reported as present 'often' or 'always.'
Exclusion criteria	Neurological signs or symptoms, spondylolisthesis, spinal stenosis, scoliosis >20 degrees, rheumatoid arthritis/spondylitis, previous vertebral fracture, psychiatric condition, cognitive impairment or pain related litigation.

Name of/given name of trial	**York BP, *n* = 187[114]**
Country	United Kingdom
Interventions	• Exercise programme – delivered as a group intervention over eight one hour session over a four week period. The sessions comprised of 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 form their GP.
Recruitment	Recruited from GP practices.
Inclusion criteria	Patients aged between 18-60 years with LBP which has lasted at least four weeks but less than six 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, *n* = 172**[115]
Country	Australia
Interventions	In both arms patients received 12 one hour sessions over an eight week period. Home exercises were encouraged in both groups. The home exercises and treatment sessions totalled 20 hours. • Graded activity – The aim of graded activity was to get patients to engage in activities they found difficult due to back pain. Patients were provided with an individualised progressively increasing exercise programme to address functional problems. A cognitive behavioural approach was use by the physiotherapist. • Motor control exercise – the aim is to retain optimal control and coordination of the lumbar spine and pelvis. Stage one involves regaining basic control strategies. In stage 2 participants progressed 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 with NSLBP of at least three 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 grater 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 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, *n* = 50**[116]
Country	Sweden
Interventions	• Manual acupuncture – Needle acupuncture was used in predefined areas. There was a gradual increase in the number of needles from eight to 14 to 18 during the first three or four treatments. The de-qi feeling was sought. Treatment sessions lasted 20mins and needles were stimulated on three occasions during this time. • Electroacupunture – 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 transcutaneous electrical nerve stimulation (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 that were referred to an outpatient pain clinic during a three-year period were included.
Inclusion criteria	Patients with LBP without radiation below the knee for greater than six months, normal neurologic examination function of lumbosacral nerve.
Exclusion criteria	Those who have had previous acupuncture treatment, patients with major trauma or systemic disease and pregnancy.

Name of/given name of trial	**Kennedy, *n* = 48[117]**
Country	United Kingdom
Interventions	• Verum acupuncture plus The Back Book – acupuncture was based on a western approach. Between three and 12 session provided over a four to six week period. At each session eight to 13 needles were inserted and manually stimulated until de qi was achieved. • Sham acupuncture plus The Back Book – 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 waiting list 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 more than 12 weeks, those with a contra-indications to acupuncture or previous acupuncture treatment, any other conflicting or ongoing treatments.

3.7 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 broadly pool interventions into groups for our analyses. As a first stage we identified the control interventions and classified these as either usual care or as 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.[118] 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 where 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 developing 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 TENS, sham acupuncture and control). We listed all 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 (see *Figure 5*).
2) To explore further the potential direct and indirect comparisons a second figure was constructed (see *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 one and two 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) (see *Table 10*).

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. 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 though 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 worthwhile pursuing theses analyses. For this reason these were excluded from our final analyses.

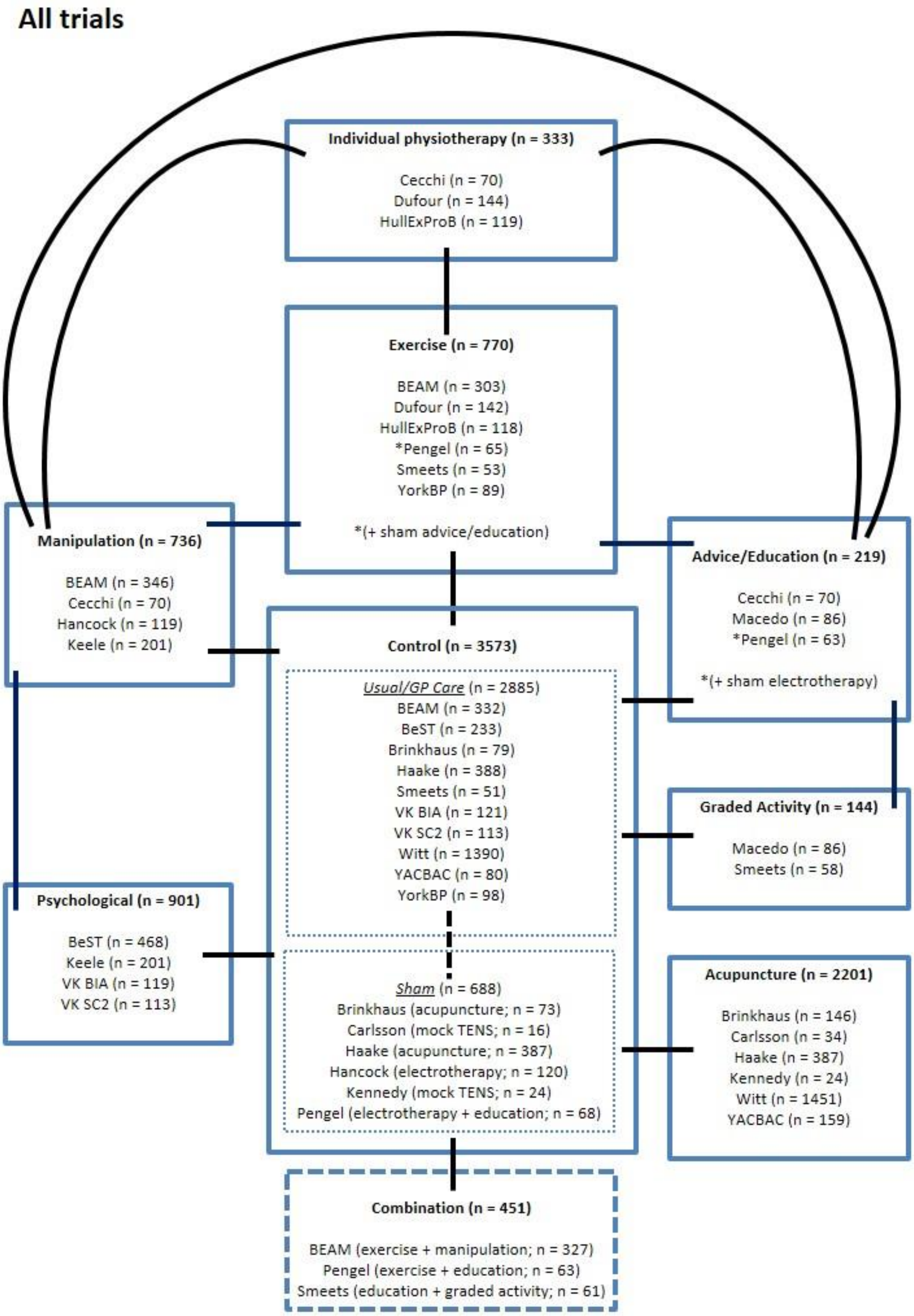
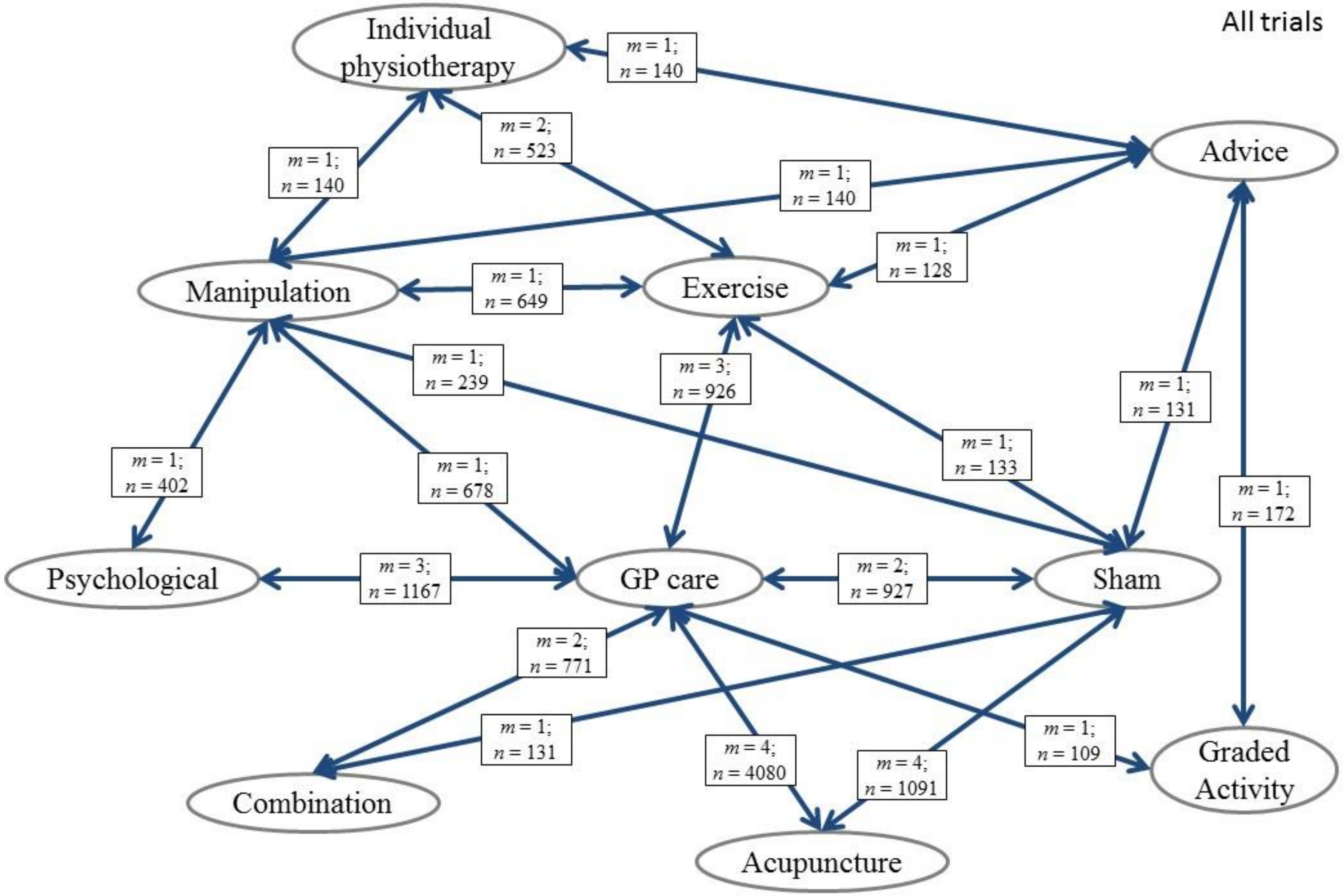


Figure 5 Step one – Classification of trials into core groups



m, Number of trials; *n*, total number of participants

Figure 6 Step two – Classification of trials with indication of number of trials and participants for direct and indirect comparisons

Table 10 Step 3 - Final grouping of treatment arms for analyses

Parent group	Subgroup	Sub-type
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 transcutaneous electrical nerve stimulation (TENS)
		Sham advice/education
Control (GP/usual care)		General practitioner (GP)
		Waiting list

CHAPTER 4 – CREATING THE REPOSITORY DATABASE AND DATA CONTROL

4.1 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 are printed in sans-serif font [like this] and the command for mapping and transformation procedures are printed in monospaced typewriter font `[like this]`. Also, coloured command fonts in the text are for ease of referencing between program commands shown in figures and text explanations.

4.2 BACKGROUND

Clinical trial datasets can be stored in a tabular format, for example, Microsoft Excel or SPSS. 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 if they were used or not. 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 many columns would have to be created for all participants.

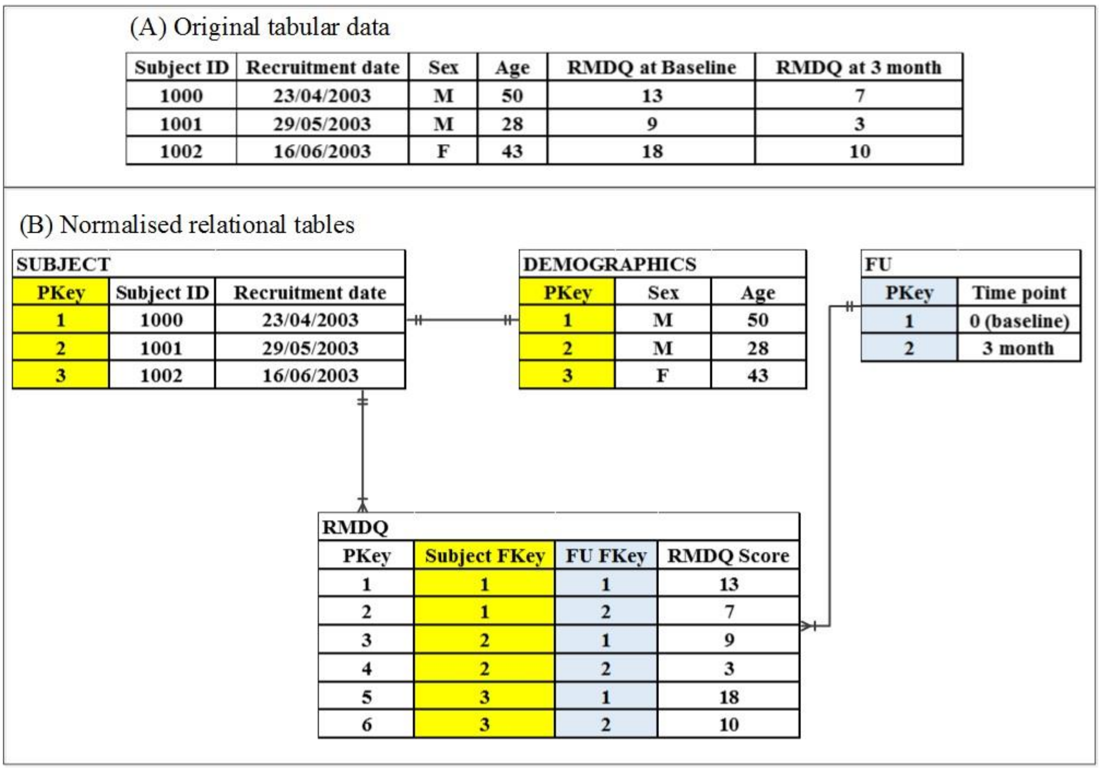
Tabular formats are only effective for the smallest of trials and quickly become inefficient and difficult to maintain when the range of data collected increases. For larger trials a more robust solution is to use a relational database. The relational database model allows individual tables to be created 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 used to create table joins.[119]

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 Roland Morris Disability Questionnaire (RMDQ) scores taken at baseline and at three-month follow-up. The data is normalised into four tables, namely, SUBJECT, DEMOGRAPHICS, RMDQ (for the RMDQ measurement) and FU. The latter 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 7(B)*). The relationship between SUBJECT and DEMOGRAPHICS is one-to-zero-or-one, that is, a subject can have zero or one demographic record. The primary key from the SUBJECT table is copied to the DEMOGRAPHICS tables thereby creating a join using a shared value.

The relationships 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 three-month follow-up time point.

To create the relationships to the RMDQ table, the primary keys from both the SUBJECT and FU tables are added as foreign keys. 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.


(A) Original tabular data

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) Normalised relational tables

SUBJECT

PKey	Subject ID	Recruitment date
1	1000	23/04/2003
2	1001	29/05/2003
3	1002	16/06/2003

DEMOGRAPHICS

PKey	Sex	Age
1	M	50
2	M	28
3	F	43

FU

PKey	Time point
1	0 (baseline)
2	3 month

RMDQ

PKey	Subject FKey	FU FKey	RMDQ Score
1	1	1	13
2	1	2	7
3	2	1	9
4	2	2	3
5	3	1	18
6	3	2	10

Figure 7 (A) A sample of original tabular format data. (B) Normalised relational interpretation of the original tabular data.

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 IT 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 that allows dynamic changes whilst enforcing data integrity.

Section 4.3 describes the architecture of the hybrid database. *Section 4.4* describes the rules used to map and transform the original source data to the repository standard. *Section 4.5* 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). *Section 4.6* describes how data from multiple RCTs were extracted, transformed and harmonised to the repository standard and finally, loaded to the repository database.

4.3 SYSTEM ARCHITECTURE

Tables and columns in a relational database can be represented as classes and attributes in an EAV model.[120] 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 only storing a pointer to the data and not the actual data itself. The entity-relationship diagram for the hybrid database is shown in *Figure 8*.

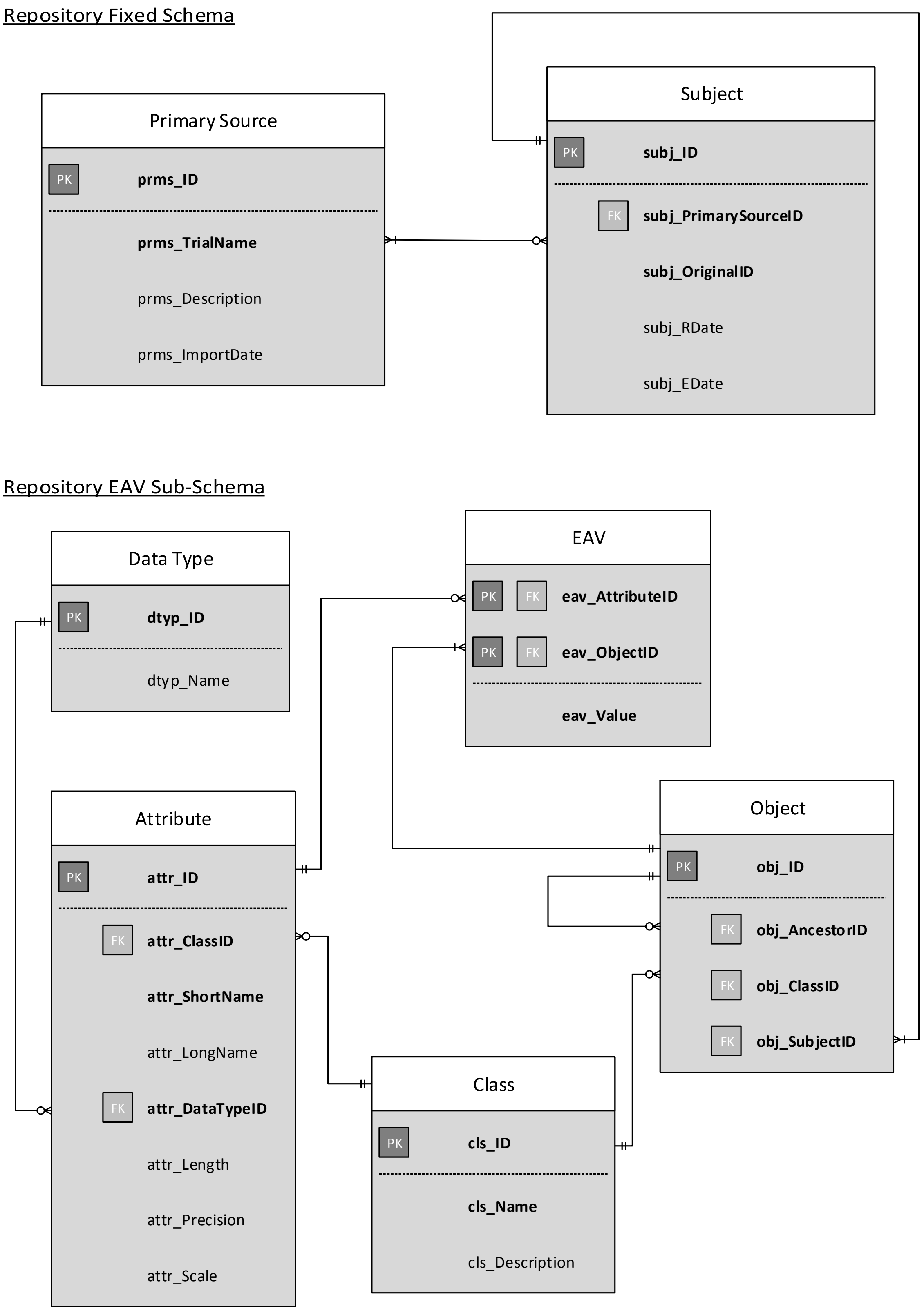


Figure 8 The entity-relationship diagram for the hybrid repository database depicting the fixed schema with the sub-schema entity-attribute-value (EAV) tables.

We anticipated that there would be some consistent data present in all 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 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.

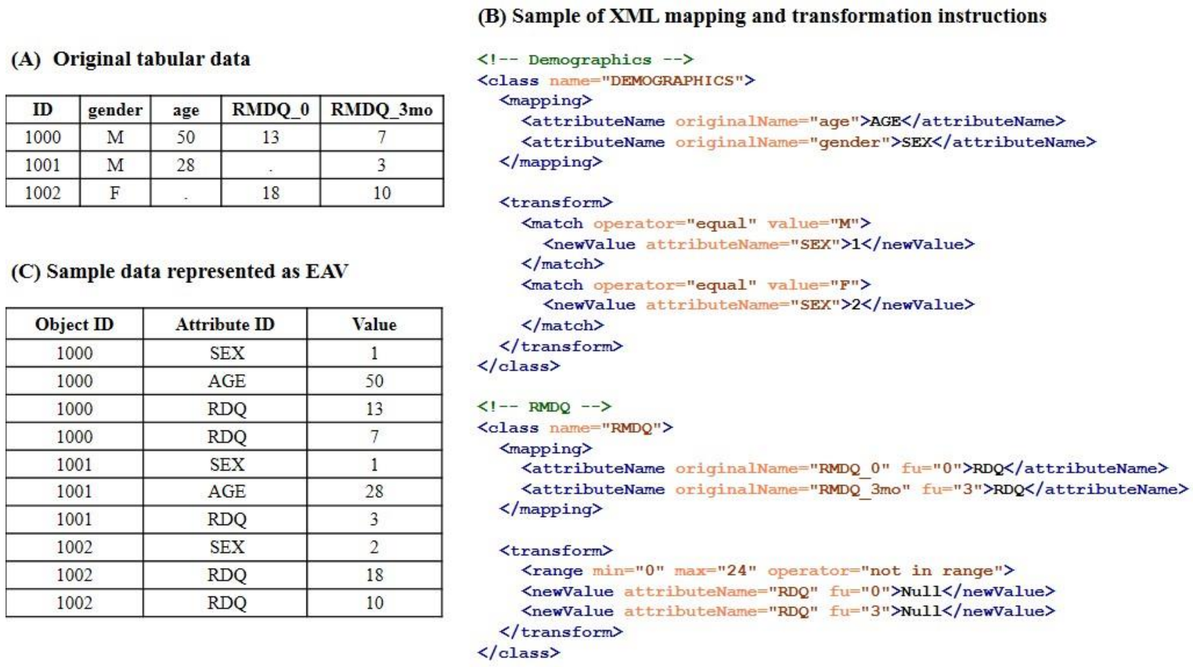
The EAV model uses a sub-schema consisting of tables for classes, attributes, objects and the EAV data. The Class table is used to hold a list of all the identified domains, for example, Roland Morris disability questionnaire, Demographics, etc. These domains generally map to a CRF but can also be used to describe a sub-set 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 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 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 is represented in an EAV table is shown in *Figure 9*. In this example, the tabular data has one row for each subject (see *Figure 9(A)*). When the data is 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 9(C)*).



(A) Original tabular data

ID	gender	age	RMDQ_0	RMDQ_3mo
1000	M	50	13	7
1001	M	28	.	3
1002	F	.	18	10

(C) Sample data represented as EAV

Object ID	Attribute ID	Value
1000	SEX	1
1000	AGE	50
1000	RDQ	13
1000	RDQ	7
1001	SEX	1
1001	AGE	28
1001	RDQ	3
1002	SEX	2
1002	RDQ	18
1002	RDQ	10

(B) Sample of XML mapping and transformation instructions

```xml
<!-- Demographics -->
<class name="DEMOGRAPHICS">
  <mapping>
    <attributeName originalName="age">AGE</attributeName>
    <attributeName originalName="gender">SEX</attributeName>
  </mapping>

  <transform>
    <match operator="equal" value="M">
      <newValue attributeName="SEX">1</newValue>
    </match>
    <match operator="equal" value="F">
      <newValue attributeName="SEX">2</newValue>
    </match>
  </transform>
</class>

<!-- RMDQ -->
<class name="RMDQ">
  <mapping>
    <attributeName originalName="RMDQ_0" fu="0">RDQ</attributeName>
    <attributeName originalName="RMDQ_3mo" fu="3">RDQ</attributeName>
  </mapping>

  <transform>
    <range min="0" max="24" operator="not in range">
    <newValue attributeName="RDQ" fu="0">Null</newValue>
    <newValue attributeName="RDQ" fu="3">Null</newValue>
  </transform>
</class>
```

Figure 9 (A) A sample of original tabular format clinical data. (B) The XML mapping and transformation instructions. (C) The sample data represented as EAV.

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 9(C)*. Also, 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.

4.4 MAPPING AND TRANSFORMATION

Early evaluation of datasets from various RCTs in the project identified large variations between variable naming and coding conventions. For example, the Roland Morris Disability Questionnaire (RMDQ) was used to measure back pain disability and participant would tick all the items that were applicable to them on that day. There are 24 items in the questionnaire and the score is the sum of all 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, ‘rdq11fu’, ‘rdq21fu’, …, ‘rdq241fu’ and ‘rdq1fu’ for items measured at the first follow-up which could have been one month or three months post randomisation depending on the protocol. In addition, some trials might use 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.

4.4.1 PILOT MAPPING AND TRANSFORMATION

A system was required to efficiently extract, transform and load (ETL) the original trial datasets 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) 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 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.

4.4.2 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.[121] 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 onto a programmer which in turn saves resources and decreases misinterpretation errors.

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

4.4.3 MAPPING CLINICAL DATA

Figure 9(B) 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 three-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 which time point an original variable belongs to.

4.4.4 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 9(A)*), 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.

Table 11 A sample of the repository standard attributes and values.

Class	Attribute short name	Attribute long name	Data type	Value	Label
DEMO-GRAPHICS	SEX	Participant's sex	Integer	1	Male
				2	Female
DEMO-GRAPHICS	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
			String	3a	Physiotherapist
			String	4M01	Non-steroidal anti-inflammatory drugs
				6	Aids and adaptations
HE	REASON	Resource reason	Integer	2	Low back pain
				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

Class	Attribute short name	Attribute long name	Data type	Value	Label
				3	Prescription
				4	Item
HE	QUANTITY		Integer	> 0	
HE	COST		Integer	> 0	
HE	PAYER	Resource payer	Integer	1	Public health service
				4	Individual

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

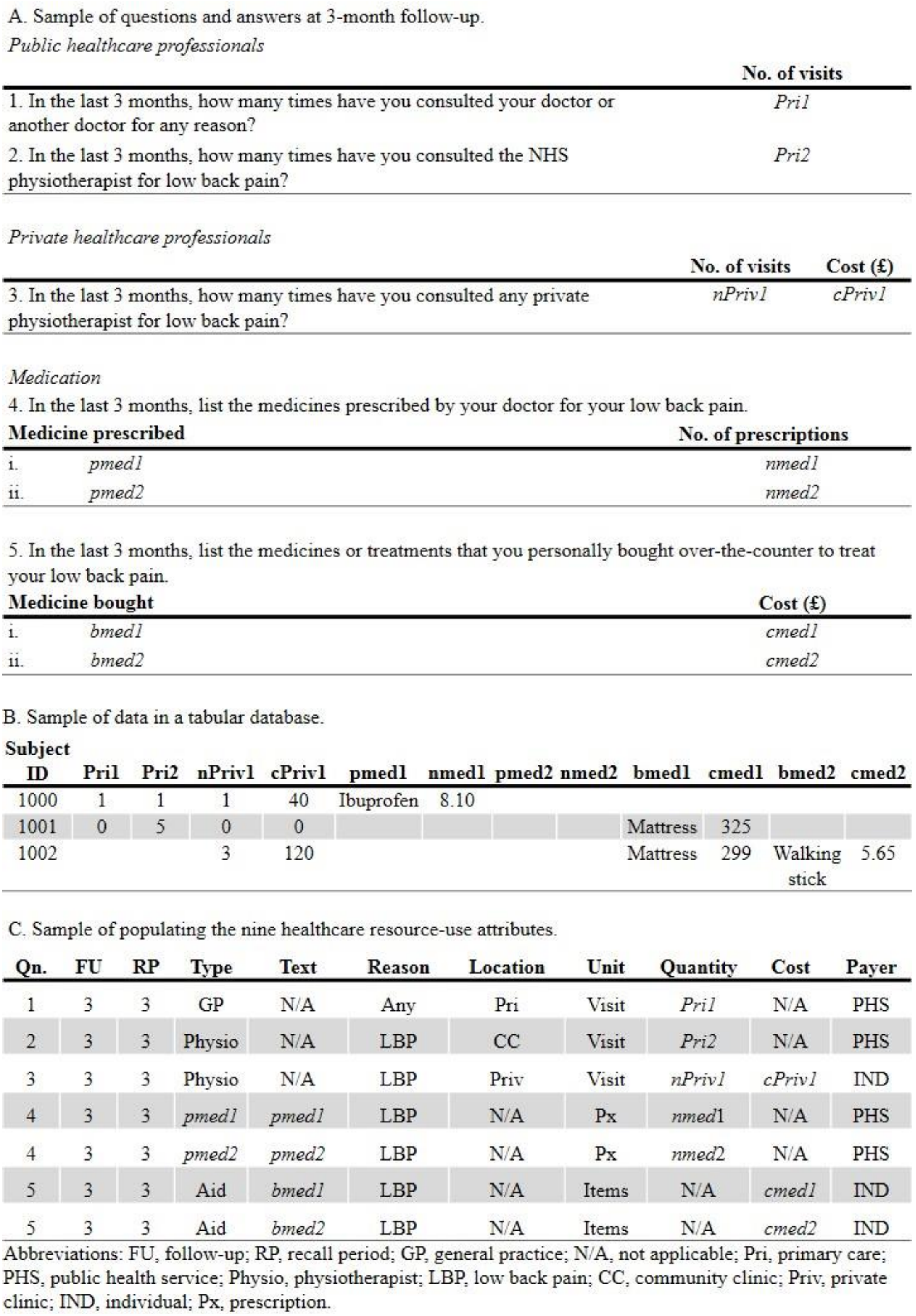
4.4.5 MAPPING AND TRANSFORMING HEALTHCARE RESOURCE-USE DATA

Mapping healthcare 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 healthcare resource-use questionnaires 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.

Figure 10 shows a simplified version of a typical healthcare resource-use questionnaire. In this example participants were asked to record all the healthcare resources they used at the three-month follow-up time point (see *Figure 10(A)*). The answers provided by the participants were stored in a tabular format that used 12 columns to capture all responses to the five questions (see *Figure 10(B)*). 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 the bought over-the-counter to treat their LBP, the number of columns required would have to be increased from 12 to 13.

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

The process for creating the transformed healthcare resource-use data involves splitting the original questions into a number of derived parts that will map to the standard attributes. For example, question one asked how many times the participant had consulted their doctor or any primary care doctor for any reason in the last three months. From 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 is ‘Primary Care Setting’, the unit of measurement is ‘Visit’, the payer is ‘Public Health Service’. All these values are derived solely on 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.



A. Sample of questions and answers at 3-month follow-up.

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?	*Pri1*
2. In the last 3 months, how many times have you consulted the NHS physiotherapist for low back pain?	*Pri2*

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?	*nPriv1*	*cPriv1*

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.	*pmed1*	*nmed1*
ii.	*pmed2*	*nmed2*

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.	*bmed1*	*cmed1*
ii.	*bmed2*	*cmed2*

B. Sample of data in a tabular database.

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. Sample of populating the nine healthcare resource-use attributes.

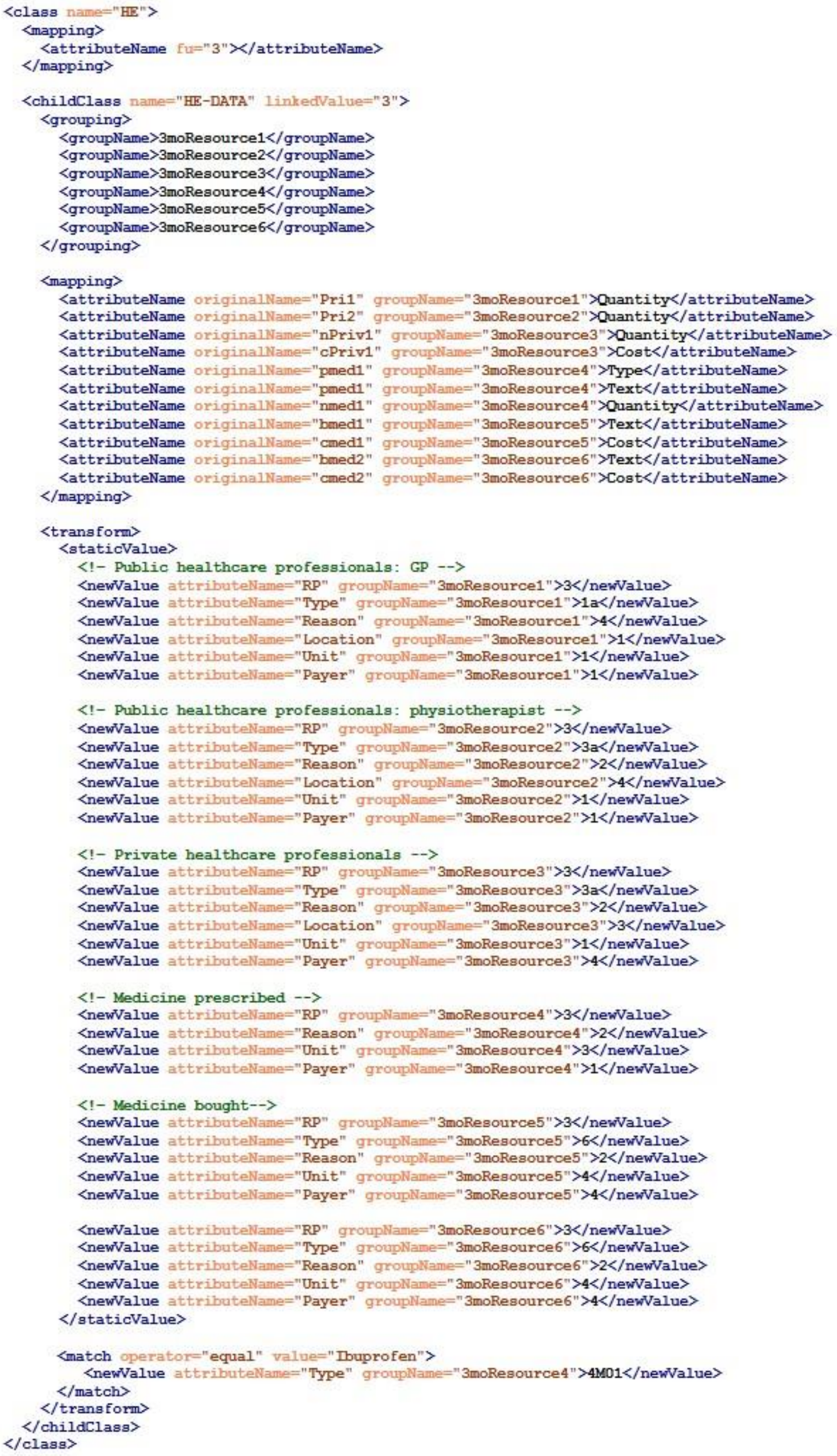
Qn.	FU	RP	Type	Text	Reason	Location	Unit	Quantity	Cost	Payer
1	3	3	GP	N/A	Any	Pri	Visit	*Pri1*	N/A	PHS
2	3	3	Physio	N/A	LBP	CC	Visit	*Pri2*	N/A	PHS
3	3	3	Physio	N/A	LBP	Priv	Visit	*nPriv1*	*cPriv1*	IND
4	3	3	*pmed1*	*pmed1*	LBP	N/A	Px	*nmed1*	N/A	PHS
4	3	3	*pmed2*	*pmed2*	LBP	N/A	Px	*nmed2*	N/A	PHS
5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

Abbreviations: FU, follow-up; RP, recall period; GP, general practice; N/A, not applicable; Pri, primary care; PHS, public health service; Physio, physiotherapist; LBP, low back pain; CC, community clinic; Priv, private clinic; IND, individual; Px, prescription.

Figure 10 (A) A sample of questions in a case report form at 3-month follow-up. (B) A sample of original tabular format healthcare resource-use data. (C) A sample of how the healthcare resource-use data populate the repository standard.

The healthcare resource-use data is 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 only used to define the time points for collecting the healthcare resource-use data. The actual resource-use data is 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 describe, 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 classed. 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 three-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 three-month follow-up time point specified in the HE class.



```xml
<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*.

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 three-month resource-use data, namely, 3moResource1, 3moResource2, 3moResource3, 3moResource4, 3moResource5, and 3moResource6. These groups represent each question in the CRF shown in *Figure 10(A)* and the data shown in *Figure 10(B)*.

The original tabular data required 13 columns across three rows to store all 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 only create objects where data exists. For example, subject #1000 will only create 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 require 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 based on the CRF and does not affect the original data. When the system processes this mapping instruction subject #1000 would have a healthcare resource-use object that show there was one GP visit made during the three-month follow-up time point (see *Figure 10(B)*).

Other `<transform/>` rules can be applied to manipulate the original healthcare 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 healthcare 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>
```

4.5 USING EAV DATA

Using the EAV/CR data in its raw state for any kind of analysis work would be extremely difficult due to the fragmented nature of the EAV schema. For analysis purposes, it is therefore necessary to piece together the data to form complete datasets that are comparable to the datasets 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 dataset 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 increase and multiple pivot operations were used for transforming object relationships. Querying the derived datasets directly was also impractical because of the huge amounts 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.

4.6 EXTRACT, TRANSFORM AND LOAD (ETL)

The bespoke extract, transform and load (ETL) application was required to read the original source data, automatically apply mapping and transformation rules from an XML document and to load the processed data into the repository. In addition to these basic functions, the ETL application was also required to permit end users to setup 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 Windows desktop application. It works by first uploading the original dataset and the XML mapping and transformation rules. The instructions defined in the XML file are applied to the original dataset and the transformed data is 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 datasets into the live database.

4.7 DATA VALIDATION

Data integrity is vital throughout the repository ETL process. To check that the mapping and transformation procedures were done correctly, the repository data was routinely checked against the original datasets. 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 inconsistency were flagged and if required, the XML instructions were amended. This process was repeated until the data was deemed to have been transformed correctly.

4.8 STORAGE

In condition of our data sharing agreements to hold the RCT datasets and to meet local governance and standard operating procedures the repository database server is held in a secure data centre with a 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.

4.9 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 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 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 email 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:
    a. Agreed to sharing the data but has asked to see a copy of the request, a copy will be sent to them via email for information purposes only.
    b. Not agreed to sharing their data, this dataset will be removed from the pooled data before provided 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 that measure the same domain, to a common scale (crosswalking). This work has now been published in Spine.[122] 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 standard deviation. 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*).

5.1 BACKGROUND

There are six participant-reported outcomes measures (PROMs) that have been used in one or more study within the repository that aim to measure back pain related disability, namely, Chronic Pain Grade (CPG) disability score which is one of the two domains in the CPG that aims to grade chronic pain status,[123] Hannover Functional Ability Questionnaire (FFbHR),[124] Oswestry disability index (ODI),[125] Pain Disability Index (PDI),[126] Patient Specific Functional Scale (PSFS)[127] and Roland Morris Disability Questionnaire (RMDQ).[28] Some trials also included generic health-related quality of life instruments such as SF-12[128] or SF-36[129] where the physical component summary (PCS) measures the physical functioning. As mentioned in *Section 6.3.3*, no common instrument was used by the trials 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 data pooling was feasible. After we had completed this work a National Institute for Health taskforce identified developing crosswalking values for 'legacy' measures of back pain outcome as a key priority for back pain research.[130]

5.2 DATA

We used data from 11 trials which 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 between two and three 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 *Section 6.3.2*. 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 standard deviation of the change. We used this in preference to the standardised effect size (change score divided by the standard deviation of the measure at baseline) so that all the standardised scores had a standard deviation of one. This enables visual comparisons to be made between all the scatterplots.

5.3 OUTCOME CONVERSION

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

$$y = \alpha + \beta x + \varepsilon \qquad (1)$$

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, where the latter is explained below.[131]

5.4 CORRELATION

Correlation was assessed by scatterplots and Pearson’s correlation coefficient with a correlation coefficient considered at least moderately high if it was greater than 0.5.

5.5 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 which could be falsely attributed to other sources, such as the treatment effect.

Similarity of responsiveness of two outcome measures was examined by categorising the change scores as negative change (change score < 0), no change (change score = 0) or positive change (change score > 0), and applying Cohen's kappa to these categorisations.[132] We considered $\kappa > 0.4$ to indicate sufficiently similar responsiveness.[133] 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 greater than 0.5) and to have at least moderately similar responsiveness (Cohen's kappa greater than 0.4).

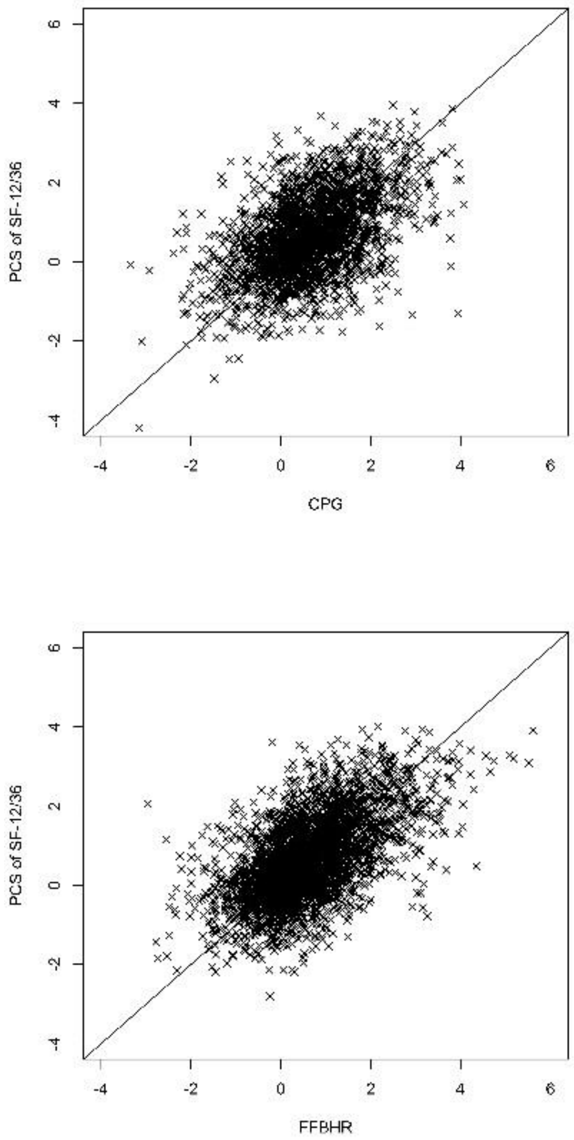
5.6 RESULTS

Eleven trials (n = 6,089) and seven instruments were included in these analyses (see *Table 12*). There were a total of 21 within trial pairwise comparisons between two outcomes. *Figure 12, Figure 13, Figure 14, Figure 15* and *Figure 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.

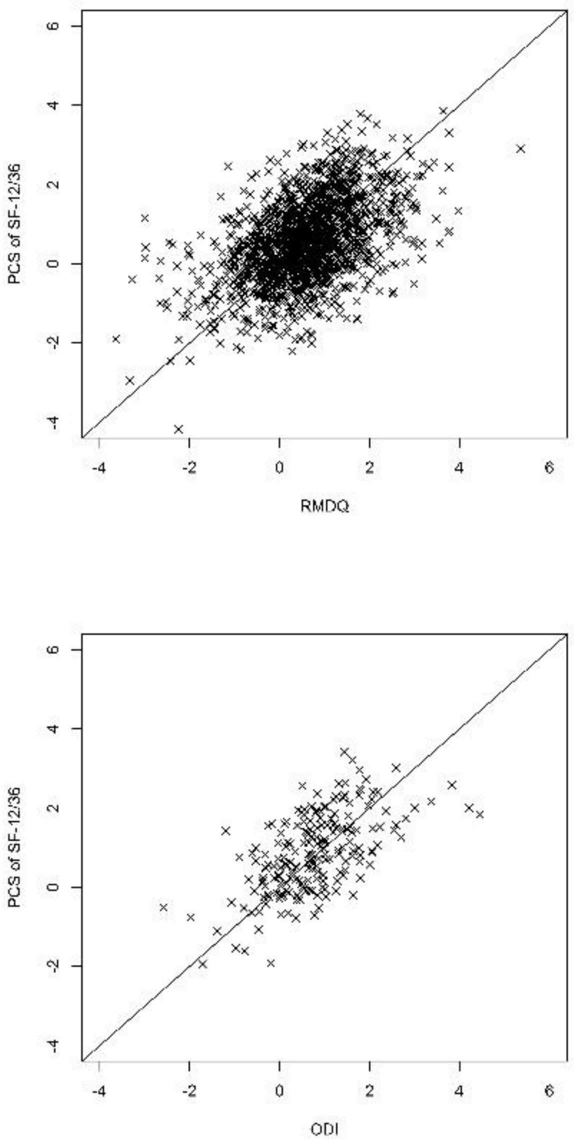
Trial	n	Outcome measures		
BeST[31]	426	RMDQ[a]	CPG[b]	PCS[c]
Brinkhaus[104]	281	PCS	FFbHR[d]	PDI[e]
Haake[102]	1,110	CPG	FFbHR	PCS
Hancock[109]	235	RMDQ	PSFS[f]	
HULLEXPRO[76]	203	RMDQ	PCS	
Macedo[115]	158	RMDQ	PCS	PSFS
Pengel[106]	232	RMDQ	PSFS	
UK BEAM[34]	885	RMDQ	CPG	PCS
VKBIA[111]	227	RMDQ	CPG	
Witt[50]	2,229	PCS	FFbHR	
YACBAC[134]	206	PCS	ODI[g]	

a RMDQ, Roland Morris disability questionnaire; b CPG, Chronic Pain Grade disability scale; c PCS, Physical Component Summary of SF-12 or SF-36; d FFbHR, Hannover Functional Ability Questionnaire; e PDI, Pain Disability Index; f PSFS, Patient Specific Functional Scale; and g ODI, Oswestry disability index.



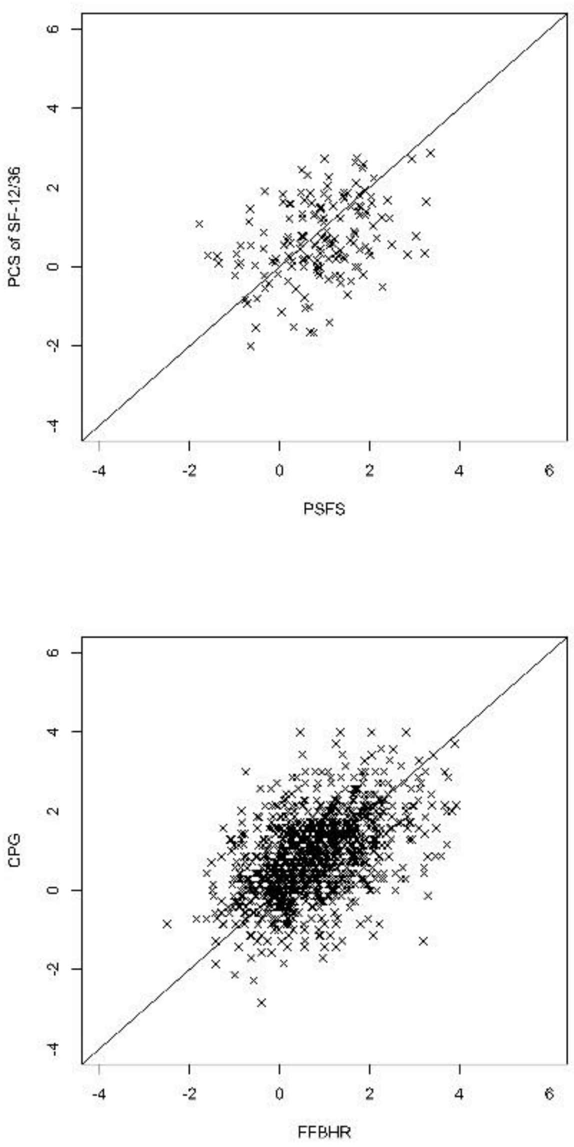
PCS, physical component scale of SF-12/36; CPG, chronic pain grade disability score; FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations.

Figure 12 Scatterplots of standardised change scores for PCS vs. CPG (*n* = 2451) and PCS vs. FFbHR (*n* = 3620) outcome measures.



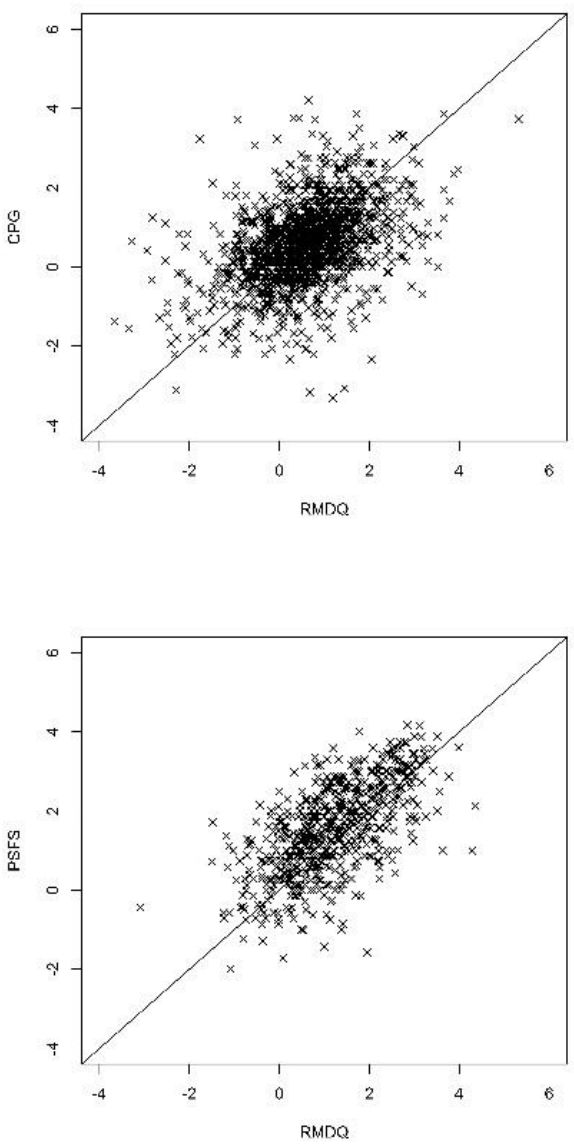
PCS, physical component scale of SF-12/36; RMDQ, Roland Morris disability questionnaire; ODI, Oswestry disability index.

Figure 13 Scatterplots of standardised change scores for PCS vs. RMDQ (*n* = 1694) and PCS vs. ODI (*n* = 206) outcome measures



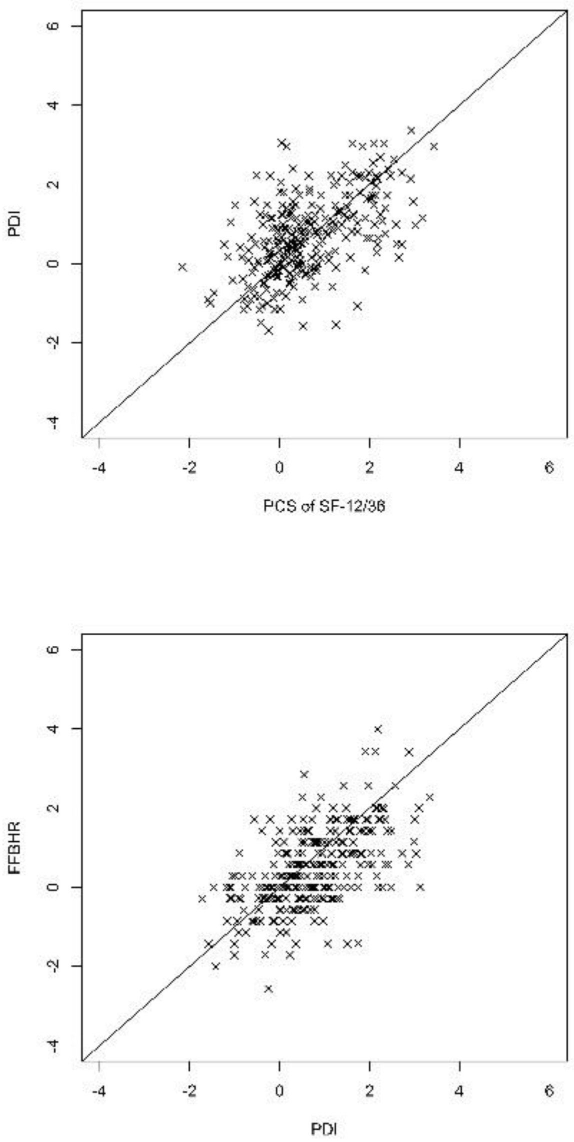
PCS, physical component scale of SF-12/36; CPG, chronic pain grade disability score; FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; PSFS, patient specific functional scale.

Figure 14 Scatterplots of standardised change scores for PCS vs. PSFS (*n* = 158) and CPG vs. FFbHR (*n* = 1110) outcome measures



CPG, chronic pain grade disability score; RMDQ, Roland Morris disability questionnaire; PSFS, patient specific functional scale.

Figure 15 Scatterplots of standardised change scores for CPG vs. RMDQ (*n* = 1661) and PSFS vs. RMDQ (*n* = 625) outcome measures



PCS, physical component scale of SF-12/36; FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; PDI, pain disability index.

Figure 16 Scatterplots of standardised change scores for PDI vs. PCS (*n* = 281), FFbHR vs. and PDI (*n* = 284) outcome measures

The correlations between outcomes ranged from 0.21 to 0.70; implying that the linear associations between them range from weak to moderately strong (see *Table 13*). Three trials had both SF-12/36 PCS and FFbHR data and their correlations were very similar, about 0.58. Another three trials had both SF-12/36 PCS and CPG and the correlations were reasonably similar, ranging from 0.41 to 0.56, and four trials had both SF-12/36 PCS and RMDQ with range 0.38 to 0.52, again similar. However, correlations between other outcomes were quite wide ranging; between CPG and RMDQ (*m* = 3 trials; range, 0.21 to 0.47) and between PSFS and RMDQ (*m* = 3; range, 0.40 to 0.70).

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

Outcome measure 1	Outcome measure 2	Trial	Pearson correlation	Cohen's kappa
CPG[a]	RMDQ[b]	BeST	0.44	0.22
		UK BEAM	0.47	0.27
		VKBIA	0.21	0.12
CPG	FFbHR[c]	Haake	0.48	0.25
PCS[d]	RMDQ	BeST	0.38	0.17
		HULLEXPROB	0.45	0.29
		Macedo	0.52	0.27
		UK BEAM	0.51	0.33
PCS	CPG	BeST	0.41	0.27
		Haake	0.49	0.27
		UK BEAM	0.56	0.31
PCS	FFbHR	Brinkhaus	0.59	0.30
		Haake	0.58	0.29
		Witt	0.59	0.27
PCS	PSFS	Macedo	0.36	0.17
PCS	ODI[e]	YACBAC	0.60	0.28
RMDQ	PSFS[f]	Hancock	0.70	0.38
		Macedo	0.40	0.26
		Pengel	0.53	0.18
PDI[g]	FFbHR	Brinkhaus	0.55	0.32
PDI	PCS	Brinkhaus	0.54	0.31

a CPG, Chronic Pain Grade disability scale; b RMDQ, Roland Morris disability questionnaire; c FFbHR, Hannover Functional Ability Questionnaire; d PCS, Physical Component Summary of SF-12 or SF-36; e ODI, Oswestry disability index; f PSFS, Patient Specific Functional Scale; g PDI, Pain Disability Index.

Cohen's kappa was less than 0.4 for all 21 comparisons. Some were similar between trials, namely for PCS and FFbHR (range, 0.27 to 0.30) and for PCS and CPG (range, 0.27 to 0.31). However, the level of agreement was never more than fair.[133] As the kappa agreement was not greater than 0.4 narrower categorisations were not investigated.

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

5.7 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 only pooled data 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 physical component score (PCS) from studies using SF-12 or SF 36.

CHAPTER 6 – PRELIMINARY STATISTICAL ANALYSES AND RESULTS

6.1 BACKGROUND

In this chapter we present the results of preliminary statistical analyses performed on the individual participant data; specifically the ANCOVA analysis 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 are presented in later chapters (see *Chapters 7-10*). We do not, in this preliminary analysis, seek to define sub-groups using multiple parameters.

6.2 STATISTICAL ANALYSIS PLAN (SAP)

In accordance with the standard operating procedure in 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 repository oversight committee (*Appendix 9*) whereas the overview of the plan is described in following sections.

6.3 DEFINITIONS

6.3.1 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 transcutaneous electrical nerve stimulation (TENS), or sham advice or education. The control arm is the non-active usual care, namely, general practitioner (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 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 included in our analyses of passive physical treatments.

6.3.2 FOLLOW-UP TIME POINT

The follow-up times are classified into short-term, mid-term and long-term. A short-term follow-up is measurement taken between two and three months post randomisation or entry to the trial. A mid-term follow-up is measurement taken at six months post randomisation or entry to the trial. A long-term follow-up is measurement taken at 12 months post randomisation or entry to the trial. Data collected at immediate follow-up (less than two months post randomisation or entry to the trial) and beyond the long-term follow-up (after 12 months post randomisation or entry to the trial) were also entered into the repository but were not considered for analysis.

6.3.2.1 Selection of follow-up time points

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

6.3.3 OUTCOME VARIABLES

6.3.3.1 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.

6.3.3.2 Health economic outcomes

For the initial economic analysis presented here, the outcome of Quality Adjusted Life Years (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 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. In order 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.[135] EQ-5D index scores can be integrated over time to estimate QALYs. QALYs were calculated for trial participants over one year of follow-up, using the area under the curve method. For each participant the area under the curve was calculated from the EQ-5D index scores 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 differences between trials were controlled for, so this potential issue was mitigated.

For one trial (Haake)[102], 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. In order to ensure 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 onto the EQ-5D index scores. The specific mapping approach applied was based on the work of Gray et al (2006);[136] in this study, a multi-nominal logit model was used to estimate the probability a particular EQ-5D dimension 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 Stata12. This mapping approach was compared to other published methods by Rowen, Brazier and Roberts (2009).[137] 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 to EQ-5D mapping might have introduced additional error 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 trial (Haake)[102] only had data up to 26 weeks. For this trial, it was assumed that the quality of life score measured at 26 weeks persisted up to 52 weeks, which allowed QALYs over one 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 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.[138]

6.3.3.3 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 measurement for physical disability and four instruments for pain (see *Appendix 9*). No single instrument was used by all RCTs to measure physical disability, 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 done on common outcomes only.

Most of the RCTs in the repository had asked participant to rate or mark on a numerical rating scale or a visual analogue scale that described either their average or worst pain at the present time or over a defined weeks or months. This item was presented either as a single standalone instrument or as an item that was part of a collective pain measurement, for example, in the McGill Pain Questionnaire where a visual analogue scale was presented as a line that anchors with ‘no pain’ at one end and ‘worst possible pain’ at the other end.[139] For the analyses of average pain, one of the following instruments from each trial, where available, was chosen (in descending order):

1. individual visual analogue scale (VAS) on average pain today,
2. average pain over the past one week,
3. average pain over the past two weeks, average pain over the past one month,

4. average pain over the past three months,
5. the individual item of the Chronic Pain Grade (CPG) pain intensity score that is equivalent to the VAS if it is available,[123]
6. the summary score of the CPG pain intensity score otherwise, or
7. the bodily pain domain of SF-12/36.[128, 129]

Where a numerical rating scale (range, 0 to 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 category, for example, with minimal, moderate or 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 was categorised into tertiles to discriminate the low and high risk or severity from the moderate risk or 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.

6.4 DATASETS

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

6.4.1 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*). Where available individual items were used to obtain the composite score for each measurement, otherwise the composite score provided to the repository were 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 either given 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 onto a sham control intervention.

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

6.4.2 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 was 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.

6.5 METHODS

6.5.1 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 standard deviation, and the categorical data were summarised as the number of participants and percentage.

6.5.2 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.[140] It can be viewed analogously as an analysis of a multicentre study where instead of multi centres in a study we have multi trials 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 where the intercept and the interaction between treatment arm and trial were modelled as random effects, and treatment arm as the fixed effect.

6.5.3 MODERATOR IDENTIFICATION

6.5.3.1 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*.

6.5.3.2 ANCOVA analysis

Analysis of covariance (ANCOVA) were performed to identify any covariate that moderates outcomes. Similarly, the one-step meta-analysis approach was used, that is, all available individual participant data were pooled into a single mixed-effects model where 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 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 analysis 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 endpoints 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- and long-term were not statistically significant. Therefore, ANCOVA was not performed for the mid- and long-term outcomes. In addition, the short-term outcomes were where the maximum clinical effects were observed between intervention and control/placebo. This is where the largest differential subgroups 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 QALY were presented. As not all of trials have the same moderators, the sample size varied depending which moderator was being assessed and for which outcome.

6.6 RESULTS

6.6.1 DESCRIPTIVE

Table 14 shows the response rates for each of the outcome of interest per treatment groups in different time points. Most trials collected data three months post randomisation or entry to the trial and this is recorded as 13 weeks whereas one RCT had specifically mentioned in their 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 outcome (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 randomised controlled trials 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.

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 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 (standard deviation, SD, 14). The average age of participants from trials that had active physical treatments was slightly lower, 44 years (*n* = 914; SD, 12) compare to the average age from trials that had passive and psychological treatments, 49 years (*n* = 3270; SD, 14) and 50 years (*n* = 1118; SD, 14), respectively. This difference is mainly due to the inclusion criteria of the trials.

Most of the participants with data in the repository had similar physical disability or functional limitation at baseline. One trial (*n* = 239) used the Oswestry disability index (ODI) as their outcome measure and the average baseline score was 33 (SD, 15), which was somewhere between no disability and moderate disability. Three trials (*n* = 4176) used the Hannover Functional Ability Questionnaire for Measuring Back-Pain Related Functional Limitations (FFbHR) and the average baseline score was 58 (SD, 21) which was slightly above moderate functional limitation. Fourteen trials (*n* = 4710) used the Roland Morris disability questionnaire (RMDQ) as their outcome measure and the average baseline score was 10 (SD, 5) which was slightly below moderate disability.

Nine trials (*n* = 6695) collected quality of life information with either the SF-12 or SF-36 instruments. The mean physical component scale (PCS) at baseline was 36 (SD, 8) and the mean mental component scale (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.

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 (*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* = 9326)**
Physical disability								
CPG-DS[a]	0	*m*=1; *n*=284	*m*=2; *n*=721	*m*=2; *n*=572	*m*=1; *n*=312	*m*=1; *n*=387	*m*=4; *n*=1052	*m*=5; *n*=3328
	4	*m*=1; *n*=228	*m*=1; *n*=315	-	*m*=1; *n*=280	-	*m*=1; *n*=262	*m*=4; *n*=1085
	8	-	-	*m*=1; *n*=109	-	-	*m*=1; *n*=120	*m*=2; *n*=229
	13	*m*=1; *n*=214	*m*=2; *n*=653	*m*=1; *n*=345	*m*=1; *n*=252	*m*=1; *n*=376	*m*=3; *n*=797	*m*=5; *n*=2637
	26	-	*m*=1; *n*=377	*m*=2; *n*=491	-	*m*=1; *n*=376	*m*=3; *n*=656	*m*=2; *n*=1900
	52	*m*=1; *n*=212	*m*=1; *n*=267	*m*=2; *n*=473	*m*=1; *n*=254	-	*m*=3; *n*=530	*m*=5; *n*=1736
	104	-	-	*m*=1; *n*=94	-	-	*m*=1; *n*=92	*m*=2; *n*=186
FFbHR[b]	0	-	*m*=3; *n*=1927	-	-	*m*=2; *n*=460	*m*=3; *n*=1789	*m*=3; *n*=4176
	6	-	*m*=1; *n*=370	-	-	*m*=1; *n*=375	*m*=1; *n*=362	*m*=1; *n*=1107
	8	-	*m*=1; *n*=140	-	-	*m*=1; *n*=70	*m*=1; *n*=74	*m*=1; *n*=284
	13	-	*m*=2; *n*=1723	-	-	*m*=1; *n*=376	*m*=2; *n*=1605	*m*=2; *n*=3704
	26	-	*m*=3; *n*=1825	-	-	*m*=2; *n*=446	*m*=3; *n*=1620	*m*=3; *n*=3891
	52	-	*m*=1; *n*=137	-	-	*m*=1; *n*=68	*m*=1; *n*=70	*m*=1; *n*=275
ODI[c]	0	-	*m*=1; *n*=159	-	-	-	*m*=1; *n*=80	*m*=1; *n*=239
	13	-	*m*=1; *n*=146	-	-	-	*m*=1; *n*=71	*m*=1; *n*=217

Outcomes	Follow-up (weeks)	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* = 9326)
	52	-	*m*=1; *n*=136	-	-	-	*m*=1; *n*=57	*m*=1; *n*=193
	104	-	*m*=1; *n*=114	-	-	-	*m*=1; *n*=50	*m*=1; *n*=164
PDI[d]	0	-	*m*=1; *n*=146	-	-	*m*=1; *n*=73	*m*=1; *n*=79	*m*=1; *n*=298
	8	-	*m*=1; *n*=140	-	-	*m*=1; *n*=70	*m*=1; *n*=74	*m*=1; *n*=284
	26	-	*m*=1; *n*=138	-	-	*m*=1; *n*=70	*m*=1; *n*=73	*m*=1; *n*=281
	52	-	*m*=1; *n*=137	-	-	*m*=1; *n*=66	*m*=1; *n*=69	*m*=1; *n*=272
PSFS[e]	0	*m*=2; *n*=150	*m*=1; *n*=119	*m*=2; *n*=148	*m*=1; *n*=62	*m*=2; *n*=188	-	*m*=3; *n*=667
	1	-	*m*=1; *n*=119	-	-	*m*=1; *n*=118	-	*m*=2; *n*=237
	2	-	*m*=1; *n*=119	-	-	*m*=1; *n*=119	-	*m*=1; *n*=238
	4	-	*m*=1; *n*=118	-	-	*m*=1; *n*=117	-	*m*=1; *n*=235
	6	*m*=1; *n*=58	-	*m*=1; *n*=54	*m*=1; *n*=57	*m*=1; *n*=59	-	*m*=1; *n*=228
	8	*m*=1; *n*=82	-	*m*=1; *n*=76	-	-	-	*m*=1; *n*=158
	12	*m*=1; *n*=57	-	*m*=1; *n*=56	*m*=1; *n*=58	*m*=1; *n*=61	-	*m*=1; *n*=232
	13	-	*m*=1; *n*=118	-	-	*m*=1; *n*=117	-	*m*=1; *n*=235
	26	*m*=1; *n*=81	-	*m*=1; *n*=74	-	-	-	*m*=1; *n*=155
	52	*m*=2; *n*=136	-	*m*=2; *n*=132	*m*=1; *n*=56	*m*=1; *n*=56	-	*m*=2; *n*=380
RMDQ[f]	0	*m*=7; *n*=907	*m*=7; *n*=1087	*m*=7; *n*=1120	*m*=3; *n*=446	*m*=3; *n*=212	*m*=6; *n*=938	*m*=14; *n*=4710
	1	-	*m*=1; *n*=119	-	-	*m*=1; *n*=118	-	*m*=1; *n*=237
	2	-	*m*=2; *n*=119	-	-	*m*=1; *n*=118	-	*m*=1; *n*=237

Outcomes	**Follow-up (weeks)**	**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* = 9326)**
	4	*m*=1; *n*=234	*m*=2; *n*=436	-	*m*=1; *n*=283	*m*=1; *n*=117	*m*=1; *n*=264	*m*=2; *n*=1334
	6	*m*=2; *n*=144	*m*=1; *n*=23	*m*=1; *n*=55	*m*=1; *n*=58	*m*=2; *n*=81	*m*=1; *n*=94	*m*=3; *n*=455
	8	*m*=1; *n*=82	-	*m*=2; *n*=186	-	-	*m*=1; *n*=120	*m*=2; *n*=388
	10	*m*=1; *n*=107	-	-	*m*=1; *n*=55	-	*m*=1; *n*=50	*m*=1; *n*=212
	12	*m*=1; *n*=58	-	*m*=1; *n*=58	*m*=1; *n*=59	*m*=1; *n*=61	-	*m*=1; *n*=236
	13	*m*=3; *n*=433	*m*=7; *n*=963	*m*=4; *n*=670	*m*=1; *n*=255	*m*=2; *n*=135	*m*=3; *n*=537	*m*=9; *n*=2993
	26	*m*=4; *n*=371	*m*=2; *n*=262	*m*=5; *n*=706	*m*=1; *n*=53	-	*m*=5; *n*=474	*m*=8; *n*=1866
	52	*m*=7; *n*=722	*m*=5; *n*=771	*m*=7; *n*=903	*m*=3; *n*=365	*m*=1; *n*=56	*m*=6; *n*=690	*m*=12; *n*=3507
	104	*m*=1; *n*=83	*m*=1; *n*=95	*m*=1; *n*=94	-	-	*m*=1; *n*=92	*m*=2; *n*=364
Troublesomeness	0	*m*=2; *n*=344	*m*=3; *n*=556	*m*=1; *n*=426	*m*=1; *n*=312	-	*m*=3; *n*=604	*m*=4; *n*=2242
	4	*m*=1; *n*=225	*m*=1; *n*=313	-	*m*=1; *n*=279	-	*m*=1; *n*=262	*m*=1; *n*=1079
	13	*m*=2; *n*=280	*m*=3; *n*=494	-	*m*=1; *n*=253	-	*m*=2; *n*=318	*m*=3; *n*=1345
	52	*m*=2; *n*=302	*m*=3; *n*=493	-	*m*=1; *n*=252	-	*m*=2; *n*=297	*m*=8; *n*=1344
	104	-	*m*=1; *n*=113	-	-	-	*m*=1; *n*=50	*m*=3; *n*=162
Pain								
CPG-PS[g]	0	*m*=1; *n*=283	*m*=2; *n*=721	*m*=2; *n*=582	*m*=1; *n*=312	*m*=1; *n*=387	*m*=4; *n*=1054	*m*=4; *n*=3339
	4	*m*=1; *n*=228	*m*=1; *n*=316	-	*m*=1; *n*=281	-	*m*=1; *n*=261	*m*=1; *n*=1086
	6	-	*m*=1; *n*=370	-	-	*m*=1; *n*=375	*m*=1; *n*=362	*m*=1; *n*=1107

Outcomes	**Follow-up (weeks)**	**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* = 9326)**
	8	-	-	*m*=1; *n*=110	-	-	*m*=1; *n*=120	*m*=1; *n*=230
	13	*m*=1; *n*=214	*m*=2; *n*=653	*m*=1; *n*=354	*m*=1; *n*=252	*m*=1; *n*=376	*m*=3; *n*=799	*m*=3; *n*=2648
	26	-	*m*=1; *n*=377	*m*=2; *n*=497	-	*m*=1; *n*=376	*m*=3; *n*=661	*m*=3; *n*=1911
	52	*m*=1; *n*=211	*m*=1; *n*=269	*m*=2; *n*=491	*m*=1; *n*=253	-	*m*=4; *n*=536	*m*=3; *n*=1760
	104	-	-	*m*=1; *n*=94	-	-	*m*=1; *n*=92	*m*=1; *n*=186
Visual analogue scale								
Average pain today	0	*m*=2; *n*=253	*m*=3; *n*=461	*m*=1; *n*=196	*m*=1; *n*=61	*m*=1; *n*=120	*m*=1; *n*=51	*m*=3; *n*=1142
	1	-	*m*=1; *n*=119	-	-	*m*=1; *n*=119	-	*m*=1; *n*=238
	2	-	*m*=1; *n*=119	-	-	*m*=1; *n*=119	-	*m*=1; *n*=238
	3	-	*m*=1; *n*=118	-	-	*m*=1; *n*=118	-	*m*=1; *n*=236
	4	*m*=1; *n*=83	*m*=1; *n*=118	*m*=1; *n*=80	-	*m*=1; *n*=118	-	*m*=2; *n*=399
	6	-	*m*=1; *n*=36	-	-	*m*=1; *n*=38	-	*m*=1; *n*=74
	8	*m*=1; *n*=81	*m*=1; *n*=24	*m*=1; *n*=79	-	*m*=1; *n*=23	-	*m*=2; *n*=207
	10	*m*=1; *n*=107	*m*=1; *n*=16	-	*m*=1; *n*=55	*m*=1; *n*=18	*m*=1; *n*=49	*m*=2; *n*=245
	11	-	*m*=1; *n*=15	-	-	*m*=1; *n*=17	-	*m*=1; *n*=32
	12	-	*m*=1; *n*=15	-	-	*m*=1; *n*=17	-	*m*=1; *n*=32

Outcomes	**Follow-up (weeks)**	**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* = 9326)**
	13	*m*=1; *n*=81	*m*=1; *n*=153	*m*=2; *n*=231	-	-	-	*m*=1; *n*=465
	17	*m*=1; *n*=79	-	*m*=1; *n*=75	-	-	-	*m*=1; *n*=154
	21	*m*=1; *n*=81	-	*m*=1; *n*=76	-	-	-	*m*=1; *n*=157
	26	*m*=2; *n*=186	-	*m*=1; *n*=75	*m*=1; *n*=53	-	-	*m*=2; *n*=314
	30	*m*=1; *n*=79	-	*m*=1; *n*=72	-	-	-	*m*=1; *n*=151
	34	*m*=1; *n*=81	-	*m*=1; *n*=73	-	-	-	*m*=1; *n*=154
	39	*m*=1; *n*=80	-	*m*=1; *n*=74	-	-	-	*m*=1; *n*=154
	43	*m*=1; *n*=78	-	*m*=1; *n*=74	-	-	-	*m*=1; *n*=152
	47	*m*=1; *n*=76	-	*m*=1; *n*=71	-	-	-	*m*=1; *n*=147
	52	*m*=2; *n*=183	*m*=1; *n*=164	*m*=2; *n*=238	*m*=1; *n*=53	-	-	*m*=6; *n*=638
Average pain over past one week	0	*m*=2; *n*=150	*m*=2; *n*=235	*m*=3; *n*=349	*m*=1; *n*=63	*m*=2; *n*=84	-	*m*=4; *n*=881
	1	-	*m*=1; *n*=235	-	-	*m*=1; *n*=119	-	*m*=1; *n*=238
	2	-	*m*=1; *n*=235	-	-	*m*=1; *n*=119	-	*m*=1; *n*=238
	3	-	*m*=1; *n*=235	-	-	*m*=1; *n*=118	-	*m*=1; *n*=237
	4	*m*=1; *n*=82	*m*=2; *n*=152	*m*=1; *n*=80	-	*m*=2; *n*=134	-	*m*=3; *n*=448
	6	*m*=1; *n*=59	*m*=1; *n*=49	*m*=1; *n*=55	*m*=1; *n*=58	*m*=2; *n*=97	-	*m*=2; *n*=306
	8	*m*=1; *n*=81	*m*=1; *n*=24	*m*=1; *n*=79	-	*m*=1; *n*=24	-	*m*=2; *n*=208

Outcomes	**Follow-up (weeks)**	**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* = 9326)**
	10	-	*m*=1; *n*=16	-	-	*m*=1; *n*=19	-	*m*=1; *n*=35
	11	-	*m*=1; *n*=11	-	-	*m*=1; *n*=17	-	*m*=1; *n*=33
	12	*m*=1; *n*=58	*m*=1; *n*=15	*m*=1; *n*=58	*m*=1; *n*=59	*m*=2; *n*=78	-	*m*=2; *n*=268
	13	*m*=1; *n*=81	*m*=2; *n*=180	*m*=2; *n*=231	-	*m*=1; *n*=9	-	*m*=3; *n*=501
	17	*m*=1; *n*=79	-	*m*=1; *n*=75	-	-	-	*m*=1; *n*=154
	21	*m*=1; *n*=81	-	*m*=1; *n*=76	-	-	-	*m*=1; *n*=157
	26	*m*=1; *n*=81	*m*=1; *n*=21	*m*=1; *n*=75	-	*m*=1; *n*=6	-	*m*=2; *n*=183
	30	*m*=1; *n*=79	-	*m*=1; *n*=72	-	-	-	*m*=1; *n*=151
	34	*m*=1; *n*=81	-	*m*=1; *n*=73	-	-	-	*m*=1; *n*=154
	39	*m*=1; *n*=80	-	*m*=1; *n*=74	-	-	-	*m*=1; *n*=154
	43	*m*=1; *n*=78	-	*m*=1; *n*=74	-	-	-	*m*=1; *n*=152
	47	*m*=1; *n*=77	-	*m*=1; *n*=71	-	-	-	*m*=1; *n*=148
	52	*m*=2; *n*=140	*m*=1; *n*=163	*m*=3; *n*=297	*m*=1; *n*=57	*m*=1; *n*=56	-	*m*=3; *n*=713
Average pain over past one month	0	-	*m*=1; *n*=24	-	-	*m*=1; *n*=24	-	*m*=1; *n*=48
	6	-	*m*=1; *n*=23	-	-	*m*=1; *n*=22	-	*m*=1; *n*=45
	13	-	*m*=1; *n*=22	-	-	*m*=1; *n*=18	-	*m*=1; *n*=40

Outcomes	**Follow-up (weeks)**	**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* = 9326)**
Worst pain today	0	*m*=1; *n*=111	-	-	*m*=1; *n*=61	-	*m*=1; *n*=51	*m*=1; *n*=223
	10	*m*=1; *n*=107	-	-	*m*=1; *n*=53	-	*m*=1; *n*=49	*m*=1; *n*=209
	26	*m*=1; *n*=103	-	-	*m*=1; *n*=53	-	-	*m*=1; *n*=156
	52	*m*=1; *n*=103	-	-	*m*=1; *n*=52	-	-	*m*=1; *n*=155
Worst pain over past one month	0	-	*m*=2; *n*=24	-	-	*m*=1; *n*=24	-	*m*=2; *n*=48
	6	-	*m*=1; *n*=23	-	-	*m*=1; *n*=22	-	*m*=2; *n*=45
	13	-	*m*=1; *n*=22	-	-	*m*=1; *n*=18	-	*m*=2; *n*=40
Quality of life								
SF-12/36[h] PCS	0	*m*=4; *n*=617	*m*=7; *n*=2544	*m*=2; *n*=507	*m*=1; *n*=305	*m*=2; *n*=460	*m*=6; *n*=2262	*m*=9; *n*=6695
	4	*m*=1; *n*=214	*m*=1; *n*=300	-	*m*=1; *n*=264	-	*m*=1; *n*=249	*m*=1; *n*=1027
	8	*m*=1; *n*=82	*m*=1; *n*=139	*m*=1; *n*=76	-	*m*=1; *n*=69	*m*=1; *n*=73	*m*=2; *n*=439
	13	*m*=3; *n*=415	*m*=6; *n*=2276	*m*=1; *n*=332	*m*=1; *n*=243	*m*=1; *n*=376	*m*=5; *n*=2006	*m*=7; *n*=5648
	26	*m*=2; *n*=185	*m*=4; *n*=1850	*m*=2; *n*=436	-	*m*=2; *n*=444	*m*=4; *n*=1711	*m*=6; *n*=4626
	52	*m*=4; *n*=469	*m*=5; *n*=719	*m*=2; *n*=449	*m*=1; *n*=235	*m*=1; *n*=68	*m*=4; *n*=545	*m*=7; *n*=2485

Outcomes	**Follow-up (weeks)**	**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* = 9326)**
	104	*m*=1; *n*=83	*m*=2; *n*=206	-	-	-	*m*=1; *n*=49	*m*=2; *n*=338
SF-12/36 MCS[i]	0	*m*=4; *n*=617	*m*=7; *n*=2544	*m*=2; *n*=507	*m*=1; *n*=305	*m*=2; *n*=460	*m*=6; *n*=2262	*m*=9; *n*=6695
	4	*m*=1; *n*=214	*m*=1; *n*=300	-	*m*=1; *n*=264	-	*m*=1; *n*=249	*m*=1; *n*=1027
	8	*m*=1; *n*=82	*m*=1; *n*=139	*m*=1; *n*=76	-	*m*=1; *n*=69	*m*=1; *n*=73	*m*=2; *n*=439
	13	*m*=3; *n*=415	*m*=6; *n*=2276	*m*=1; *n*=332	*m*=1; *n*=243	*m*=1; *n*=376	*m*=5; *n*=2006	*m*=7; *n*=5648
	26	*m*=2; *n*=185	*m*=4; *n*=1850	*m*=2; *n*=436	-	*m*=2; *n*=444	*m*=4; *n*=1711	*m*=6; *n*=4626
	52	*m*=4; *n*=469	*m*=5; *n*=719	*m*=2; *n*=449	*m*=1; *n*=235	*m*=1; *n*=68	*m*=4; *n*=545	*m*=7; *n*=2485
	104	*m*=1; *n*=83	*m*=2; *n*=206	-	-	-	*m*=1; *n*=49	*m*=2; *n*=338
Health utility								
EQ-5D-3L	0	*m*=1; *n*=85	-	-	-	-	*m*=1; *n*=94	*m*=1; *n*=179
	6	*m*=1; *n*=85	-	-	-	-	*m*=1; *n*=94	*m*=1; *n*=179
	26	*m*=1; *n*=77	-	-	-	-	*m*=1; *n*=86	*m*=1; *n*=163
	52	*m*=1; *n*=82	-	-	-	-	*m*=1; *n*=88	*m*=1; *n*=170

a CPG-DS, chronic pain grade disability score; b FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; c ODI, Oswestry disability index; d PDI, pain disability index; e PSFS, patient specific functional scale; f RMDQ, Roland Morris disability questionnaire; g CPG-PS, chronic pain grade pain intensity score; h PCS, physical component scale of SF-12/36; i MCS, mental component scale of SF-12/36.

Table 15 Demographics and clinical characteristics at baseline by treatment arms.

Characteristics	Active physical (*m* = 7; *n* = 914)	Passive physical (*m* = 12; *n* = 3,270)	Psychological (*m* = 7; *n* = 1,120)	Combination (*m* = 3; *n* = 451)	Sham (*m* = 6; *n* = 688)	Control (*m* = 10; *n* = 2,885)	All (*m* = 19; *n* = 9,328)
Demographics							
Age, years							
No. of trials, *m*	7	12	7	3	6	10	19
n	914	3,270	1,118	451	688	2,885	9,326
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
Sex							
No. of trials, *m*	7	12	7	3	6	10	19
Female (%)	497 (54.4)	1,907 (58.3)	655 (58.5)	237 (52.6)	412 (59.9)	1,641 (56.9)	5,349 (57.4)
Male (%)	417 (45.6)	1,363 (41.7)	464 (41.5)	214 (47.5)	276 (40.1)	1,243 (43.1)	3,977 (42.6)
Ethnicity							
No. of trials, *m*	1	1	4	-	-	4	5
White (%)	65 (75.6)	159 (100.0)	667 (87.8)	-	-	478 (89.4)	1,369 (88.9)
Mixed	-	-	4 (0.5)	-	-	3 (0.6)	7 (0.5)
Black	-	-	26 (3.4)	-	-	21 (3.9)	47 (3.1)

Characteristics	Active physical (*m* = 7; *n* = 914)	Passive physical (*m* = 12; *n* = 3,270)	Psychological (*m* = 7; *n* = 1,120)	Combination (*m* = 3; *n* = 451)	Sham (*m* = 6; *n* = 688)	Control (*m* = 10; *n* = 2,885)	All (*m* = 19; *n* = 9,328)
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)
Smoking status							
No. 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)
Employment status							
No. of trials, *m*	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)	1,745 (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)	1,060 (30.8)
BMI[a]							
No. of trials, *m*	2	4	2	-	2	2	5
n	222	811	156	-	453	462	2,104
Mean	27.03	26.60	26.52	-	26.45	26.42	26.57

Characteristics	**Active physical (*m* = 7; *n* = 914)**	**Passive physical (*m* = 12; *n* = 3,270)**	**Psychological (*m* = 7; *n* = 1,120)**	**Combination (*m* = 3; *n* = 451)**	**Sham (*m* = 6; *n* = 688)**	**Control (*m* = 10; *n* = 2,885)**	**All (*m* = 19; *n* = 9,328)**
SD	5.31	4.60	5.22	-	4.73	4.48	4.73
Physical disability							
CPG-DS[b] *(0 to 100; 100=worst)*[a]							
No. of trials, *m*	1	2	2	1	1	5	4
n	284	721	572	312	387	1,052	3,328
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
FFbHR[c] *(0 to 100; 100=best)*							
No. of trials, *m*	-	3	-	-	2	3	3
n	-	1,927	-	-	460	1,789	4,176
Mean	-	58.33	-	-	48.01	59.38	57.64
SD	-	20.63	-	-	16.14	20.69	20.5
ODI [d] *(0 to 100; 100=worst)*							
No. of trials, *m*	-	1	-	-	-	1	1
n	-	159	-	-	-	80	239

Characteristics	**Active physical (*m* = 7; *n* = 914)**	**Passive physical (*m* = 12; *n* = 3,270)**	**Psychological (*m* = 7; *n* = 1,120)**	**Combination (*m* = 3; *n* = 451)**	**Sham (*m* = 6; *n* = 688)**	**Control (*m* = 10; *n* = 2,885)**	**All (*m* = 19; *n* = 9,328)**
Mean	-	33.72	-	-	-	31.36	32.93
SD	-	15.40	-	-	-	14.24	15.03
PDI[e] *(0 to 70; 70=worst)*							
No. of trials, *m*	-	1	-	-	1	1	1
n	-	146	-	-	73	79	298
Mean	-	28.92	-	-	31.53	30.95	30.10
SD	-	11.12	-	-	11.14	13.27	11.75
PSFS[f] *(0 to 10; 10=best)*							
No. of trials, *m*	2	1	2	1	2	-	3
n	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
RMDQ[g] *(0 to 24; 24=worst)*							
No. of trials, *m*	7	7	7	3	3	6	14
n	907	1,087	1,120	446	212	938	4,710
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

Characteristics	**Active physical (*m* = 7; *n* = 914)**	**Passive physical (*m* = 12; *n* = 3,270)**	**Psychological (*m* = 7; *n* = 1,120)**	**Combination (*m* = 3; *n* = 451)**	**Sham (*m* = 6; *n* = 688)**	**Control (*m* = 10; *n* = 2,885)**	**All (*m* = 19; *n* = 9,328)**
Troublesomeness							
No. of trials, *m*	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	1,008
Very troublesome (%)	115	205	151	107	-	211	789
Extremely troublesome (%)	39	72	38	22	-	54	225
Pain							
CPG-PS[h] *(0 to 100; 100=worst)*[a]							
No. of trials, *m*	1	2	3	1	1	5	5
n	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
Average pain (0 to 100; 100=worst)[b]							
No. of trials, *m*	4	6	6	3	5	6	12

Characteristics	**Active physical (*m* = 7; *n* = 914)**	**Passive physical (*m* = 12; *n* = 3,270)**	**Psychological (*m* = 7; *n* = 1,120)**	**Combination (*m* = 3; *n* = 451)**	**Sham (*m* = 6; *n* = 688)**	**Control (*m* = 10; *n* = 2,885)**	**All (*m* = 19; *n* = 9,328)**
n	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							
SF-12/36 PCS[i] *(0 to 100; 100=best)*							
No. of trials, *m*	4	7	2	1	2	6	9
n	617	2,544	507	305	460	2,262	6,695
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
SF-12/36 MCS[j] *(0 to 100; 100=best)*							
No. of trials, *m*	4	7	2	1	2	6	9
n	617	2,544	507	305	460	2,262	6,695
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							

Characteristics	**Active physical (*m* = 7; *n* = 914)**	**Passive physical (*m* = 12; *n* = 3,270)**	**Psychological (*m* = 7; *n* = 1,120)**	**Combination (*m* = 3; *n* = 451)**	**Sham (*m* = 6; *n* = 688)**	**Control (*m* = 10; *n* = 2,885)**	**All (*m* = 19; *n* = 9,328)**
EQ-5D-3L (-0.11 to 1;1=best)							
No. of trials, *m*	4	4	2	2	-	5	7
n	593	740	652	371	-	724	3,080
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							
DASS^k-DE (0 to 42; 42=worst)							
No. of trials, *m*	1	-	1	1	1	-	1
n	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^l							
No. of trials, *m*	2	1	-	1	-	2	2
Type N^m (%)	135 (36.49)	122 (36.75)	-	116 (37.54)	-	184 (44.88)	557 (39.20)
Type R^n (%)	147 (39.73)	147 (44.28)	-	120 (38.83)	-	158 (38.54)	572 (40.25)

Characteristics	**Active physical (*m* = 7; *n* = 914)**	**Passive physical (*m* = 12; *n* = 3,270)**	**Psychological (*m* = 7; *n* = 1,120)**	**Combination (*m* = 3; *n* = 451)**	**Sham (*m* = 6; *n* = 688)**	**Control (*m* = 10; *n* = 2,885)**	**All (*m* = 19; *n* = 9,328)**
Type DD[o] (%)	55 (14.86)	41 (12.35)	-	46 (14.89)	-	49 (11.95)	191 (13.44)
Type DS[p] (%)	33 (8.92)	22 (6.63)	-	27 (8.74)	-	19 (4.63)	101 (7.11)
HADS[q]-DE (0 to 21; 21=worst)							
No. of trials, *m*	-	-	1	-	-	1	1
n	-	-	464	-	-	231	695
Mean	-	-	6.04	-	-	5.54	5.87
SD	-	-	3.81	-	-	3.6	3.75
MZDI[r] (0 to 69; 69=worst)							
No. of trials, *m*	2	2	1	1	-	2	3
n	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
Anxiety							
DASS[k]-AN (0 to 42; 42=worst)							
No. of trials, *m*	1	-	1	1	1	-	1

Characteristics	**Active physical (*m* = 7; *n* = 914)**	**Passive physical (*m* = 12; *n* = 3,270)**	**Psychological (*m* = 7; *n* = 1,120)**	**Combination (*m* = 3; *n* = 451)**	**Sham (*m* = 6; *n* = 688)**	**Control (*m* = 10; *n* = 2,885)**	**All (*m* = 19; *n* = 9,328)**
n	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
HADS[q]*-AN (0 to 21; 21=worst)*							
No. of trials, *m*	-	-	1	-	-	1	1
n	-	-	458	-	-	230	688
Mean	-	-	8.22	-	-	7.49	7.98
SD	-	-	4.3	-	-	4.43	4.35
Fear avoidance							
ALBPSQ[s]*-FA (0 to 30; 30=worst)*							
No. of trials, *m*	2	-	2	1	1	-	2
n	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

Characteristics	**Active physical (*m* = 7; *n* = 914)**	**Passive physical (*m* = 12; *n* = 3,270)**	**Psychological (*m* = 7; *n* = 1,120)**	**Combination (*m* = 3; *n* = 451)**	**Sham (*m* = 6; *n* = 688)**	**Control (*m* = 10; *n* = 2,885)**	**All (*m* = 19; *n* = 9,328)**
FABQ[t]-PC (0 to 24; 24=worst)							
No. of trials, *m*	2	3	1	1	2	4	5
n	366	840	443	311	506	1,016	3,482
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
TSK[u] (16 to 68; 68=worst)							
No. of trials, *m*	2	1	4	2	1	3	5
n	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)							
CSQ[v]-CAT (0 to 36; 36=worst)							
No. of trials, *m*	1	1	2	-	-	-	2
n	86	193	282	-	-	-	561
Mean	10.84	7.83	9.62	-	-	-	9.19

Characteristics	**Active physical (*m* = 7; *n* = 914)**	**Passive physical (*m* = 12; *n* = 3,270)**	**Psychological (*m* = 7; *n* = 1,120)**	**Combination (*m* = 3; *n* = 451)**	**Sham (*m* = 6; *n* = 688)**	**Control (*m* = 10; *n* = 2,885)**	**All (*m* = 19; *n* = 9,328)**
SD	7.61	6.65	7.22	-	-	-	7.16
PRSS[w]-CAT (0 to 45; 45=worst)							
No. of trials, *m*	1	1	1	1	2	-	2
n	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[v]-CSS (0 to 36; 36=best)							
No. of trials, *m*	-	1	1	-	-	-	1
n	-	198	196	-	-	-	394
Mean	-	25.13	25.33	-	-	-	25.23
SD	-	6.23	6.64	-	-	-	6.43
PRSS[w]-CSS (0 to 45; 45=best)							
No. of trials, *m*	1	2	1	1	2	-	2

Characteristics	**Active physical (*m* = 7; *n* = 914)**	**Passive physical (*m* = 12; *n* = 3,270)**	**Psychological (*m* = 7; *n* = 1,120)**	**Combination (*m* = 3; *n* = 451)**	**Sham (*m* = 6; *n* = 688)**	**Control (*m* = 10; *n* = 2,885)**	**All (*m* = 19; *n* = 9,328)**
n	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 [x] *(0 to 60; 60=best)*							
No. of trials, *m*	3	1	3	1	1	1	4
n	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[y] *(0 to 39; 39=worst)*							
No. of trials, *m*	2	2	1	1	-	2	3
n	372	526	195	310	-	411	1,814
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 to 33; 33=worst)							

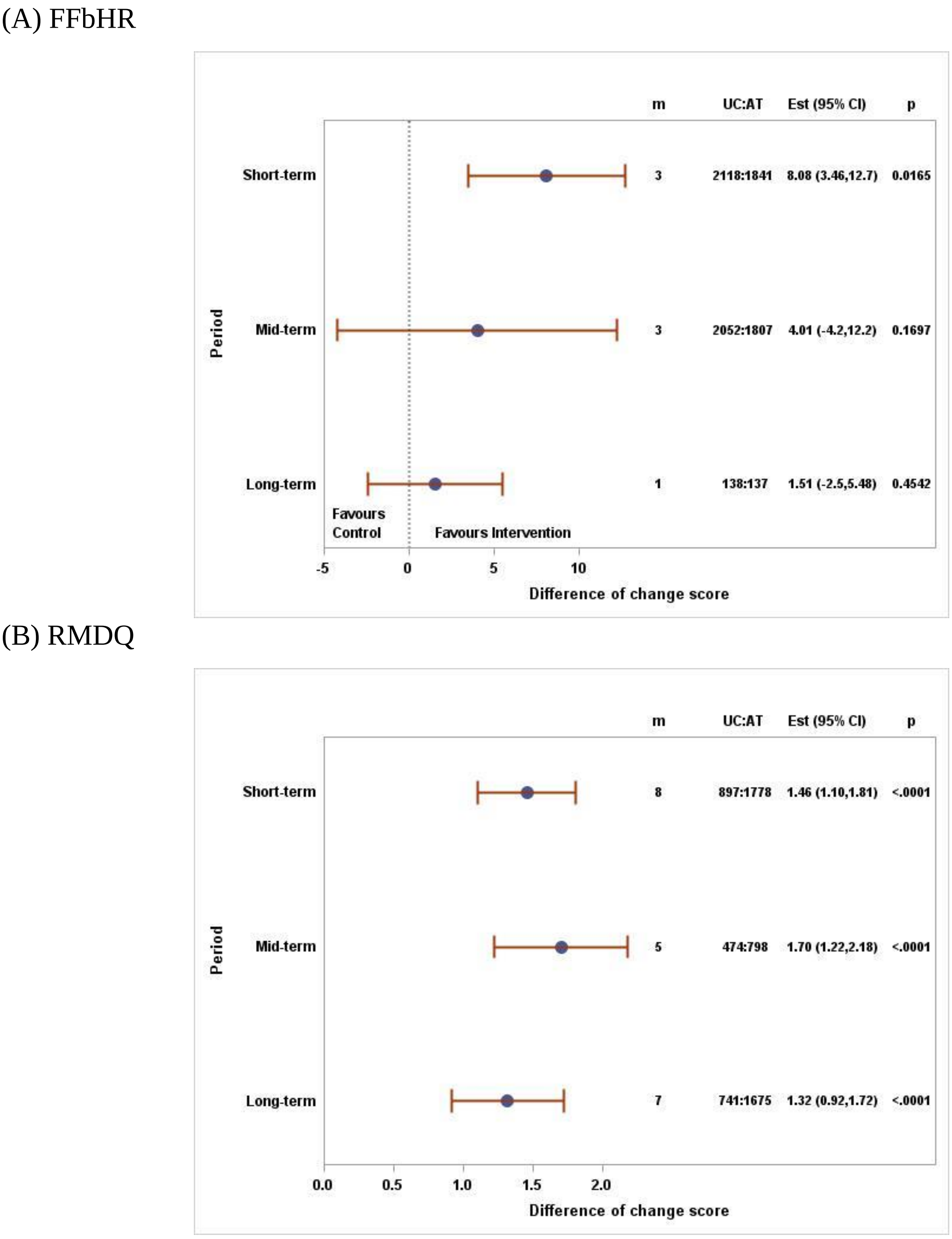
Characteristics	**Active physical (*m* = 7; *n* = 914)**	**Passive physical (*m* = 12; *n* = 3,270)**	**Psychological (*m* = 7; *n* = 1,120)**	**Combination (*m* = 3; *n* = 451)**	**Sham (*m* = 6; *n* = 688)**	**Control (*m* = 10; *n* = 2,885)**	**All (*m* = 19; *n* = 9,328)**
No. of trials, *m*	-	1	1	-	-	-	1
n	-	185	170	-	-	-	355
Mean	-	14.21	14.26	-	-	-	14.24
SD	-	6.10	6.36	-	-	-	6.22
SES^{z}-SE (10 to 40; 40=worst)							
No. of trials, *m*	-	1	-	-	1	1	1
n	-	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)							
McGill-AF (0 to 12; 12=worst)							
No. of trials, *m*	-	1	1	-	-	-	1
n	-	192	187	-	-	-	379
Mean	-	4.21	4.25	-	-	-	4.23
SD	-	3.31	3.36	-	-	-	3.33

Characteristics	**Active physical (*m* = 7; *n* = 914)**	**Passive physical (*m* = 12; *n* = 3,270)**	**Psychological (*m* = 7; *n* = 1,120)**	**Combination (*m* = 3; *n* = 451)**	**Sham (*m* = 6; *n* = 688)**	**Control (*m* = 10; *n* = 2,885)**	**All (*m* = 19; *n* = 9,328)**
SES-AF (14 to 56; 56=worst)							
No. of trials, *m*	-	1	-	-	1	1	1
n	-	146	-	-	73	79	298
Mean	-	50.19	-	-	50.88	50.01	50.31
SD	-	8.38	-	-	8.17	9.34	8.57

a BMI. body mass index; b CPG-DS, chronic pain grade disability score; c FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; d ODI, Oswestry disability index; e PDI, pain disability index; f PSFS, patient specific functional scale; g RMDQ, Roland Morris disability questionnaire; h CPG-PS, chronic pain grade pain intensity score; i PCS, physical component scale of SF-12/36; j MCS, mental component scale of SF-12/36; k DASS, depression anxiety stress scales; l DRAM, distress and risk assessment method; m Type N, normal; n Type R, at risk; o Type DD, distressed-depressive; p Type DS, distressed-somatic; q HADS, hospital anxiety and depression scale; r MZDI, modified Zung depression index; s ALBPSQ, acute low back pain screening questionnaire; t FABQ, fear-avoidance beliefs questionnaire; u TSK, Tampa scale for kinesiophobia; v CSQ, coping strategy questionnaire; w PRSS, pain related self-statement; x PSEQ, pain self-efficacy questionnaire; y MSPQ, modified somatic perception questionnaire; z SES, pain experience scale.

6.6.2 ONE-STEP META-ANALYSIS

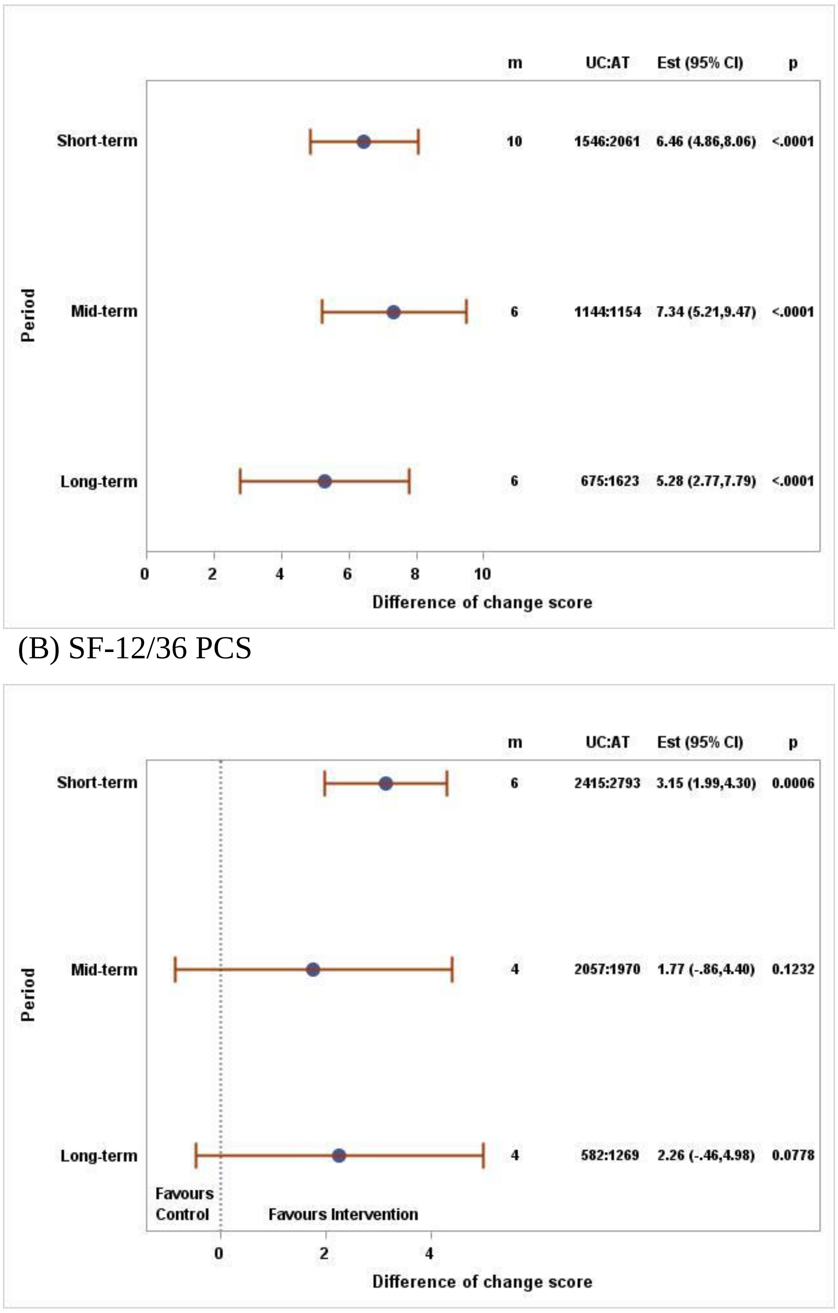
Boxplots of change of outcome measures from baseline to short-, mid- and long-term follow-up by treatment arms show that participants in all groups are behaving as expected with all groups improving over time (data not shown). This observation was examined further in the one-step meta-analysis (adjusting for study effects) and the results are shown in *Figure 17, Figure 18* and *Figure 19* and *Table 16*. There was a statistically significant difference between control and intervention for all outcomes at the short-term follow-up.


(A) FFbHR

(B) RMDQ

Abbreviations: m, number of trials; UC, number of participants in the control arm; AT, number of participants in the intervention arm; Est (95% CI), estimated treatment efficacy and 95% confidence interval; *p*, *p*-value.

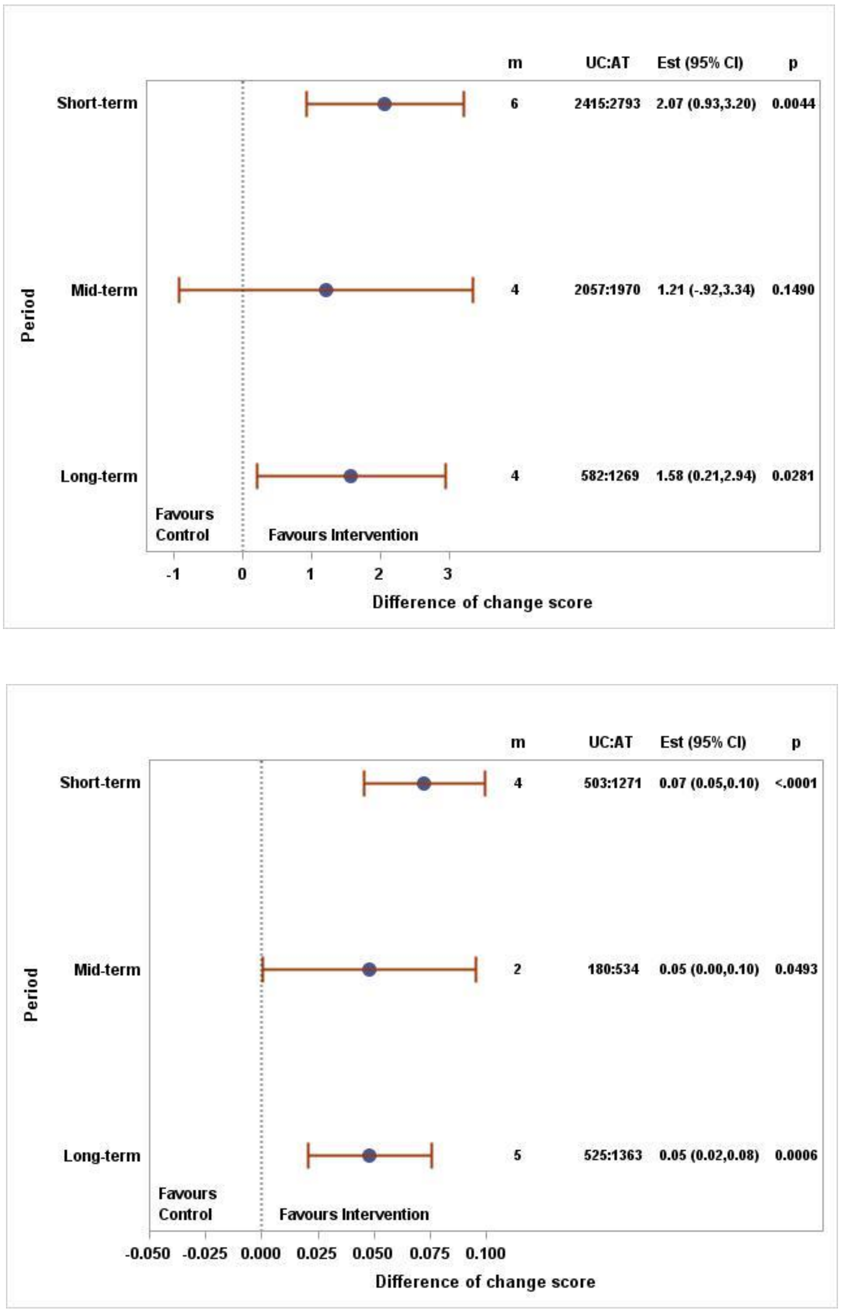
Figure 17 The estimated efficacy between control (non-active usual care and sham) and intervention treatments from one-step meta-analysis for (A) Hannover functional ability questionnaire for measuring back-pain related functional limitations (FFbHR) and (B) Roland-Morris Disability Questionnaire (RMDQ)


(A) Average pain (VAS)

(B) SF-12/36 PCS

Abbreviations: m, number of trials; UC, number of participants in the control arm; AT, number of participants in the intervention arm; Est (95% CI), estimated treatment efficacy and 95% confidence interval; *p*, *p*-value.

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 visual analogue scale) and (B) physical component scale of SF-12/36 (PCS)



(A) SF-12/36 MCS

(B) EQ-5D

Abbreviations: m, number of trials; UC, number of participants in the control arm; AT, number of participants in the intervention arm; Est (95% CI), estimated treatment efficacy and 95% confidence interval; *p*, *p*-value.

Figure 19 The estimated efficacy between control (non-active usual care and sham) and intervention treatments from one-step meta-analysis for (A) mental component scale of SF-12/36 (MCS) and (B) EQ-5D.

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% confidence interval).[a]

Outcomes	**No. of trials, *m***	**Intervention**	**Control[b]**	**Difference[c]**	***p*-value**
FFbHR[d]	3	*n* = 1841	*n* = 2118		
		13.88	5.80	8.08	0.0165
		(1.24, 26.51)	(-6.93, 18.53)	(3.46, 12.69)	
RMDQ[e]	8	*n* = 1778	*n* = 897		
		4.43	2.97	1.46	<0.0001
		(1.56, 7.29)	(0.10, 5.84)	(1.10, 1.81)	
Average pain[f]	10	*n* = 2061	*n* = 1546		
		18.03	11.57	6.46	<0.0001
		(8.65, 27.41)	(2.18, 20.97)	(4.86, 8.06)	
PCS[g]	6	*n* = 2793	*n* = 2415		
		6.86	3.72	3.15	0.0006
		(4.90, 8.83)	(1.75, 5.68)	(1.99, 4.30)	
MCS[h]	6	*n* = 2793	*n* = 2415		
		2.69	0.62	2.07	0.0044
		(1.54, 3.84)	(-0.55, 1.79)	(0.93, 3.20)	
EQ-5D	4	*n* = 1271	*n* = 503		
		0.1065	0.03422	0.072	<0.0001
		(0.008, 0.205)	(-0.059, 0.127)	(0.04538, 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 FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; e RMDQ, Roland Morris disability questionnaire; f Obtained from either visual analogue scale or pain intensity score of chronic pain grade scale (see *Section 6.3.3.3*); g PCS, physical component scale of SF-12/36; h MCS, mental component scale of SF-12/36.

6.6.3 ANCOVA ANALYSIS

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, MCS. There were three trials with FFbHR short-term outcome 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 participant was weakly statistically significant (p=0.2018). Participants with lower value of FFbHR at baseline (more physical disability) had 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 ANCOVA analysis for short-term outcomes (change from baseline to short-term follow-up)[a].

Outcome	Covariates	No. of trials, *m*	No. of participants, *AT*[b]:*UC*[c]	Estimate (interaction term)	LCI[d]	UCI[e]	*p*-value
FFbHR[f]							
	Age	3	1841:2118	-0.051	-0.131	0.028	0.2018
	Sex (male vs. female)[l]	3	1841:2118	-0.684	-2.851	1.483	0.5361
	FFBHR	3	1841: 2118	-0.177	-0.229	-0.125	<.0001
	PCS (<50 vs. ≥50)[m]	3	1718:2000	2.521	-2.361	7.403	0.3114
	PCS (continuous)	3	1718:2000	-0.318	-0.451	-0.186	<.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
RMDQ[g]							
	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

Outcome	Covariates	No. of trials, *m*	No. of participants, *AT*[b]:*UC*[c]	Estimate (interaction term)	LCI[d]	UCI[e]	*p*-value
	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[n]			-0.295	-1.713	1.123	0.6832
	Moderate risk[o]			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[p]			0.387	-2.271	3.046	0.7747
	Moderate[q]			2.030	-0.461	4.521	0.1099
	Coping	3	620:348				0.6797
	Positive[r]			0.428	-1.127	1.982	0.5895
	Moderate[s]			0.729	-0.904	2.362	0.3813
	Fear avoidance	7	1706:858				0.1933
	Positive[t]			0.786	-0.125	1.697	0.0907
	Moderate[u]			0.714	-0.225	1.653	0.1361
Average pain[h]							

Outcome	**Covariates**	**No. of trials, *m***	**No. of participants, *AT*[b]:*UC*[c]**	**Estimate (interaction term)**	**LCI[d]**	**UCI[e]**	***p*-value**
	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
	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

Outcome	Covariates	No. of trials, *m*	No. of participants, *AT*[b]:*UC*[c]	Estimate (interaction term)	LCI[d]	UCI[e]	*p*-value
	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
SF-12/36 PCS[i]							
	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

Outcome	Covariates	No. of trials, *m*	No. of participants, *AT*[b]:*UC*[c]	Estimate (interaction term)	LCI[d]	UCI[e]	*p*-value
	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
	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[v]			0.542	-1.989	3.072	0.6746
	Moderate[w]			0.249	-1.907	2.405	0.8206
SF-12/36 MCS[j]							
	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

Outcome	Covariates	No. of trials, *m*	No. of participants, *AT*[b]:*UC*[c]	Estimate (interaction term)	LCI[d]	UCI[e]	*p*-value
	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
	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
EQ-5D							

Outcome	Covariates	No. of trials, *m*	No. of participants, *AT*[b]:*UC*[c]	Estimate (interaction term)	LCI[d]	UCI[e]	*p*-value
	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
	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

Outcome	Covariates	No. of trials, *m*	No. of participants, *AT*[b]:*UC*[c]	Estimate (interaction term)	LCI[d]	UCI[e]	*p*-value
QALY[k]							
	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

a Mixed effects models with intercept, trials and interaction between treatments and trials as random effects, and covariate and interaction between covariates; b AT, number of patients in the intervention arm (active physical, passive physical, psychological, or combination); c UC, number of patients in the control arm (usual care/GP or sham); d LCI, lower limit of the 95% confidence interval; e UCI upper limit of the 95% confidence interval; f FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; g RMDQ, Roland Morris disability questionnaire; h Obtained from either visual analogue scale or pain intensity score of chronic pain grade scale (see *Section 6.3.3.3*); i PCS, physical component scale of SF-12/36; j MCS, mental component scale of SF-12/36; k QALY, quality-adjusted life year; l estimate of the treatment effect for male was less as opposed to female; m estimate of the treatment effect for participants with SF-12/36 PCS score lower than general norm (<50) was greater as opposed to those with score at or above the general norm (≥50); n estimate of the treatment effect for participants with low risk of anxiety was less as opposed to those with the high risk; o estimate of the treatment effect for participants with moderate risk of anxiety was greater as opposed to those with the high risk; p estimate of the treatment effect for participants with positive attitude of catastrophising (low catastrophising score) was greater as opposed to those with the negative attitude (high catastrophising score); q estimate of the treatment effect for participants with moderate attitude of catastrophising was greater

Outcome	Covariates	No. of trials, *m*	No. of participants, *AT*[b]:*UC*[c]	Estimate (interaction term)	LCI[d]	UCI[e]	*p*-value

as opposed to those with the negative attitude; r estimate of the treatment effect for participants with positive attitude of coping strategy (high coping score) was greater as opposed to those with the negative attitude (low coping score); s estimate of the treatment effect for participants with moderate attitude of coping strategy was greater as opposed to those with the negative attitude; t estimate of the treatment effect for participants with positive belief (low fear avoidance) of fear avoidance belief was greater as opposed to those with the negative attitude; u estimate of the treatment effect for participants with moderate belief of fear avoidance was greater as opposed to those with the negative attitude; v estimate of the treatment effect for participants with least general somatic symptoms was greater as opposed to those with more general somatic symptoms; w estimate of the treatment effect for participants with moderate general somatic symptoms was greater as opposed to those with more general somatic symptoms.

.

6.6.3.1 RMDQ

There were up to eight trials with RMDQ short-term outcome and the explanatory covariates provided by them were age, sex, RMDQ, average pain, PCS, MCS, EQ-5D, anxiety level, depression level, catastrophising, coping strategy and fear avoidance at baseline. Seven trials 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 over the short-term where those with either positive or moderate attitude (lower fear avoidance score) had greater treatment effect than those with negative attitude (high fear avoidance score). Although the covariate catastrophising was not statistically significant (p = 0.236) in predicting the change of RMDQ at 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 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.

6.6.3.2 Pain

Ten trials provided an average pain short-term outcome. The list of covariates that were considered in the analysis of covariance were age, sex, RMDQ, 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 at 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 negative attitude. However, those with low risk of anxiety had less treatment effect as opposed to those with high risk of anxiety. Similarly, those with positive attitude of coping strategy had less treatment effect than those with negative strategy. 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 at p = 0.1451. The estimated treatment effect decreased as PCS increased, that is, as participant's physical functioning score got worse, they had greater treatment effect (p = 0.1587). The interaction term between treatment and MCS was also weakly statistically significant (*p* = 0.1677) where participants with higher (better) mental component score 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.

6.6.3.3 MCS and PCS

There were six trials 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 (<50) had worse treatment effect and this was possibly statistically significant at *p* = 0.0504. Also, age and PCS at baseline were significant where those who were younger and those with substantial physical limitations had larger treatment effect. Therefore, age, PCS and MCS scores 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 at short-term. Those with higher physical disability and more psychologically distress had a greater treatment effect. Therefore, both FFbHR and MCS scores at baseline were included for the prediction rule analyses for the change of SF-12/36 MCS at short-term.

6.6.3.4 EQ-5D

Four trials provided health utility measured by EQ-5D at short-term. The covariates examined in the analysis of covariance were age, sex, RMDQ, average pain, PCS, MCS, EQ-5D, anxiety level, depression level and fear avoidance. Of these, seven of them were statistically or weakly significant in predicting the change of EQ-5D at short-term and they are sex, RMDQ, average pain, PCS, MCS, anxiety level and fear avoidance at baseline. Female had greater treatment effect (*p* = 0.1543) and so were those with worse physical disability (RMDQ, *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 larger treatment effect. Therefore, these were considered for the prediction rule analyses.

6.6.3.5 QALY

There were up to six trials with QALY data. Age, and baseline scores of RMDQ and PCS were possibly statistically significant in moderating QALY. 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 by treatment interaction was significant (*p*=0.13) at our pre-specified level of 0.2. The coefficient of 0.003 was positive. The scale on the RMDQ 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 (*Chapter 9*). The coefficient of PCS by treatment interaction was -0.001 (*p*=0.12). The negative coefficient indicates 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).

6.6.3.6 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 only conclude that overall; that back pain disability moderates effect size on back pain disability outcomes (FFbHR moderates FFbHR) that physical state and back pain moderate effect size on physical outcomes, (PCS and FFbHR moderate PCS), that psychological state moderates effect size on psychological outcomes (MCS moderates MCS), that overall psychological state and anxiety moderate effect size on quality of life (PCS and anxiety moderate EQ-5D), and that 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 test 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 different methods we have applied to addressing this problem. In particular we have been 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 which was part of this programme grant.[141] The other methods are described in later chapters (see *Chapters 8-10).*

7.1 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).[94, 96] 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 participant data (IPD) meta-analyses 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 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 dataset. The method then searches all possible binary splits for every covariate to find the best split that maximises some splitting criterion. Suppose sex is identified as the first best split. The method therefore splits the root node using the sex covariate to form two child nodes; 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 since 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 ≤ 50 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.

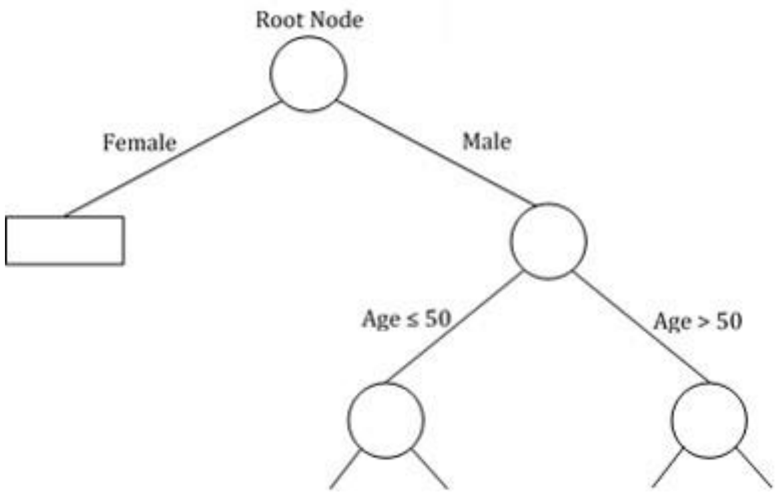


Figure 20 Example of a tree structure.

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 whilst the IPD-SIDES method is driven by identifying the spilt that maximises the overall treatment benefit in one of the subgroups formed from the split.

7.2 IPD-INTERACTION TREE (IPD-IT)

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 either result in tree structure with just the root node, that is, no moderators identified, or a larger tree structure that stems from the root node, that is, some moderators identified. In the latter case, the subgroups identified by the final selected tree are interpreted using its terminal nodes.

1. 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*. Since we are now applying this method to individual patient data 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, i.e. fitting a fixed-effect model.[141] 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 dataset
- 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 to 3 until either a full tree is grown or until some stopping criterion is satisfied, for example, minimum number of observations in a node is 30.

2. 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 sub-trees from which the optimal final sub-tree will be chosen. A more detailed description of the pruning procedure can be found elsewhere.[94, 141, 142]

3. Selecting the best tree

Once the sequence of sub-trees 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 sub-trees, 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, for validating the trees.[143] As a guideline, LeBlanc and Crowley suggested that around 25 to 100 bootstrap samples is sufficient. The sub-tree 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.

7.3 IPD-SUBGROUP IDENTIFICATION BASED ON A DIFFERENTIAL EFFECT SEARCH (IPD-SIDES)

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 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.

1. 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 dataset
- Iteration:
    - Step 1 - Evaluate the splitting criterion for all splits of every covariate, excluding any covariates already 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, provided 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 to 3 until either a pre-specified maximum number of levels is reached or if no more splits can be formed i.e. 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 dataset. 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 that 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 to 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-analyses setting. For this reason, a new splitting criterion was proposed which uses the same fixed effect model described earlier for the IPD-IT method but instead uses the p-value of the interaction effect where a smaller p-value is indicative of a larger interaction effect. If a covariate has more than two distinct cut-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 where covariates with a larger number of splits have a greater probability of being chosen as the splitting variable.[144, 145] The method adjusts the p-value by applying a Sidak-based multiplicity adjustment as described in the original SIDES method paper.[96]

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):

$$p_c \leq \gamma \cdot p_p \tag{2}$$

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; where the recommended value is three. This means that any identified subgroups will at most be defined by L covariates; 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 where 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.[96] Hence, once the relative improvement parameter values are in place, a child node is only retained provided its treatment effect p-value is less than or equal to the right hand side of the continuation criterion.

2. 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.[96] Comparing the unadjusted p-value to the adjusted p-value gives a good indication as to whether the identified subgroups are spurious or not.

7.4 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, Hannover functional ability questionnaire for measuring back-pain related functional limitations (FFbHR), mental component scale of SF-12/36 (MCS), physical component scale of SF-12/36 (PCS) and Roland Morris Disability Questionnaire (RMDQ).

In addition to the above outcome measures, we also looked at the quality adjusted life years (QALYs) health economics outcome. This analysis provides proof of principal that the analytical techniques are robust when used with real data rather than simply in the simulated datasets in which we originally developed our techniques.[141]

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 dataset. 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 (*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 pre-specified 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 i.e. 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 three with a restriction of $p \leq 0.20$ placed on the splitting criterion. This is the same constraint we set in the identification of 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 one in steps of 0.2 at the first level and then from zero to one 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 1,000 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% confidence interval (CI). All analyses were performed using R version 3.0.3.

7.5 RESULTS

7.5.1 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.

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 UK BEAM[b] (*n* = 910), BeST[c] (*n* = 467)	Age, sex, anxiety, fear avoidance, MCS, PCS, average pain and RMDQ at baseline
EQ-5D	*m* = 2; *n* = 1339 UK BEAM (*n* = 883), BeST (*n* = 456)	Age , sex, anxiety, fear avoidance, MCS, PCS RMDQ and average pain at baseline
FFbHR[d]	*m* = 3; *n* = 3718 Brinkhaus[e] (*n* = 284), Haake[f] (*n* = 1110), Witt[g] (*n* = 2324)	Age, sex, PCS, FFbHR and MCS at baseline
MCS[h]	*m* = 3; *n* = 3630 Brinkhaus (*n* = 281), Haake (*n* = 1110), Witt (*n* = 2239)	Age, sex, FFbHR, MCS and PCS at baseline
PCS[i]	*m* = 6; *n* = 5208 UK BEAM (*n* = 893), BeST (*n* = 470), Brinkhaus (*n* = 281), Haake (*n* = 1110), Witt (*n* = 2248), YACBAC[j] (*n* = 206)	Age, sex, MCS and PCS at baseline
RMDQ[k]	*m* = 7; *n* = 2564 UK BEAM (*n* = 951), BeST (*n* = 488), Hancock[l] (*n* = 235), Pengel[m] (*n* = 236), Smeets[n] (*n* = 212), VK BIA[o] (*n* = 229), VK SC2[p] (*n* = 213)	Age, sex, fear avoidance and RMDQ at baseline
QALY[q]	*m* = 4; *n* = 1514 UK BEAM (*n* = 728), BeST (*n* = 468), Smeets (*n* = 151), YorkBP[r] (*n* = 167)	Age and RMDQ at baseline

a Change from baseline to short-term follow-up (between two and three months post-randomisation or entry to the trial); b UK BEAM (Exercise, spinal manipulation, combined, best care); c BeST (Cognitive behavioural approach, control); d FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; e Brinkhaus (Acupuncture, minimal acupuncture, waiting list); f Haake (Verum acupuncture, sham acupuncture, conventional therapy); g Witt (Acupuncture, control); h MCS, mental component scale of SF-12/36; i PCS, physical component scale of SF-12/36; j YACBAC (Traditional acupuncture, usual care); k RMDQ, Roland Morris disability questionnaire; l Hancock (Spinal manipulation, placebo spinal manipulation, advice); m Pengel (Exercise, sham exercise, advice, sham advice); n Smeets (Active physical therapy, cognitive behavioural treatment, combined treatment, waiting list); o Von Korff BIA (Brief individualised programme, usual care); Von Korff SC2[p] (Self-care, usual care); q QALY, quality adjusted life year which was measured over one year of follow-up using the area under the curve method; r York BP (Exercise, control)

7.5.1.1 Subgroups identified by the IPD-IT method

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

7.5.1.2 Subgroups identified by the IPD-SIDES 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 vs control/placebo (see *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.

Table 19 Candidate subgroups identified by the IPD-SIDES method for the intervention vs control/placebo comparison[a]

Subgroups	*n*	Treatment effect (95% confidence interval, CI)	Interaction effect	Unadjusted *p*-value
Outcome: short-term FFbHR[b]				
Overall treatment effect (95% CI): 8.93 (7.81, 10.05)				
Candidate 1				
FFbHR ≤ 54.2	1,709	11.31 (9.38, 13.23)	4.69	< 0.001
FFbHR > 54.2	2,009	6.62 (5.46, 7.78)		
Candidate 2				
FFbHR ≤ 54.2 AND Age ≤ 60	1,043	13.17 (10.56, 15.77)	5.03	0.019
FFbHR ≤ 54.2 AND Age > 60	666	8.14 (5.47, 10.80)		
Candidate 3				
FFbHR ≤ 54.2 AND Age ≤ 66	1,367	12.26 (10.06, 14.46)	5.14	0.043
FFbHR ≤ 54.2 AND Age > 66	342	7.12 (3.42, 10.82)		
Outcome: short-term MCS[c]				
Overall treatment effect (95% CI): 2.61 (1.92, 3.29)				
Candidate 1				
MCS ≤ 54.4	2,541	3.46 (2.62, 4.30)	2.62	0.002
MCS > 54.4	1,089	0.84 (0.01, 1.67)		

Subgroups	***n***	**Treatment effect (95% confidence interval, CI)**	**Interaction effect**	**Unadjusted *p*-value**
Outcome: short-term PCS[d]				
Overall treatment effect (95% CI): 3.48 (3.01, 3.96)				
Candidate 1				
MCS > 50.9	2,082	4.09 (3.32, 4.87)	0.97	0.033
MCS ≤ 50.9	3,126	3.12 (2.54, 3.71)		
Candidate 2				
MCS > 50.9 AND Sex = Female	1,125	4.72 (3.67, 5.78)	1.38	0.097
MCS > 50.9 AND Sex = Male	957	3.34 (2.20, 4.48)		
Candidate 3				
MCS > 50.9 & PCS ≤ 43.2	1,666	4.62 (3.75, 5.49)	2.61	0.020
MCS > 50.9 & PCS > 43.2	416	2.01 (0.69, 3.33)		
Candidate 4				
MCS > 50.9 & PCS ≤ 40.0	1,457	4.89 (3.96, 5.82)	2.61	0.007
MCS > 50.9 & PCS > 40.0	625	2.28 (1.12, 3.44)		

a The first row of each candidate subgroup is the selected subgroup with enhanced treatment effect; b FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations ranging from 0-100 where a lower score represents greater disability; c MCS, mental component scale of SF-12/36 ranging from 0-100 where a lower score represents worse mental functioning; d PCS, physical component scale of SF-12/36 ranging from 0-100 where a lower score represents worse physical functioning.

7.5.1.2.1 Short-term Hannover functional ability questionnaire for measuring back-pain related functional limitations (FFbHR) 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% confidence interval, CI, 7.81 to 10.05). Three candidate subgroups with enhanced treatment effect were identified by the IPD-SIDES procedure. Those with baseline FFbHR ≤ 54.2 had a treatment effect of 11.31 (95% CI, 9.38 to 13.23), those with baseline FFbHR ≤ 54.2 and age ≤ 60 had a treatment effect of 13.17 (95% CI, 10.56 to 15.77) and those with FFbHR ≤ 54.2 and age ≤ 66 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.*

7.5.1.2.2 Short-term mental component scale of SF-12/36 (MCS) 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 ≤ 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.*

7.5.1.2.3 Short-term physical component scale of SF-12/36 (PCS) 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 > 50.9 had a treatment effect of 4.09 (95% CI, 3.32 to 4.87), those with baseline MCS > 50.9 and who are female had a treatment effect of 4.72 (95% CI, 3.67 to 5.78), those with baseline MCS > 50.9 and baseline PCS ≤ 43.2 had a treatment effect of 4.62 (95% CI, 3.75 to 5.49) and finally those with baseline MCS > 50.9 and baseline PCS ≤ 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.

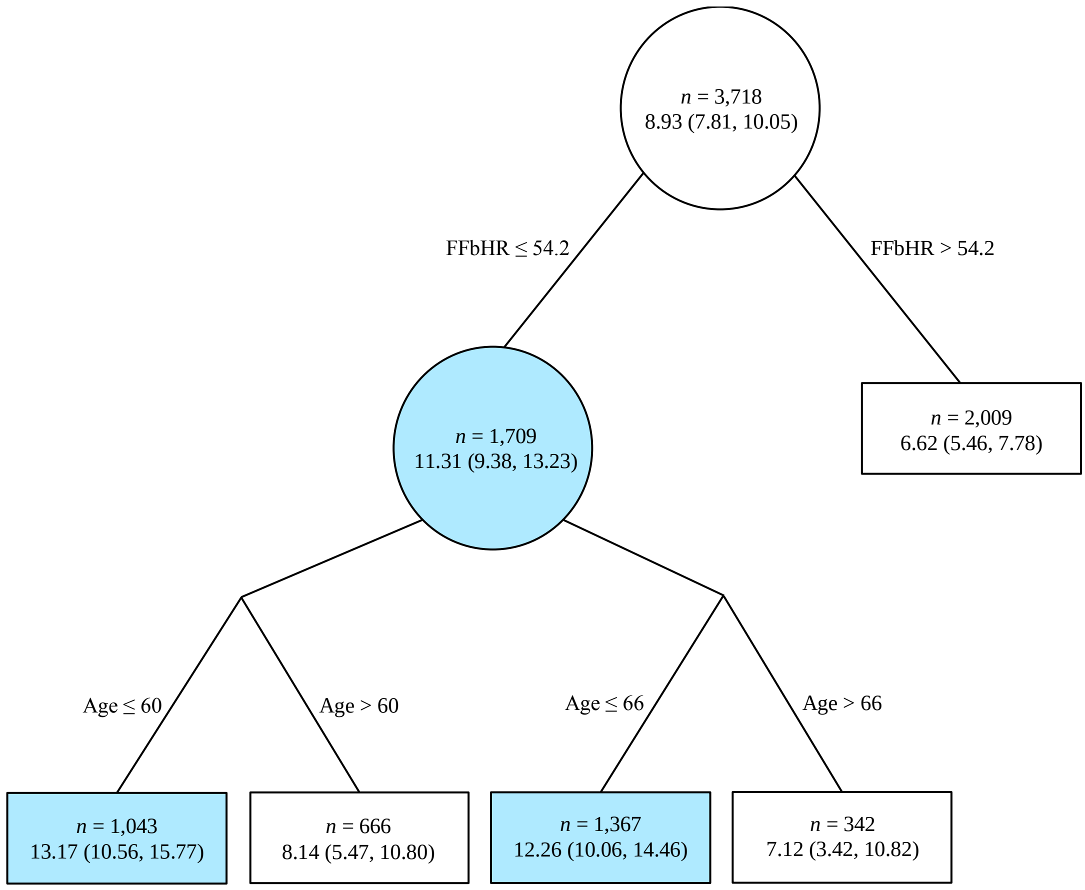


Figure 21 Candidate subgroups identified (shaded blue) by the IPD-SIDES method when applied to change from baseline to short-term Hannover functional ability questionnaire for measuring back-pain related functional limitations (FFbHR – range 0-100; lower score implies greater disability) outcome for the intervention against control/placebo comparison. Results presented as treatment effect (95% confidence interval) in each node.

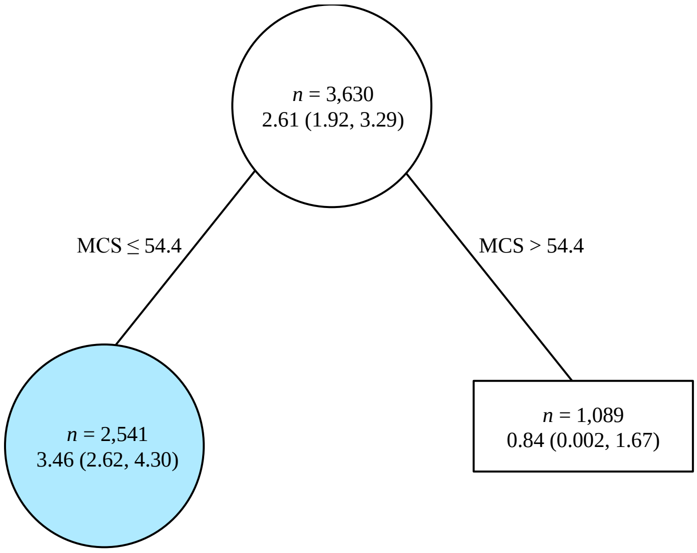


Figure 22 Candidate subgroup identified (shaded blue) by the IPD-SIDES method when applied to change from baseline to short-term SF-12/36 MCS outcome (range 0-100; lower score implies worse mental functioning) for the intervention against control/placebo comparison. Results presented as treatment effect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

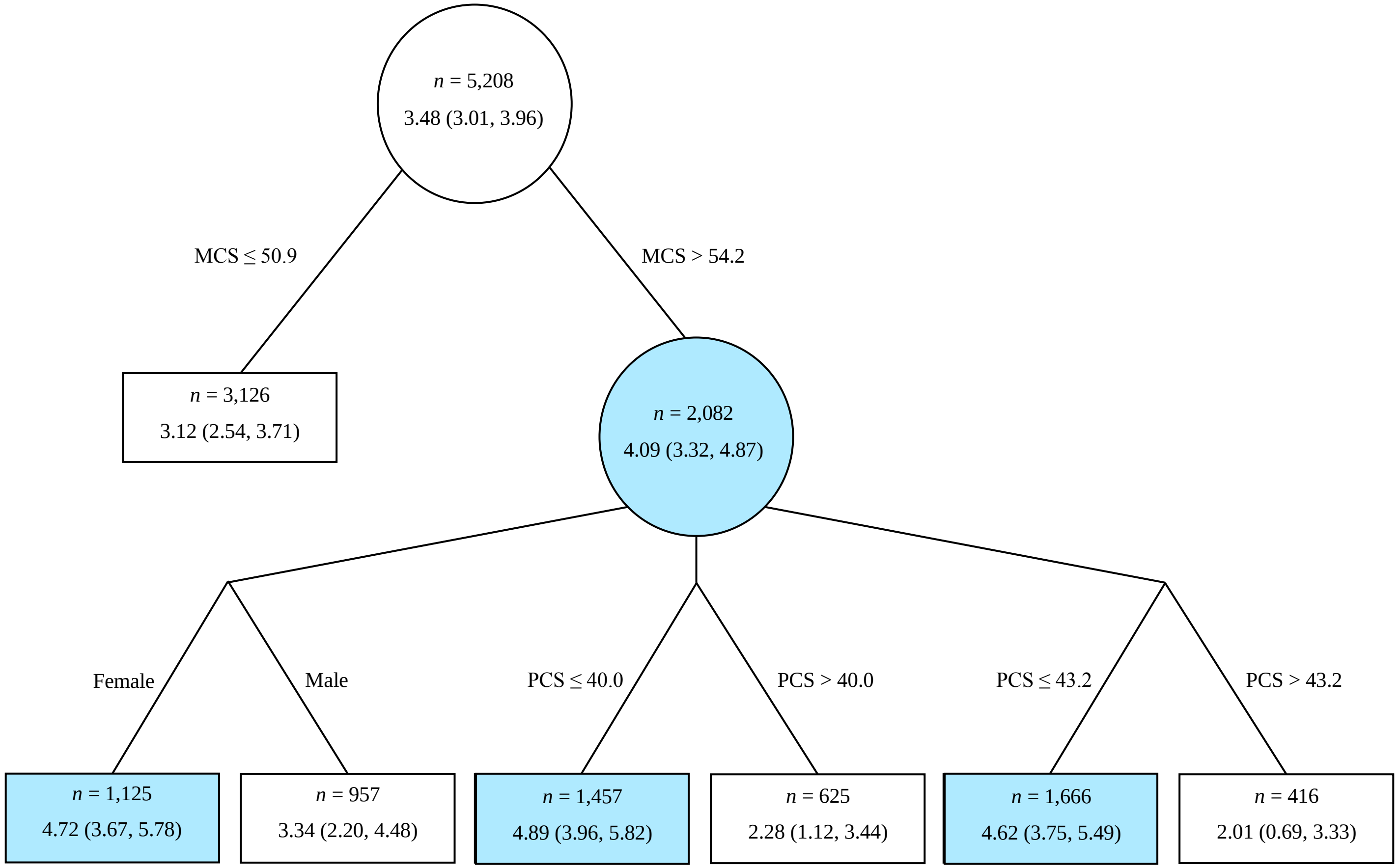


Figure 23 Candidate subgroups identified (shaded blue) by the IPD-SIDES method when applied to change from baseline to short-term SF-12/36 PCS outcome (range 0-100; lower score implies worse physical functioning) for the intervention against control/placebo comparison. Results presented as treatment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

7.5.2 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.

7.5.2.1 Subgroups identified by the IPD-IT 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.

7.5.2.2 Subgroups identified by the IPD-SIDES 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 (see *Table 21*), psychological against non-active usual care (see *Table 22*) and sham against non-active usual care (see *Table 23*). No candidate subgroups were identified for the active physical against non-active usual care comparison.

7.5.2.2.1 Passive physical vs non-active usual care results

Short-term FFbHR 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 baseline FFbHR ≤ 54.2 had a treatment effect of 12.86 (95% CI, 10.81 to 14.91), those with baseline FFbHR ≤ 54.2 and age ≤ 57 had a treatment effect of 15.86 (95% CI, 12.80 to 18.92), those with FFbHR ≤ 54.2 and age ≤ 53 had a treatment effect of 16.67 (95% CI, 13.16 to 20.18) and those with baseline FFbHR ≤ 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 SF-12/36 MCS 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 baseline MCS ≤ 54.3 had a treatment effect of 3.76 (95% CI, 2.97 to 4.55), those with MCS ≤ 54.3 and PCS ≤ 43.9 had a treatment effect of 4.27 (95% CI, 3.39 to 5.15) and those with MCS ≤ 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 SF-12/36 PCS 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 baseline PCS ≤ 43.6 had a treatment effect of 4.39 (95% CI, 3.78 to 4.99), those with baseline PCS ≤ 43.6 and age ≤ 44 had a treatment effect of 5.35 (95% CI, 4.21 to 6.49), those with baseline PCS ≤ 37.8 had a treatment effect of 4.61 (95% CI, 3.90 to 5.32), those with PCS ≤ 37.8 and age ≤ 62 had a treatment effect of 5.08 (95% CI, 4.21 to 5.94), those with baseline PCS ≤ 37.8 and MCS > 44.0 had a treatment effect of 5.48 (95% CI, 4.55 to 6.41), those with PCS ≤ 37.8 and MCS > 51.8 had a treatment effect of 5.77 (95% CI, 4.63 to 6.91), those with PCS ≤ 37.8, MCS > 51.8 and are 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 PCS ≤ 40.3 and MCS > 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*

7.5.2.2.2 Psychological vs non-active usual care results

Short-term RMDQ 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 baseline RMDQ > 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 when compared to usual care control*

7.5.2.2.3 Sham vs non-active usual care results

Short-term SF-12/36 MCS 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 at baseline had a treatment effect of 3.42 (95% CI, 1.80 to 5.04) and those with baseline PCS ≤ 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 suggest that younger people and those with worse physical status at baseline have a greater benefit from sham treatment on psychological distress when compared to a usual care control*

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

	Short-term outcome measures									
	FFbHR[a]		**RMDQ[b]**		**MCS[c]**		**PCS[d]**		**QALY[e]**	
Comparison	**Trials***	**Variables**	**Trials***	**Variables**	**Trials***	**Variables**	**Trials***	**Variables**	**Trials***	**Variables**
Active vs non-active usual care	-	-	*m* = 2; *n* = 576 UK BEAM (*n* = 421), Smeets (*n* = 155)	Fear avoidance, age, sex, RMDQ, average pain today, EQ5D, HADS anxiety, HADS depression	-	-	-	-	*m* = 2; *n* = 496 UK BEAM (*n* = 329), YorkBP (*n* = 167)	Age, RMDQ

	Short-term outcome measures									
	FFbHR[a]		RMDQ[b]		MCS[c]		PCS[d]		QALY[e]	
Comparison	**Trials***	**Variables**	**Trials***	**Variables**	**Trials***	**Variables**	**Trials***	**Variables**	**Trials***	**Variables**
Passive vs non-active usual care	*m* = 3; *n* = 3272 Brinkhaus (*n* = 214), Haake (*n* = 734), Witt (*n* = 2324)	Age, PCS, FFbHR, sex, MCS	-	-	*m* = 5; *n* = 3879 UK BEAM (*n* = 479), Brinkhaus (*n* = 212), Haake (*n* = 734), Witt (*n* = 2248), YACBAC (*n* = 206)	MCS, age, sex, PCS	*m* = 5; *n* = 3879 UK BEAM (*n* = 479), Brinkhaus (*n* = 212), Haake (*n* = 734), Witt (*n* = 2248), YACBAC (*n* = 206)	Age, MCS, PCS, sex	*m* = 3; *n* = 1209 UK BEAM (*n* = 379), Haake (*n* = 716), YACBAC (n=114)	Age, PCS
Psychological vs non-active usual care	-	-	*m* = 3; *n* = 928 BeST (*n* = 487), VKBIA (*n* = 229), VKSC2 (*n* = 212)	Fear avoidance, age, sex, RMDQ, average pain today	-	-	-	-	-	-

	Short-term outcome measures									
	FFbHR[a]		RMDQ[b]		MCS[c]		PCS[d]		QALY[e]	
Comparison	**Trials***	**Variables**	**Trials***	**Variables**	**Trials***	**Variables**	**Trials***	**Variables**	**Trials***	**Variables**
Sham vs non-active usual care	*m* = 2; *n* = 881 Brinkhaus (*n* = 144), Haake (*n* = 737)	Age, PCS, FFbHR, sex, MCS	-	-	*m* = 2; *n* = 879 Brinkhaus (*n* = 142), Haake (*n* = 737)	MCS, age, sex, PCS	*m* = 2; *n* = 879 Brinkhaus (*n* = 142), Haake (*n* = 737)	Age, MCS, PCS, sex	-	-

a FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; b RMDQ, Roland Morris disability questionnaire; c MCS, mental component scale; d PCS, physical component scale; e QALY, quality adjusted life years;

* UK BEAM (Exercise, spinal manipulation, combined, best care); Smeets (Active physical therapy, cognitive behavioural treatment, combined treatment, waiting list); York BP (Exercise, control); Brinkhaus (Acupuncture, minimal acupuncture, waiting list); Haake (Verum acupuncture, sham acupuncture, conventional therapy); Witt (Acupuncture, control); YACBAC (Traditional acupuncture, usual care); BeST (Cognitive behavioural approach, control); Von Korff BIA (Brief individualised programme, usual care); Von Korff SC2 (Self-care, usual care);

Table 21 Candidate subgroups identified by the IPD-SIDES method for the passive physical vs non-active usual care comparison.[a, b]

Subgroups	*n*	Treatment effect (95% confidence interval, CI)	Interaction effect	Unadjusted *p*-value
Outcome: short-term FFbHR				
Overall treatment effect (95% CI): 9.95 (8.80, 11.11)				
Candidate 1				
FFbHR ≤ 54.2	1,424	12.86 (10.81, 14.91)	5.45	<0.001
FFbHR > 54.2	1,848	7.41 (6.23, 8.59)		
Candidate 2				
FFbHR ≤ 54.2 AND Age ≤ 57	731	15.86 (12.80, 18.92)	6.63	0.002
FFbHR ≤ 54.2 AND Age > 57	693	9.23 (6.64, 11.82)		
Candidate 3				
FFbHR ≤ 54.2 AND Age ≤ 53	571	16.67 (13.16, 20.18)	6.85	0.001
FFbHR ≤ 54.2 AND Age > 53	853	9.83 (7.43, 12.22)		
Candidate 4				
FFbHR ≤ 41.7	792	15.03 (12.06, 18.01)	6.71	<0.001
FFbHR > 41.7	2,480	8.32 (7.19, 9.45)		
Outcome: short-term MCS				
Overall treatment effect (95% CI): 2.96 (2.31, 3.61)				
Candidate 1				
MCS ≤ 54.3	2,714	3.76 (2.97, 4.55)	2.82	<0.001
MCS > 54.3	1,165	0.93 (0.10, 1.76)		
Candidate 2				
MCS ≤ 54.3 AND PCS ≤ 43.9	2,171	4.27 (3.39, 5.15)	2.43	0.019
MCS ≤ 54.3 AND PCS > 43.9	543	1.85 (0.11, 3.59)		
Candidate 3				

Subgroups	***n***	**Treatment effect (95% confidence interval, CI)**	**Interaction effect**	**Unadjusted *p*-value**
MCS ≤ 51.3	2,327	3.83 (2.96, 4.70)	2.57	<0.001
MCS > 51.3	1,552	1.26 (0.52, 1.99)		
Outcome: short-term PCS				
Overall treatment effect (95% CI): 4.10 (3.56, 4.63)				
Candidate 1				
PCS ≤ 43.6	3,103	4.39 (3.78, 4.99)	1.61	0.013
PCS > 43.6	776	2.77 (1.87, 3.67)		
Candidate 2				
PCS ≤ 43.6 AND Age ≤ 44	942	5.35 (4.21, 6.49)	1.45	0.040
PCS ≤ 43.6 AND Age > 44	2,161	3.90 (3.20, 4.60)		
Candidate 3				
PCS ≤ 37.8	2,326	4.61 (3.90, 5.32)	1.23	0.025
PCS > 37.8	1,553	3.37 (2.66, 4.09)		
Candidate 4				
PCS ≤ 37.8 AND Age ≤ 62	1,682	5.08 (4.21, 5.94)	1.97	0.016
PCS ≤ 37.8 AND Age > 62	644	3.11 (1.94, 4.28)		
Candidate 5				
PCS ≤ 37.8 AND MCS > 44.0	1,396	5.48 (4.55, 6.41)	1.80	0.011
PCS ≤ 37.8 AND MCS ≤ 44.0	930	3.68 (2.64, 4.71)		
Candidate 6				
PCS ≤ 37.8 AND MCS > 51.8	932	5.77 (4.63, 6.91)	1.78	0.012
PCS ≤ 37.8 AND MCS ≤ 51.8	1,394	3.99 (3.11, 4.87)		
Candidate 7				
PCS ≤ 37.8 AND MCS > 51.8 AND Sex = Female	520	6.64 (5.12, 8.16)	1.73	0.167

Subgroups	*n*	Treatment effect (95% confidence interval, CI)	Interaction effect	Unadjusted *p*-value
PCS ≤ 37.8 AND MCS > 51.8 AND Sex = Male	412	4.91 (3.17, 6.65)		
Candidate 8				
PCS ≤ 40.3	2,715	4.51 (3.85, 5.16)	1.61	0.006
PCS > 40.3	1,164	2.90 (2.11, 3.68)		
Candidate 9				
PCS ≤ 40.3 AND MCS > 51.5	1,086	5.43 (4.37, 6.48)	1.38	0.042
PCS ≤ 40.3 AND MCS ≤ 51.5	1,629	4.05 (3.24, 4.85)		

a The baseline FFbHR score ranges from 0-100 where a lower score represents greater disability. The baseline MCS and PCS scores range from 0-100 where 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, 1.91)				
Candidate 1				
RMDQ > 4	697	1.72 (1.12, 2.31)	1.07	0.038
RMDQ ≤ 4	231	0.65 (-0.11, 1.40)		

a The baseline RMDQ score ranges from 0-24 where 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, 4.04)				
Candidate 1				
Age ≤ 65	705	3.42 (1.80, 5.04)	4.32	0.019
Age > 65	174	-0.90 (-4.16, 2.35)		
Candidate 2				
PCS ≤ 42.0	791	3.10 (1.55, 4.65)	4.99	0.043
PCS > 42.0	88	-1.89 (-6.07, 2.28)		

a The baseline PCS score ranges from 0-100 where a lower score represents worse physical functioning; b The first row of each candidate subgroup is the selected subgroup with enhanced treatment effect

CHAPTER 8 – METHODOLOGY AND STATISTICAL DEVELOPMENTS 2: SUBGROUP IDENTIFICATION USING AN ADAPTIVE REFINEMENT BY DIRECTED PEELING (ARDP) ALGORITHM

8.1 BACKGROUND

The adaptive risk group refinement introduced by LeBlanc *et al*. 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 which are easy to interpret.[146] The approach is based on a so-called 'adaptive refinement by directed peeling' (ARDP) algorithm. Starting with the whole data set the algorithms peels off fractions of the data in a series of locally optimal steps optimising a prognostic indicator (for example, median survival in the paper by LeBlanc *et al*..[146] We aim to identify subgroups of participants that 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*. 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 interaction tree (IT) and subgroup identification based on a differential effect search (SIDES) methods (see *Chapter 7*). In the following we describe the modified ARDP algorithm for individual participant data meta-analysis.

8.2 ADAPTIVE REFINEMENT BY DIRECTED PEELING IN IPD META-ANALYSIS (ARDP-MA)

The ARDP-MA algorithm to construct a region that predicts the best or worst response to treatment consists of the following steps:

1. In order 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 $B_j^m, j = 1, \ldots, L$. 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 to 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 to 5 until the size of the remaining region is not smaller than r.

Figure 24 is a schematic illustrating 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. Restricting the number of covariates to four we could do far more extensive searches by considering all 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 endpoints as we only assume that appropriate regression models can be fitted modelling the outcome. For instance, Gaussian linear models could be applied to continuous outcomes, logistic regression to binomial outcomes, Cox 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.[146]

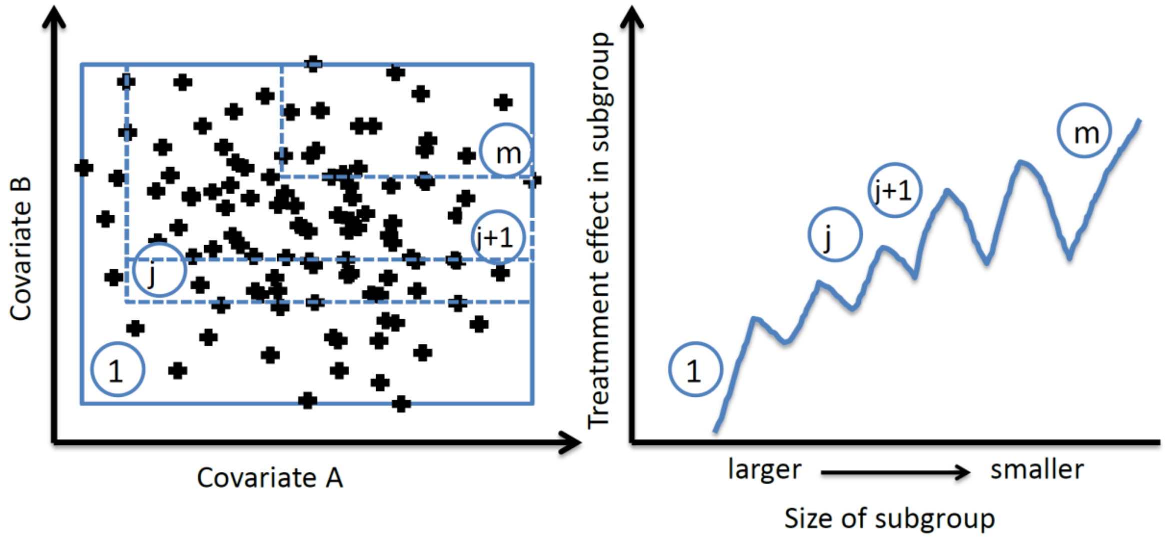


Figure 24 Schematic of the ARDP algorithm to identify subgroups of treatment responders. Here the subgroups are defined by thresholds for the two covariates A and B.

8.3 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 *Section 7.4*, two sets of analyses were performed. The first one was to confirm proof of concept where 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, Hannover functional ability questionnaire for measuring back-pain related functional limitations (FFbHR), mental component scale of SF-12/36 (MCS), physical component scale of SF-12/36 (PCS) and Roland

Morris Disability Questionnaire (RMDQ). 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.

8.4 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.

8.4.1 ANALYSIS 1 OVERALL COMPARISON TREATMENT VS. 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.

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 (n=926), BeST (*n* = 498), Haake (*n* = 1,110)	Age, average pain, PCS and MCS at baseline
EQ-5D	*m* = 2; *n* = 1,365 UK BEAM (*n* = 890), BeST (*n* = 475)	RMDQ, average pain, PCS and MCS at baseline
FFbHR[c]	*m* = 3; *n* = 3718 Brinkhaus (n=284), Haake (*n* = 1,110), Witt (*n* = 2,324)	Age, FFbHR, PCS and MCS at baseline
MCS[d]	*m* = 3; *n* = 3,630 Brinkhaus (*n* = 281), Haake (*n* = 1,110), Witt (n=2,239)	Age, FFbHR, PCS and MCS at baseline
PCS[e]	*m* = 6; *n* = 5,208 UK BEAM (*n* = 893), BeST (*n* = 470), Brinkhaus (*n* = 281), Haake (*n* = 1,110), Witt (*n* = 2,248), YACBAC (*n* = 206)	Age, PCS and MCS at baseline
RMDQ[f]	*m* = 8; *n* = 2,675 UK BEAM (n=995), BeST (n=514), Hancock (*n* = 235), Kennedy (*n* = 40), Pengel (*n* = 236), Smeets (*n* = 212), VKBIA (*n* = 230), VKSC2 (*n* = 213)	Age and RMDQ at baseline

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 two and three months post-randomisation or entry to the trial); c FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; d MCS, mental component scale of SF-12/36; e PCS, physical component scale of SF-12/36; f RMDQ, Roland Morris disability questionnaire

8.4.1.1 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 score) seemed to gain benefit in short-term average pain.

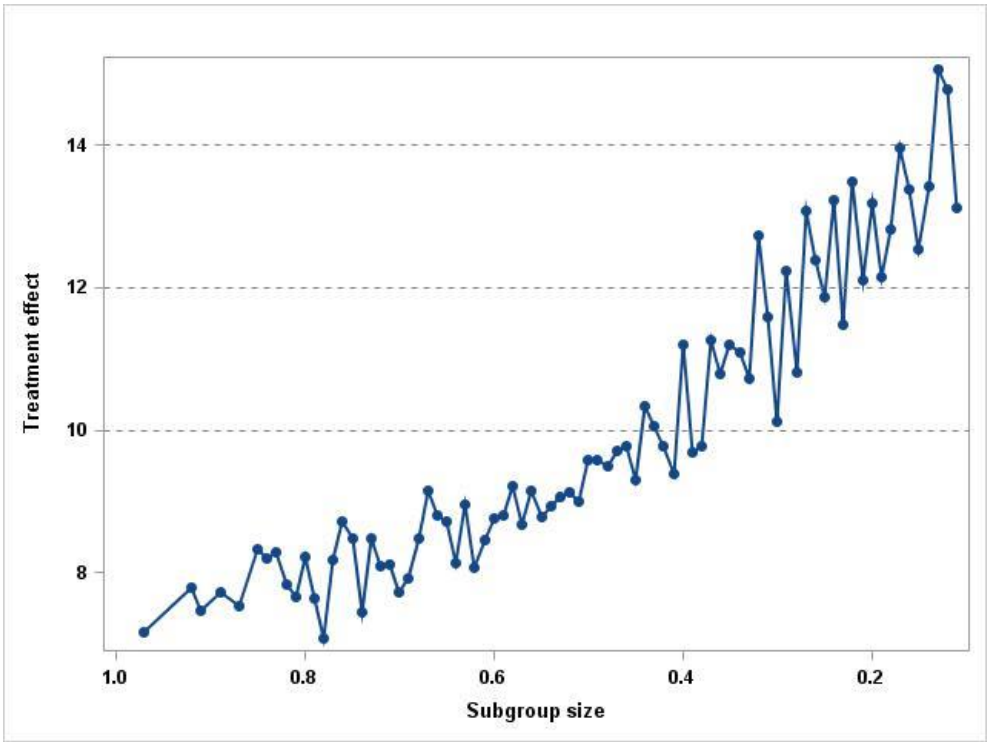


Figure 25 Trajectory plot for the treatment effect against the size of the constructed region for the average pain short-term outcome.

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

Subgroup size	Age (<)	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, physical component scale of SF-12/36; b MCS, mental component scale of SF-12/36; c for about 10.6% of the population with age < 50, average pain score > 50, SF-12/36 PCS < 33.62 and SF-12/36 MCS > 38.21, the treatment effect was 14.04.

8.4.1.2 Short-term EQ-5D outcome

Figure 26 shows the trajectory plot for the short-term outcome of health utility measured by EQ-5D. As seen in *Table 26* approximately 90% of the initial 1,365 participants (corresponding to PCS < 68 and MCS < 60, regardless of the average pain and RMDQ 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 were excluded from the analysis. There was a markedly increased of treatment effect for about 20% of the population (corresponding to PCS < 31, MCS < 72, average pain > 0 and RMDQ > 6) where the average treatment effect was about 0.160.

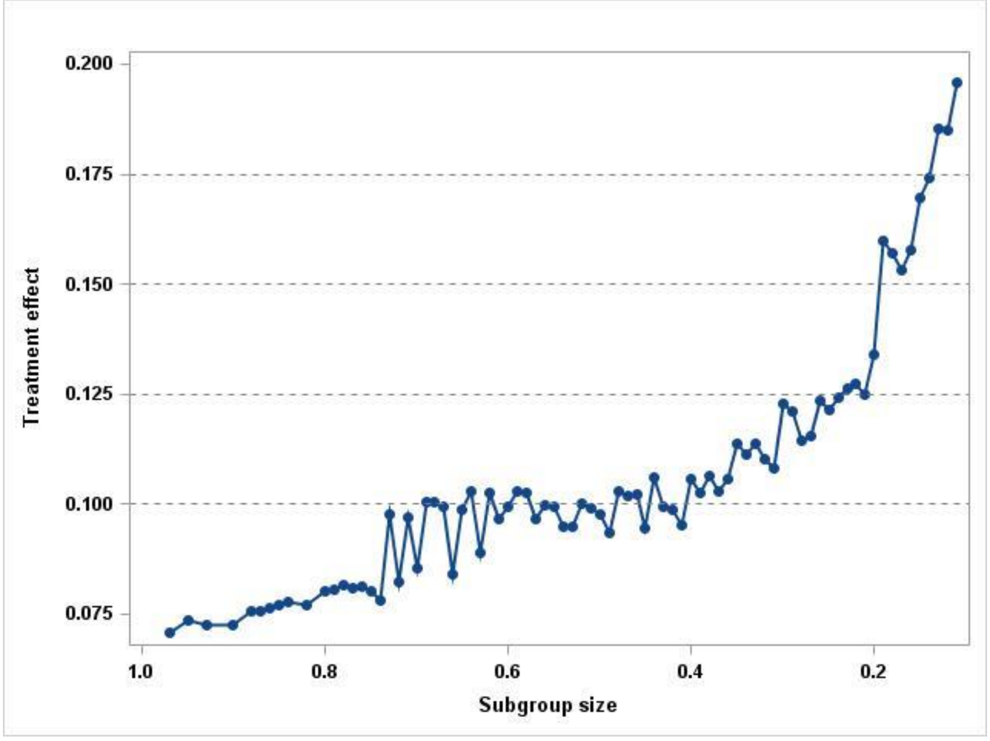


Figure 26 Trajectory plot for the treatment effect against the size of the constructed region for the EQ-5D short-term outcome

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[c] (>)	Treatment effect
0.101[d]	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

Subgroup size	PCS[a] (<)	MCS[b] (<)	Pain (>)	RMDQ[c] (>)	Treatment effect
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
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, physical component scale of SF-12/36; b MCS, mental component scale of SF-12/36; c RMDQ, Roland Morris disability questionnaire; d for about 10.1% of the population with SF-12/36 PCS < 35.66, SF-12/36 MCS < 60.35, average pain > 0 and RMDQ > 14, the treatment effect was 0.208.

8.4.1.3 Short-term FFbHR outcome

Figure 27 shows the trajectory plot for the treatment effect against the size of the constructed region for the change of FFbHR 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 high value of PCS at baseline, that is, the remaining 90% in the subgroup correspond to any age, FFBHR < 100, PCS < 48 and MCS < 72. The average treatment effect was 8.5 (see *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 < 29, PCS < 68 and MCS < 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 greater 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, FFbHR < 54, PCS < 40 and MCS < 72) would benefit greater on the disability outcome at short-term, a similar result to that observed in the IPD-SIDES Analysis 1 where participants with FFbHR ≤ 54.2 and age ≤ 66 had an enhanced treatment effect (see *7.5.1.2*). 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 younger would gain greater improvement in short-term functional ability as measured by the FFbHR.**

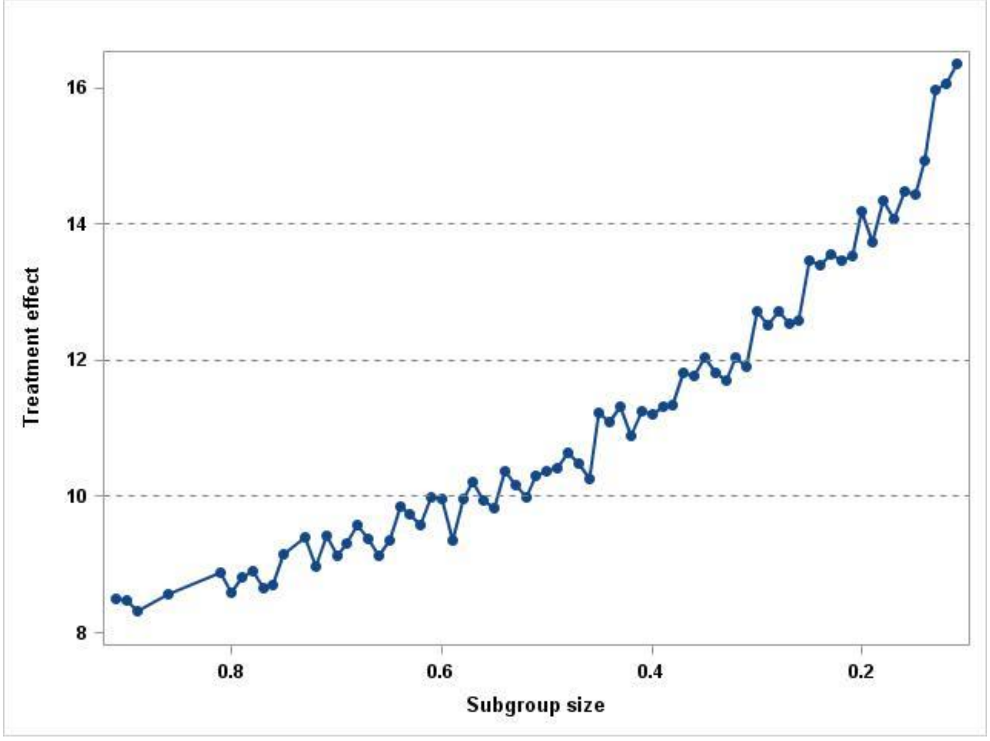


Figure 27 Trajectory plot for the treatment effect against the size of the constructed region for the FFbHR short-term outcome.

Table 27 Thresholds for selected size of subgroup for the short-term FFbHR outcome as seen in *Figure 27.*

Subgroup size	Age (<)	FFbHR[a] (<)	PCS[b] (<)	MCS[c] (<)	Treatment effect
0.102[d]	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
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

Subgroup size	Age (<)	FFbHR[a] (<)	PCS[b] (<)	MCS[c] (<)	Treatment effect
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 FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; b PCS, physical component scale of SF-12/36; c MCS, mental component scale of SF-12/36; d for about 10.2% of the population with age < 91, FFbHR < 29.17, SF-12/36 PCS < 67.75 and SF-12/36 MCS < 56.82, the treatment effect was 16.79.

8.4.1.4 Short-term SF-12/36 MCS 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, PCS and MCS. The average treatment effect of approximately 90% of the initial 3,630 participants (corresponding to age > 16, FFbHR < 100, PCS < 48 and MCS < 72) was 2.23 and this increased to 5.98 for approximately 10% of the participants (corresponding to age > 16, FFbHR < 100, PCS < 29 and MCS < 51). Approximately 55% of the participants (corresponding to age > 31, FFbHR < 63, PCS < 44 and MCS < 72) had an average treatment effect of 3 units. A smaller region consisting of 30% of the participants (corresponding to age > 54, FFbHR < 75, PCS < 44 and MCS < 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 where 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 younger would gain greater improvement in short-term psychological outcome as measured by the SF-12/36 MCS.**

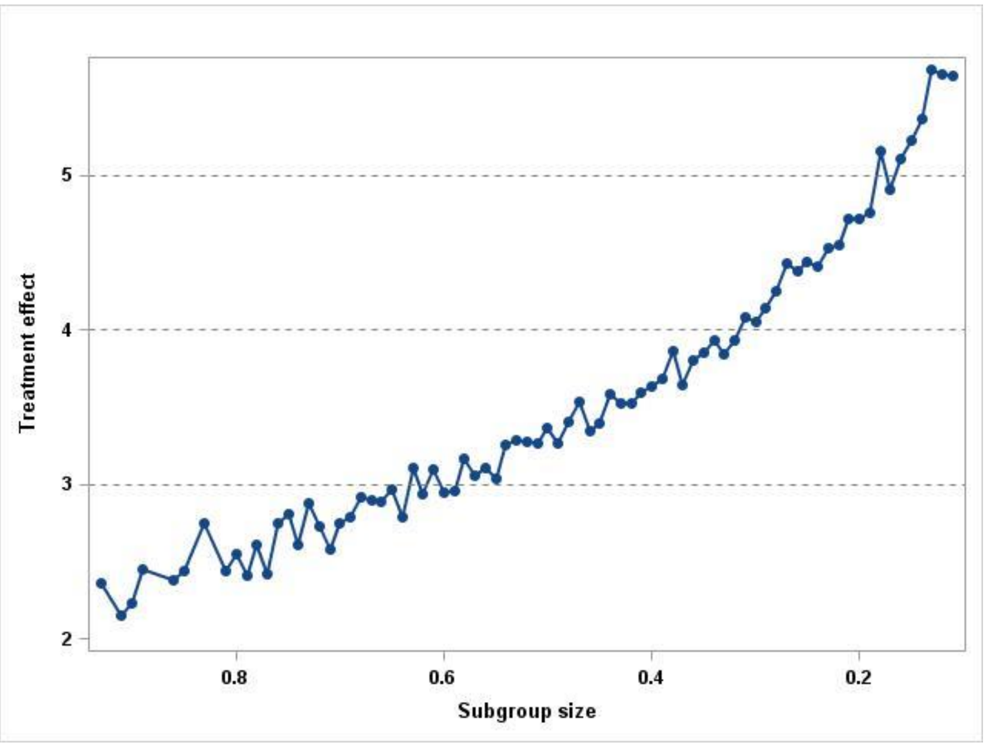


Figure 28 Trajectory plot for the treatment effect against the size of the constructed region for the SF-12/36 MCS short-term outcome.

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

Subgroup size	Age (>)	FFbHR[a] (<)	PCS[b] (<)	MCS[c] (<)	Treatment effect
0.108[d]	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 FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; b PCS, physical component scale of SF-12/36; c MCS, mental component scale of SF-12/36; d for about 10.8% of the population with age > 16, FFbHR < 100, SF-12/36 PCS < 28.84 and SF-12/36 MCS < 50.61, the treatment effect was 5.98.

8.4.1.5 Short-term SF-12/36 PCS 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.

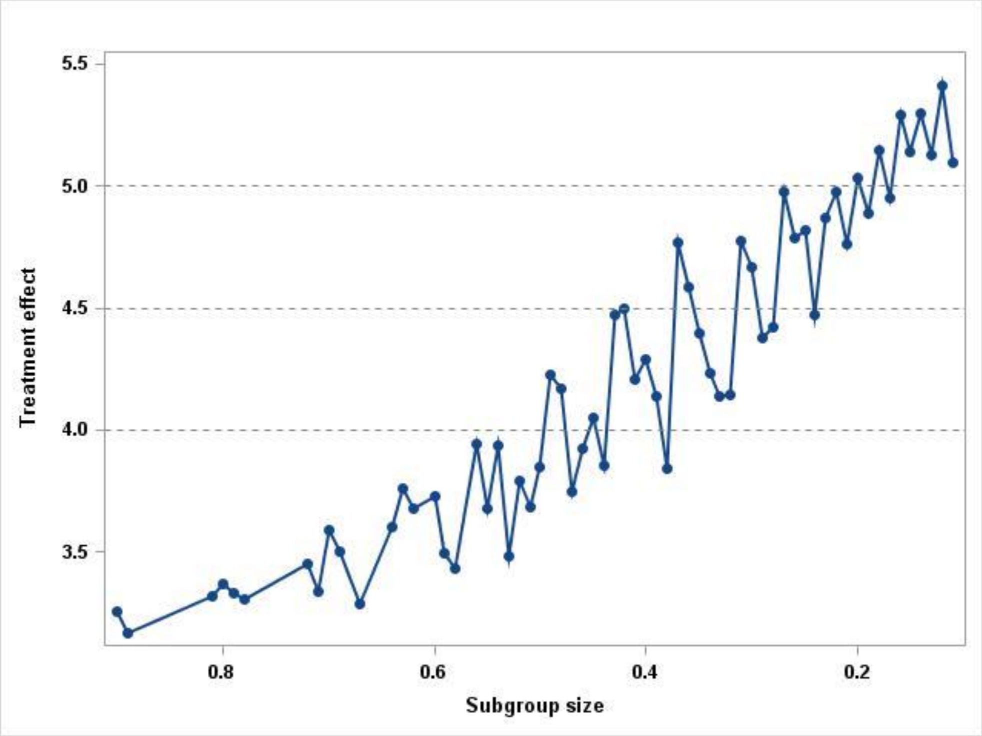


Figure 29 Trajectory plot for the treatment effect against the size of the constructed region for the SF-12/36 PCS short-term outcome.

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

Subgroup size	Age (<)	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
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

Subgroup size	Age (<)	PCS[a] (<)	MCS[b] (>)	Treatment effect
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, physical component scale of SF-12/36; b MCS, mental component scale of SF-12/36; c for about 11.0% of the population with age < 54, SF-12/36 PCS < 40.45 and SF-12/36 MCS > 56.82, the treatment effect was 5.30.

8.4.1.6 Short-term RMDQ outcome

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

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

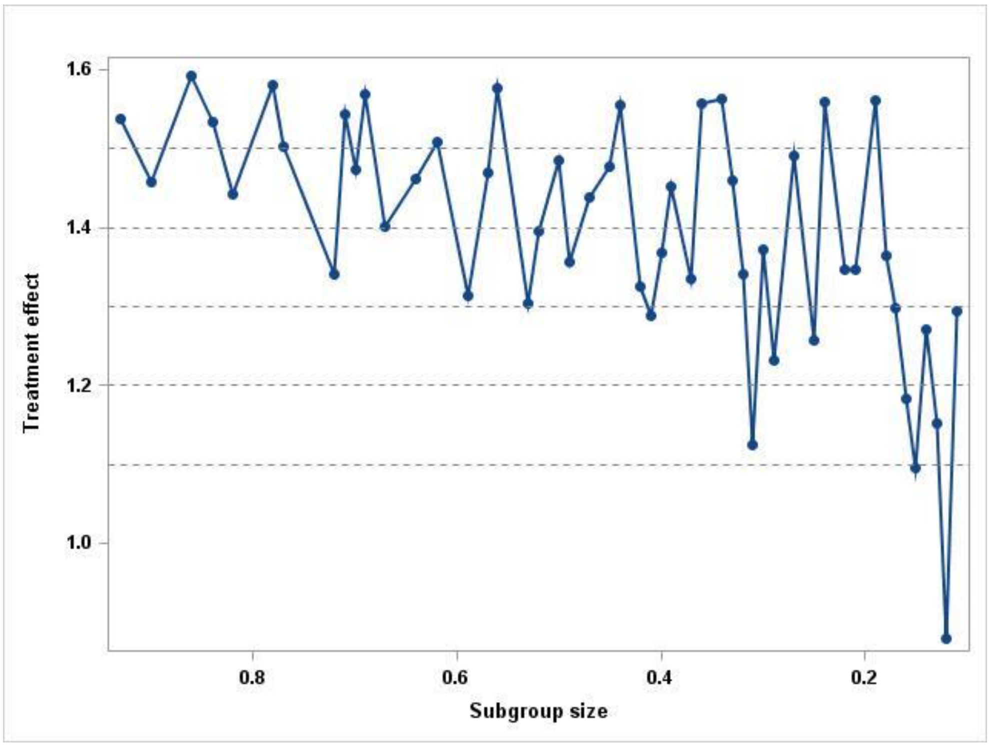


Figure 30 Trajectory plot for the treatment effect against the size of the constructed region for the RMDQ short-term outcome.

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

Subgroup size	Age (<)	RMDQ[a] (<)	Treatment effect
0.110[b]	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 RMDQ, Roland Morris disability questionnaire; b for about 11.0% of the population with age < 45 and RMDQ < 5, the treatment effect was 1.13.

8.4.2 Analysis 2: Pairwise comparisons

Similar to the analyses seen in *Section 7.5.2*, 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 were 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 outcomes measures.

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	FFbHR[a]		RMDQ[b]		MCS[c]		PCS[d]	
Comparison	**Trials**	**Variables**	**Trials**	**Variables**	**Trials**	**Variables**	**Trials**	**Variables**
Active physical vs. non-active usual care[e]			*m* = 2; *n* = 622 UK BEAM (*n* = 465), Smeets (*n* = 157)	Age and RMDQ at baseline				
Passive physical vs. non-active usual care[e]	*m* = 3; *n* = 3,272 Brinkhaus (n=214), Haake (*n* = 734), Witt (*n* = 2,324)	Age, FFbHR, PCS and MCS at baseline			*m* = 5; *n* = 3,879 UK BEAM (*n* = 479), Brinkhaus (*n* = 212), Haake (*n* = 734), Witt (*n* = 2,248), YACBAC (*n* = 206)	Age, PCS and MCS at baseline	*m* = 5; *n* = 3,879 UK BEAM (*n* = 479), Brinkhaus (*n* = 212), Haake (*n* = 734), Witt (*n* = 2,248), YACBAC (*n* = 206)	Age, PCS and MCS at baseline
Psychological vs. non-active usual care[e]			*m* = 3; *n* = 957 BeST (*n* = 514), VK BIA (*n* = 230), VK SC2 (*n* = 213)	Age and RMDQ at baseline				

Outcome	FFbHR[a]		RMDQ[b]		MCS[c]		PCS[d]	
Comparison	**Trials**	**Variables**	**Trials**	**Variables**	**Trials**	**Variables**	**Trials**	**Variables**
Sham vs. non-active usual care[e]	*m* = 2; *n* = 881 Brinkhaus (*n* = 144), Haake (*n* = 737)	Age, FFbHR, PCS and MCS at baseline			*m* = 2; *n* = 879 Brinkhaus (*n* = 142), Haake (*n* = 737)	Age, PCS and MCS at baseline	*m* = 2; *n* = 879 Brinkhaus (*n* = 142), Haake (*n* = 737)	Age, PCS and MCS at baseline

a FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; b RMDQ, Roland Morris disability questionnaire; c MCS, mental component scale of SF-12/36; d PCS, physical component scale of SF-12/36; e Control treatment is usual care/GP or waiting list.

8.4.2.1 Active physical vs. non-active usual care

8.4.2.1.1 Short-term RMDQ 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 similar result seen in *Section 8.4.1.6*, 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.

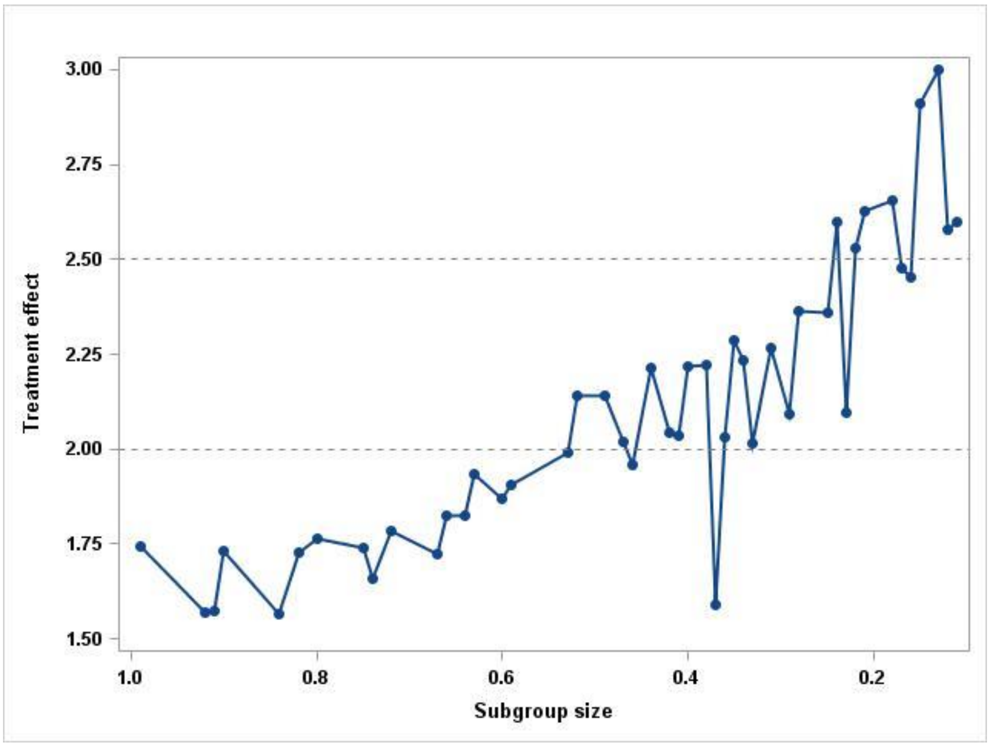


Figure 31 Trajectory plot for the treatment effect between active physical and non-active usual care against the size of the constructed region for the RMDQ short-term outcome.

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

Subgroup size	Age (>)	RMDQ[a] (>)	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[b]	5	1.73

a RMDQ, Roland Morris disability questionnaire; b minimum age=19

8.4.2.2 Passive physical vs. non-active usual care

8.4.2.2.1 Short-term FFbHR 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 FFbHR < 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, FFbHR < 50, PCS < 68 and MCS < 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.

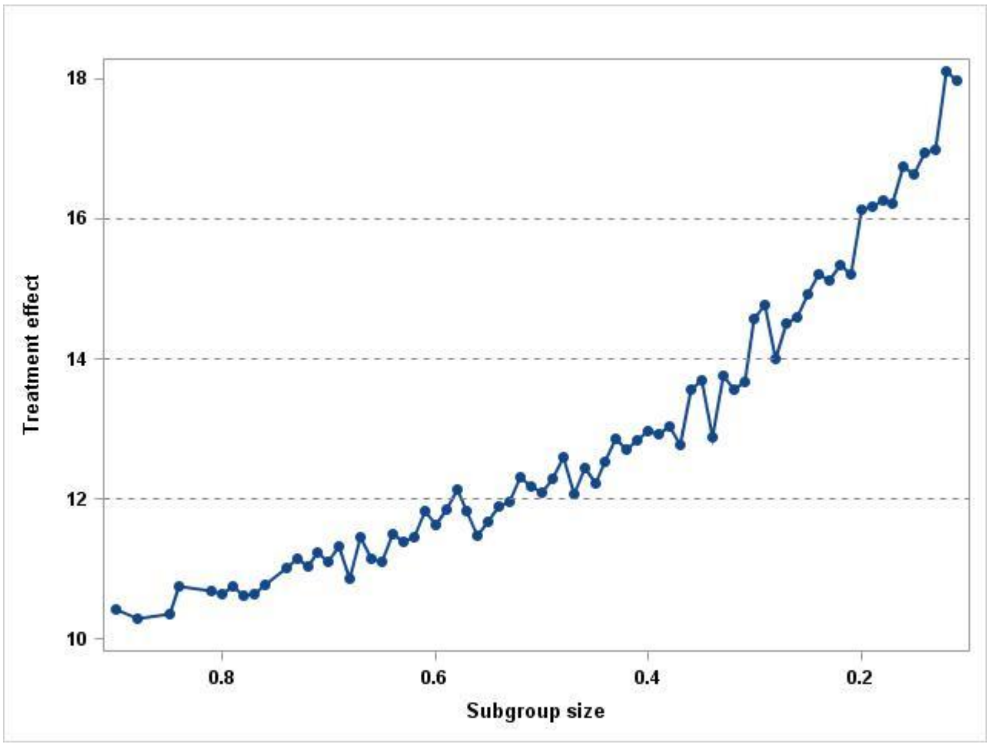


Figure 32 Trajectory plot for the treatment effect between passive physical and non-active usual care against the size of the constructed region for the FFbHR short-term outcome.

Table 33 Thresholds for selected size of subgroup for the short-term FFbHR as seen in *Figure 32*.

Subgroup size	Age (<)	FFbHR[a] (<)	PCS[b] (<)	MCS[c] (<)	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 FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; b PCS, physical component scale of SF-12/36; c MCS, mental component scale of SF-12/36

8.4.2.2.2 Short-term SF-12/36 MCS 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 in *8.4.1.4* where approximately 90% of the initial 3,879 participants (corresponding to age < 68, PCS < 68 and MCS < 71) had an average treatment effect of 3.06 (see *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 only applicable to a small proportion of participants, approximately 10% of them (corresponding to age < 51, PCS < 44 and MCS < 38). That is, only younger participants with substantial physical limitations and psychological distress would benefit from greater improvement in passive physical treatment against control.

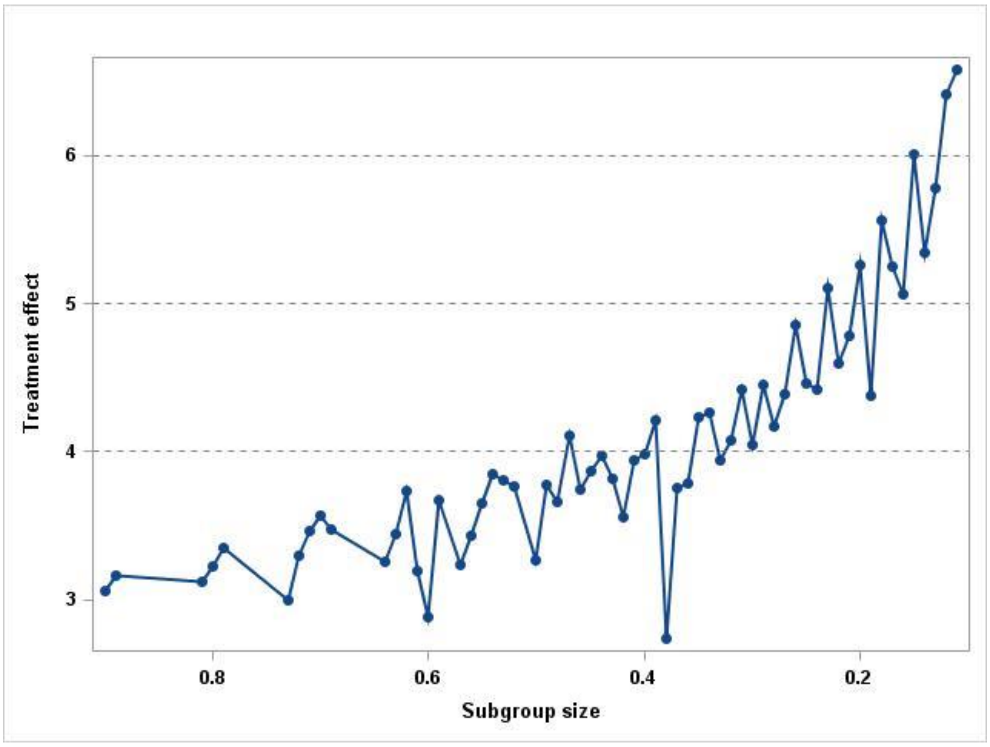


Figure 33 Trajectory plot for the treatment effect between passive physical and non-active usual care against the size of the constructed region for the SF-12/36 MCS short-term outcome.

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

Subgroup size	Age (<)	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, physical component scale of SF-12/36; b MCS, mental component scale of SF-12/36; c Maximum age =91

8.4.2.2.3 Short-term SF-12/36 PCS 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 29* summarised the average treatment for selected constructed regions with the corresponding thresholds for the comparison seen in *Figure 34*.

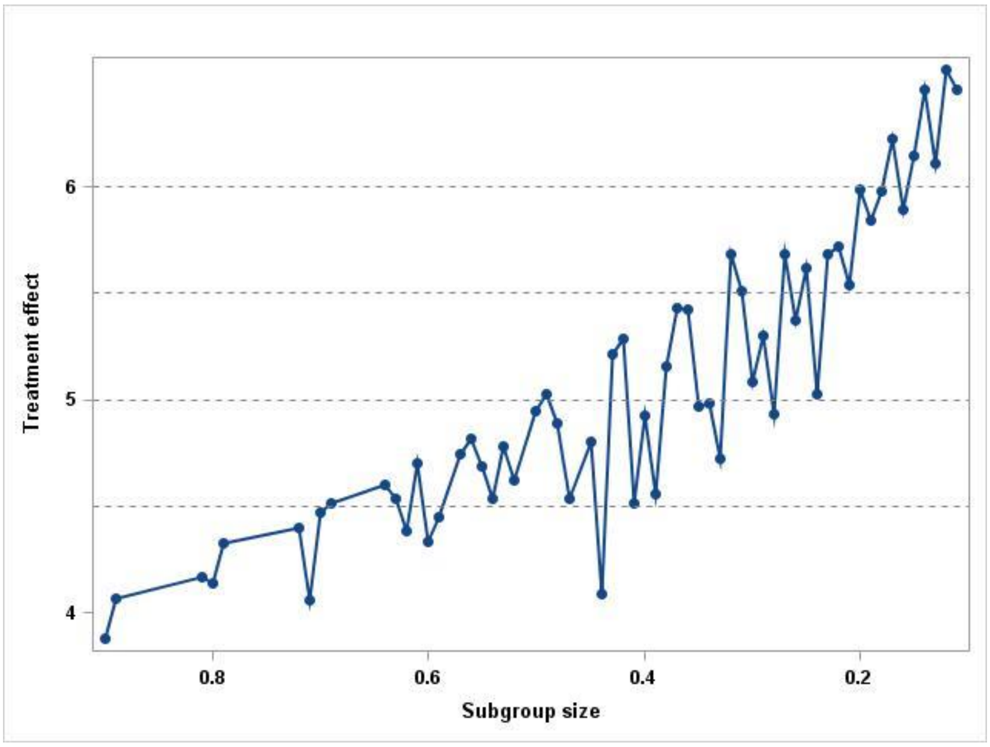


Figure 34 Trajectory plot for the treatment effect between passive physical and non-active usual care against the size of the constructed region for the SF-12/36 PCS short-term outcome.

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

Subgroup size	Age (<)	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, physical component scale of SF-12/36; b MCS, mental component scale of SF-12/36; ; c Maximum age =91

8.4.2.3 Psychological vs. non-active usual care

8.4.2.3.1 Short-term RMDQ 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 that seen in *Section 8.4.1.6*, that is, there was no subgroup that would gain a substantial improvement in treatment effect.

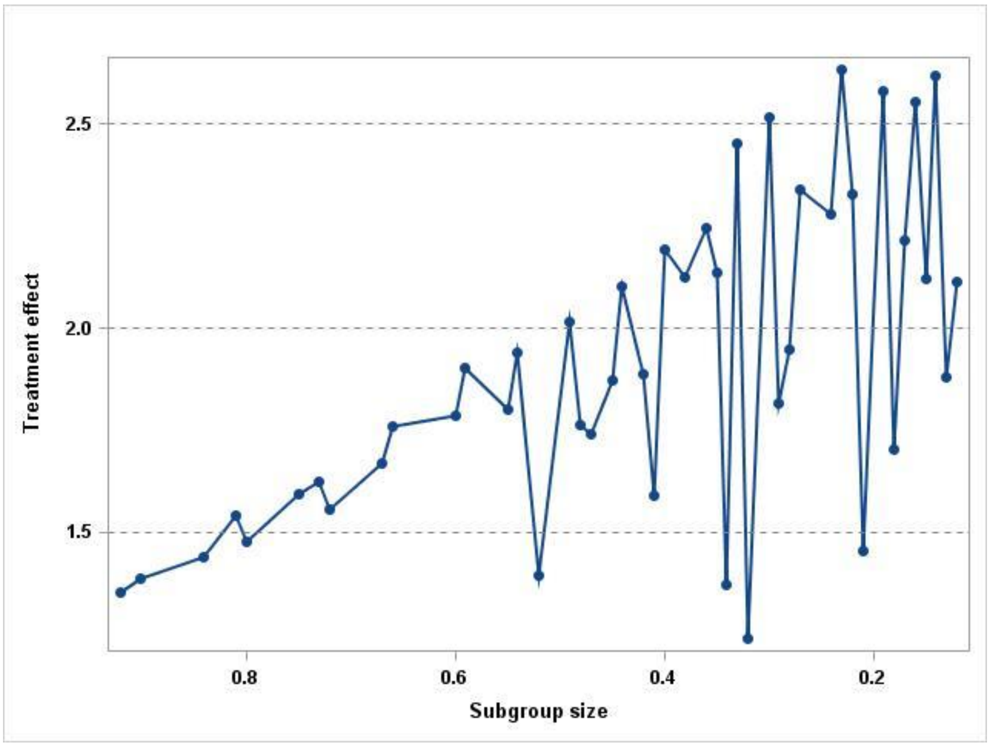


Figure 35 The size of the constructed region for the RMDQ short-term outcome.

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

Subgroup size	Age (<)	RMDQ[a] (>)	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[b]	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 RMDQ, Roland Morris disability questionnaire; b maximum age=85

8.4.2.1 Sham vs. non-active usual care

8.4.2.1.1 Short-term FFbHR outcome

Three trials were included in the comparison between passive physical and non-active usual care. All three trials had acupuncture as the therapist delivered intervention. Of these two of them 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 seen 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.

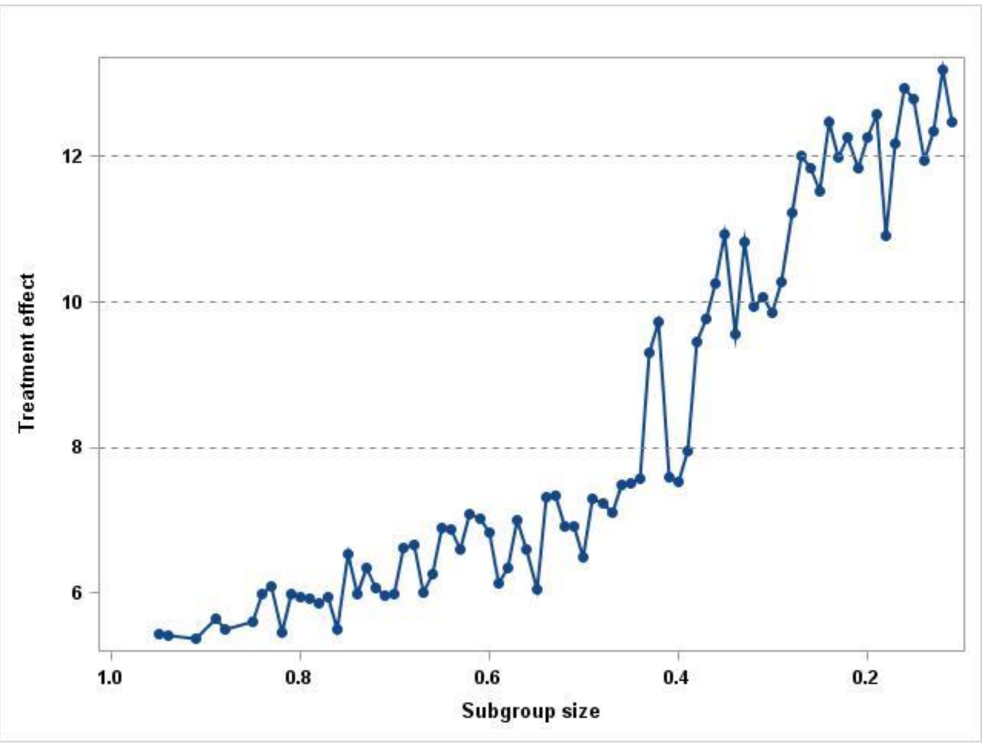


Figure 36 The size of the constructed region for the FFbHR short-term outcome.

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

Subgroup size	Age (<)	FFbHR[a] (<)	PCS[b] (<)	MCS[c] (<)	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 [d]	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 FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; b PCS, physical component scale of SF-12/36; c MCS, mental component scale of SF-12/36; d maximum age = 87

8.4.2.1.2 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 (see *Table 38*). Nevertheless, there was a markedly higher treatment effect which was 6.22 for approximately 20% of the participants (corresponding to PCS < 36 and MCS < 39, regardless of age).

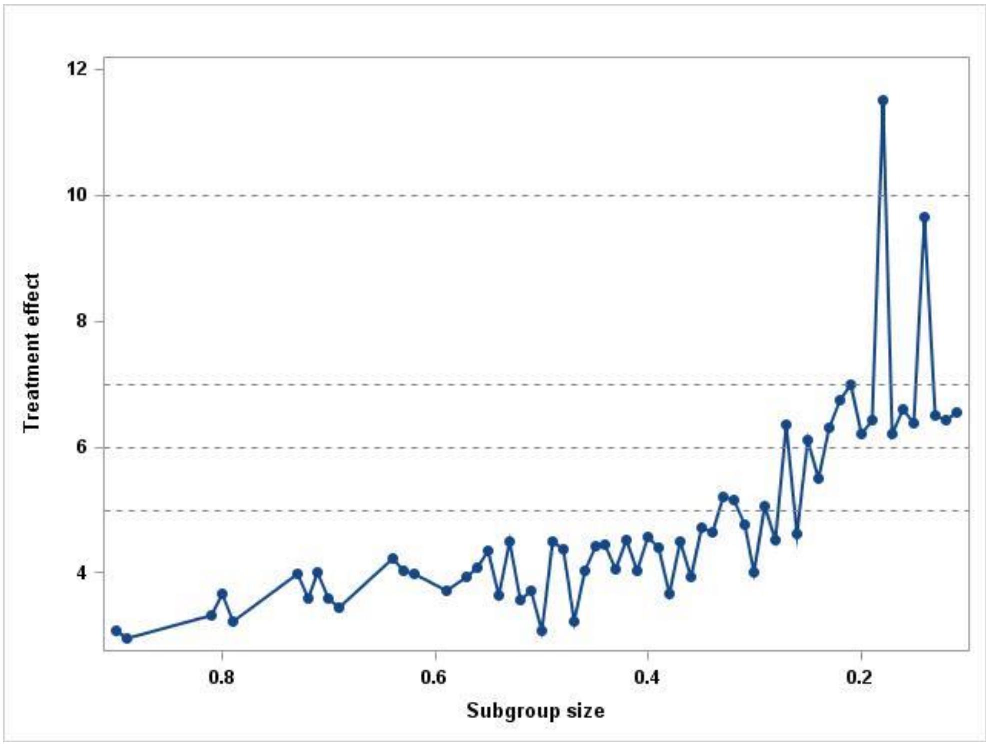


Figure 37 Trajectory plot for the treatment effect between sham and non-active usual care against the size of the constructed region for the SF-12/36 MCS short-term outcome.

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

Subgroup size	Age (<)	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[d]	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, physical component scale of SF-12/36; b MCS, mental component scale of SF-12/36; c maximum age = 87

8.4.2.1.3 Short-term PCS outcome

The trajectory plot for the treatment effect between sham and non-active usual care is shown in *Figure 38* and *Table 39* summarised 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.

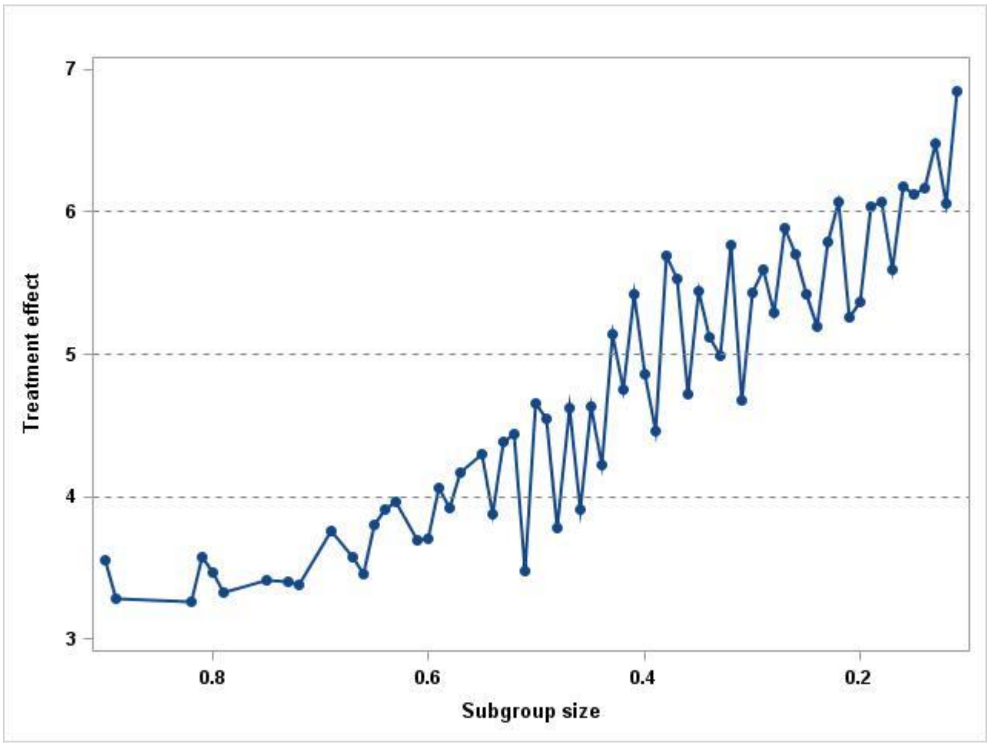


Figure 38 Trajectory plot for the treatment effect between sham and non-active usual care against the size of the constructed region for the SF-12/36 PCS short-term outcome.

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

Subgroup size	Age (<)	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, physical component scale of SF-12/36; b MCS, mental component scale of SF-12/36; c maximum age = 87

CHAPTER 9 – METHODOLOGY AND STATISTICAL DEVELOPMENTS 3: IDENTIFICATION OF COST-EFFECTIVE SUBGROUPS BY DIRECTED PEELING

9.1 INTRODUCTION

The economic analysis 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 healthcare 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 to £30,000 per QALY, as recommended by NICE), given the estimated QALY gain from treatment.[147]

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 to noise ratio: the directed peeling approach of LeBlanc et al, which works by 'peeling' a fraction of patients (with the least favourable effect) from the subgroup in a series of steps.[146] 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 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 vs best care, active physical interventions vs best care, passive physical interventions vs best care and active physical vs passive physical. Psychological interventions were not included in the comparison as only one trial had EQ-5D data necessary to calculate a QALY and a control arm. Data for comparisons against a 'sham' treatment arm were also excluded from this analysis.

9.2 METHODS

9.2.1 QALYs

The outcome used for the analysis was the Quality Adjusted Life Year (QALY). We calculated QALYs for individuals based on EQ-5D utility scores at baseline, short, medium and long term follow up (up to one year). For trials with SF-36/12 outcomes but no EQ-5D, we used a mapping algorithm[136] to estimate EQ-5D scores. QALYs were estimated using an area under the curve approach adjusting for baseline EQ-5D scores (see *Section 6.3.3.2*).

9.2.2 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 effect 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 *Section 6.3.3*). The sign on the moderator by treatment interaction coefficient dictated whether 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.

9.2.3 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 effect 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 was 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 one for the remaining individuals (the ‘in’ group) and zero 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 effect 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 pre-set minimum number (n_{min}), the algorithm stopped. Otherwise the above steps were repeated.

9.2.4 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 trial 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 = $\lambda * Q - C$, where λ is a set cost-effectiveness threshold (£20,000 per QALY). This NMB variable was then used as 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 a 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 BEAM 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 top 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 £1,000 (0.05*£20,000) for the treatment, using the lower threshold or £1,500 (0.05*£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 one year, was estimated at £541(SD: £768) from the UK BEAM economic analysis: £147 for the intervention and £394 relating to other healthcare costs (UK BEAM).[34] Estimates for other treatments varied, ranging from £422 (£187 for the intervention, £235 for other healthcare costs) for a psychological intervention (BeST 2010)[31] to £486 (SD: £907) comprised of £41 for the intervention and £445 relating to other healthcare costs, for active therapies (UK BEAM)[31].

9.3 RESULTS

Six analyses were run (see *Table 40*), dictated by the moderators with significant treatment interaction terms in the QALY analysis of covariance. These included the following comparisons: all interventions versus control; active physical versus control; passive physical versus control; and active physical versus passive physical. As noted above, analysis 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 versus control comparison: the first to include as many trials as possible with QALY data (age and PCS as moderators).

Table 40 ARDP-MA, analyses conducted on economic outcomes

Analysis	Outcome variable	Moderators included	Trials included	Sample size *I : C*
All interventions vs control				
9.3.1	QALY[a]	Age, PCS[b]	UK BEAM[c]; BeST[d]; YACBAC[e]; Haake	1,273 : 715
9.3.2	QALY	Age, RMDQ[f]	UK BEAM; BeST; York; Smeets	1,092 : 422
9.3.3	QALY	Age, PCS, RMDQ	UK BEAM; BeST	827 : 323
Active physical interventions vs control				
9.3.4	QALY	Age, RMDQ	UK BEAM; York	232 : 264
Passive physical interventions vs control				
9.3.5	QALY	Age, PCS	UK BEAM, YACBAC, Haake	643 : 566
Active physical vs passive physical interventions				
9.3.6	QALY	Age, RMDQ	UK BEAM, HullExProB	232 : 288

a QALY, quality adjusted life year; b PCS, physical component scale of SF-12/36; c Back pain Exercise And Manipulation; d BeST, Back Skills Training Trial; e YACBAC York Acupuncture Back Pain Trial; f RMDQ, Roland Morris Disability Questionnaire

9.3.1 All interventions vs control. Moderators: age and PCS

The algorithm trace is shown in *Figure 39*. The *y*-axis shows the estimated treatment effect for the subgroup, i.e. 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 based on PCS score. 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 between 54 and 84 years old with a PCS score 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.

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 £1,248 (per patient over the course of a year) if all patients were to be offered treatment, or £1,704 if only patients in the 10% subgroup were to be offered treatment. If the threshold of £30,000 was applied this will be £1,872 and £2,556 respectively. However, these results do not incorporate any measure of uncertainty and should only be considered as 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

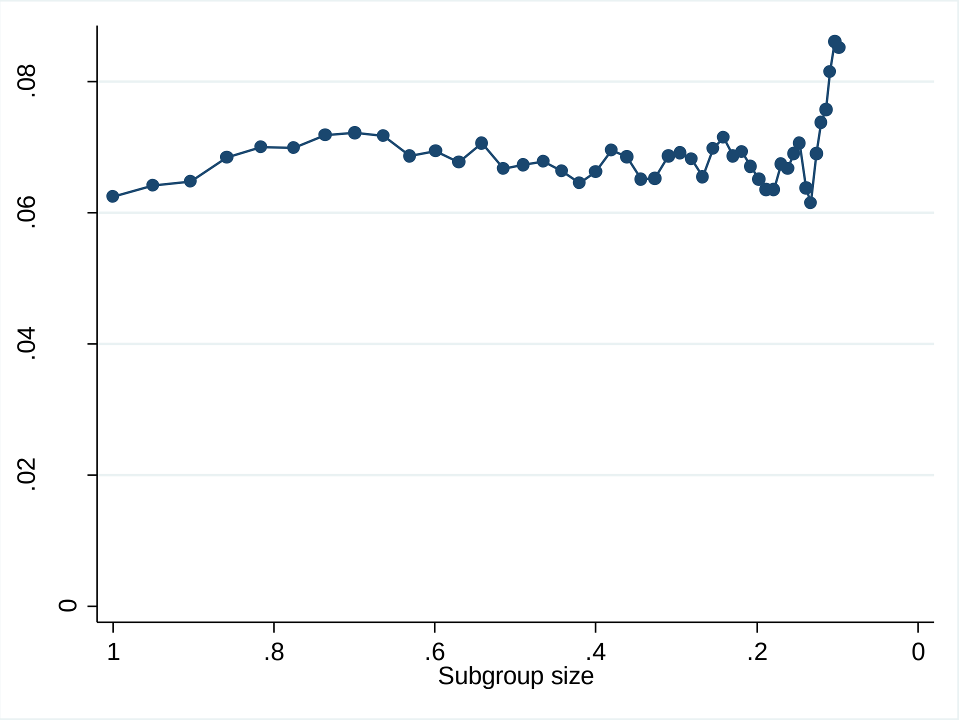


Figure 39 Mean treatment effect in subgroup

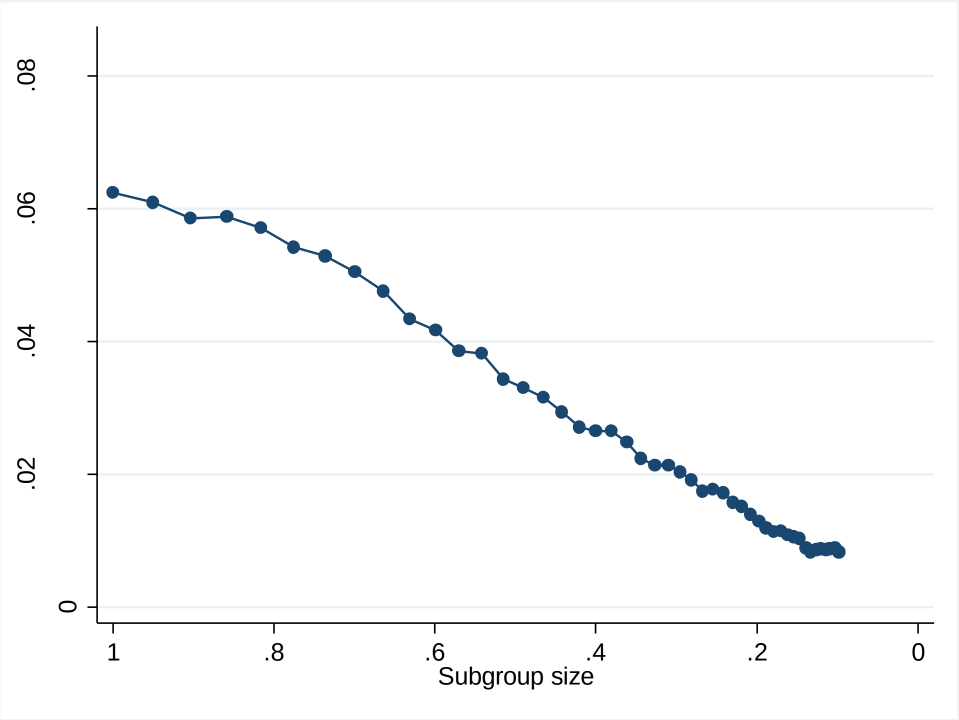


Figure 40 Weighted mean treatment effect across treatment subgroup and non-treatment subgroup

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

Iteration	Moderator	Direction peeled	Proportion in subgroup	≈*n*	Incremental QALYs[a]		Age		PCS[b]	
					Subgroup	All	Min	Max	Min	Max
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

Iteration	Moderator	Direction peeled	Proportion in subgroup	≈*n*	Incremental QALYs[a]		Age		PCS[b]	
					Subgroup	All	Min	Max	Min	Max
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
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 QALY, quality adjusted life year; b PCS, physical component scale of SF-12/36

9.3.2 ALL INTERVENTIONS VS CONTROL. MODERATORS: AGE AND RMDQ

Figure 41 and *Figure 42* illustrate the peeling trace with moderators 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 favoured peeling 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 to age. Due to some correlation with RMDQ and age, some older patients were removed from the treatment subgroup as the algorithm peeled based on RMDQ.

The peeling trace for analysis ii is shown in *Table 42*. The subgroup at 20% of the initial sample comprised participants aged over 34 with an RMDQ score of 13 or higher. A modest improvement in QALYs gained from treatment can be seen for this subgroup: from 0.043 if the whole population where 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 £1,520 respectively for the whole population and for the subgroup where a threshold of £20,000 is applied, or £1,290 and £2,280 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 at baseline appear to achieve moderately better response to treatment.

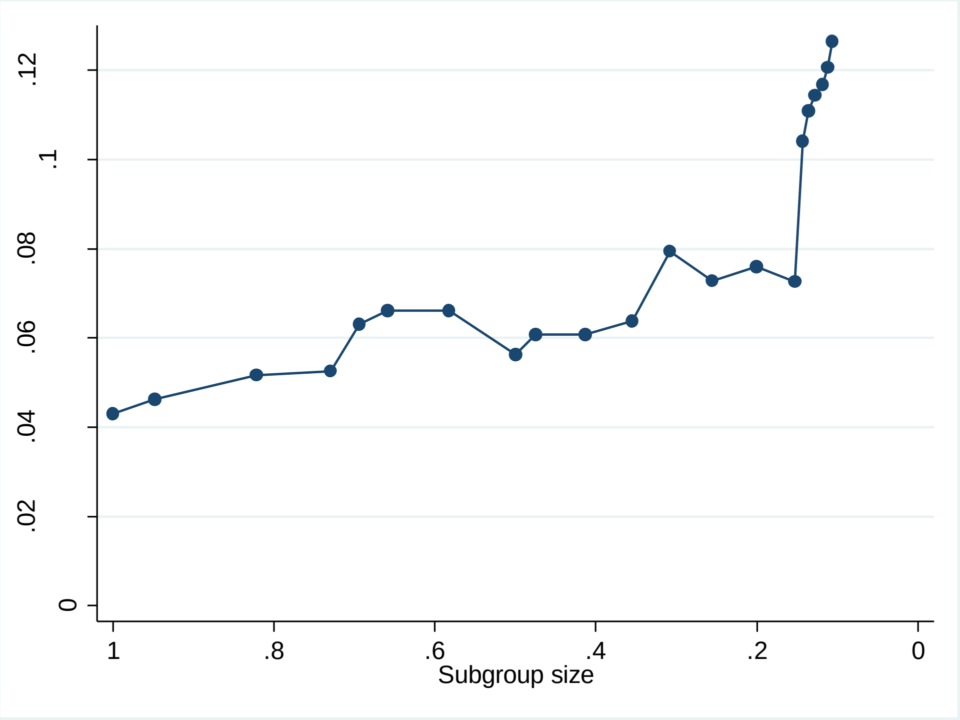


Figure 41 Mean treatment effect in subgroup

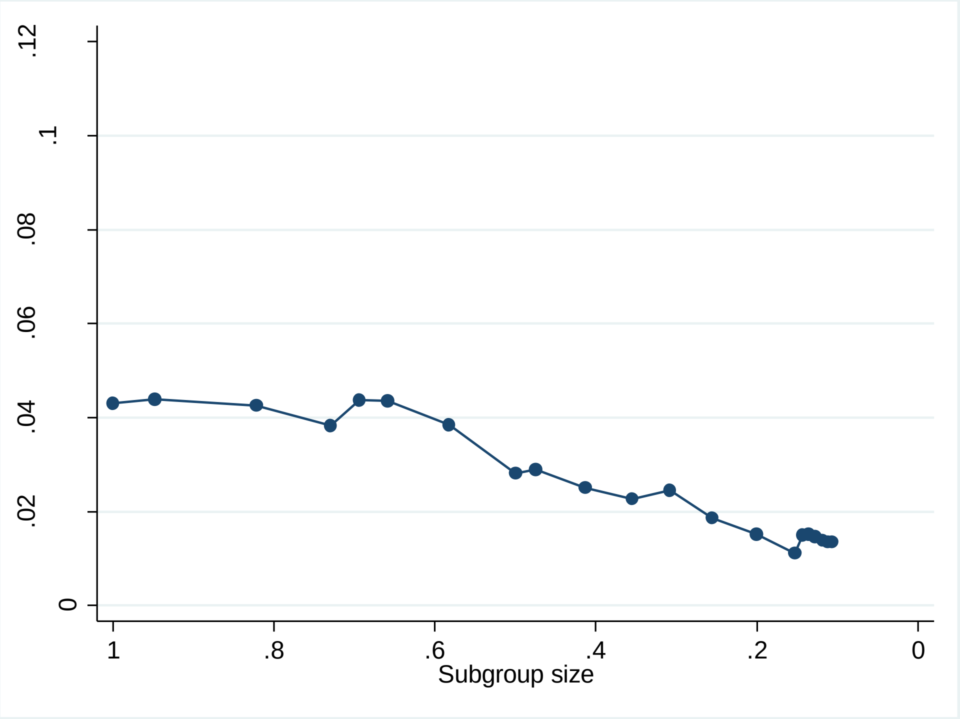


Figure 42 Weighted mean treatment effect across treatment subgroup and non-treatment subgroup

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

Iteration	Moderator	Direction peeled	Proportion in subgroup	≈*n*	Incremental QALYs[a]		Age		RMDQ[b]	
					Subgroup	All	Min	Max	Min	Max
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

a QALY, quality adjusted life year; b RMDQ, Roland Morris disability questionnaire

9.3.3 ALL INTERVENTIONS VS CONTROL. MODERATORS: AGE, PCS AND RMDQ

Figure 43 and *Figure 44* illustrate the peeling results for the analysis with age, PCS and RMDQ. As some trials did not have available PCS scores and others did not have RMDQ scores, the sample was restricted to two trials. The results of the peeling trace are very similar to those of analysis ii). The algorithm chose to peel almost exclusively on RMDQ and age. PCS was employed 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) PCS scores were removed from the treatment subgroup; this is unsurprising as RMDQ 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 was comprised of participants aged 44 to 82 with an RMDQ score over 12 and a PCS score 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 £1,008 and £1,962 for the whole population and the refined subgroup respectively. At £30,000 per QALY, these figures are £1,512 and £2,943 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 at baseline appear to have moderately better response to treatment

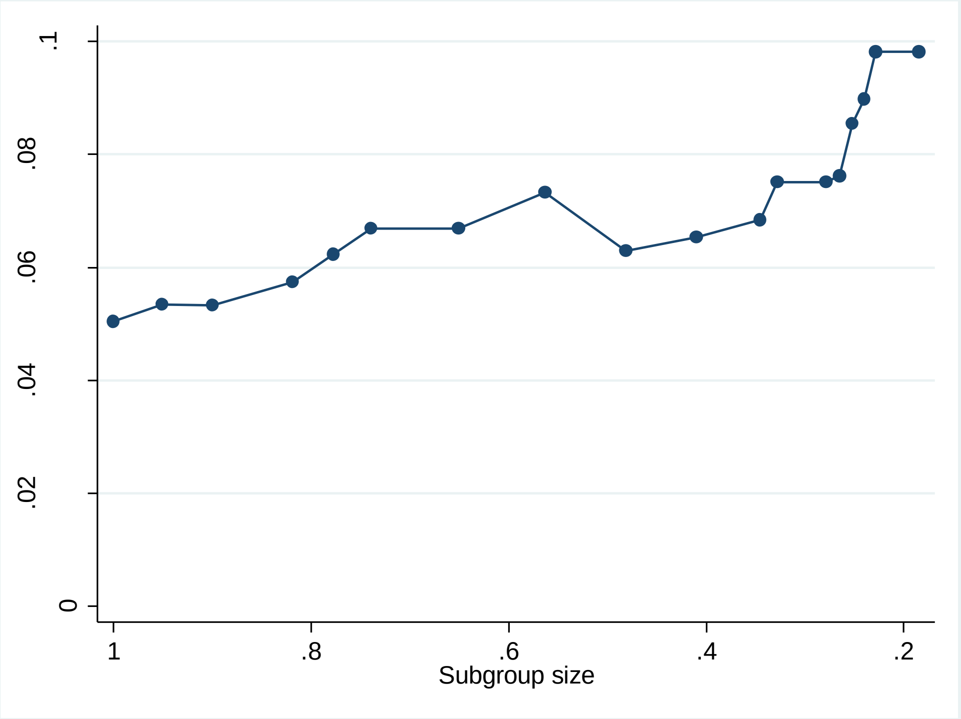


Figure 43 Mean treatment effect in subgroup

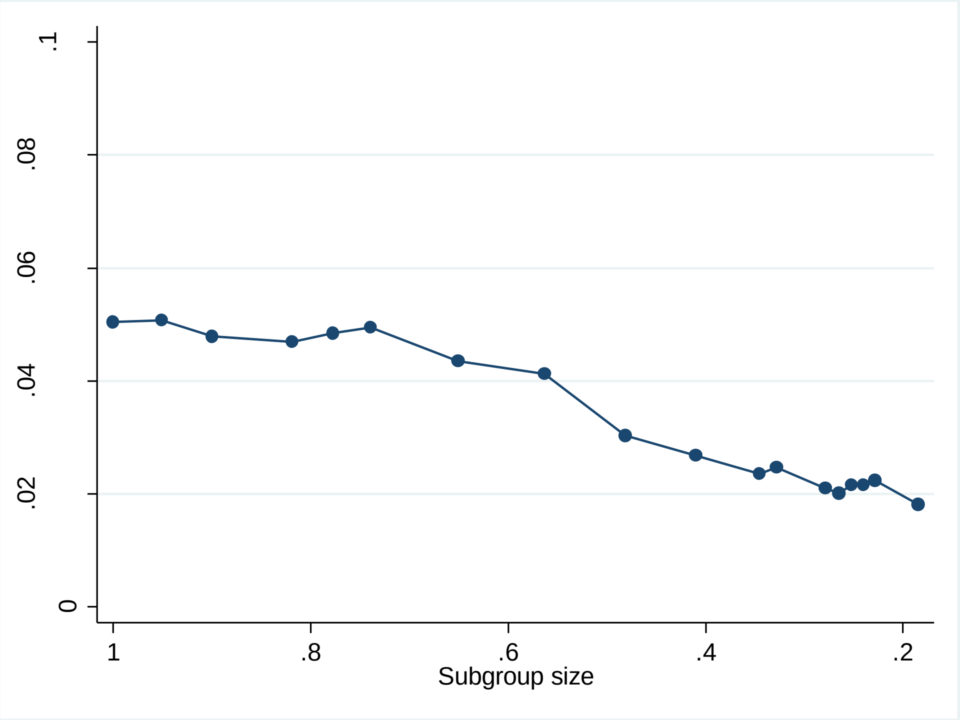


Figure 44 Weighted mean treatment effect across treatment subgroup and non-treatment subgroup

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

Iteration	Moderator	Direction peeled	Proportion in subgroup	≈*n*	Incremental QALYs[a]		Age		PCS[b]		RMDQ[c]	
					Subgroup	All	Min	Max	Min	Max	Min	Max
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 QALY, quality adjusted life year; b PCS, physical component scale of SF-12/36; c RMDQ, Roland Morris disability questionnaire

9.3.4 Active physical intervention vs control. Moderators: age and RMDQ

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. *Figure 45* shows the peeling trace with RMDQ 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 patients with worse baseline RMDQ scores achieve better treatment outcomes. At iteration 10, age was peeled on, removing patients who were younger.

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 active physical treatment in these analyses.

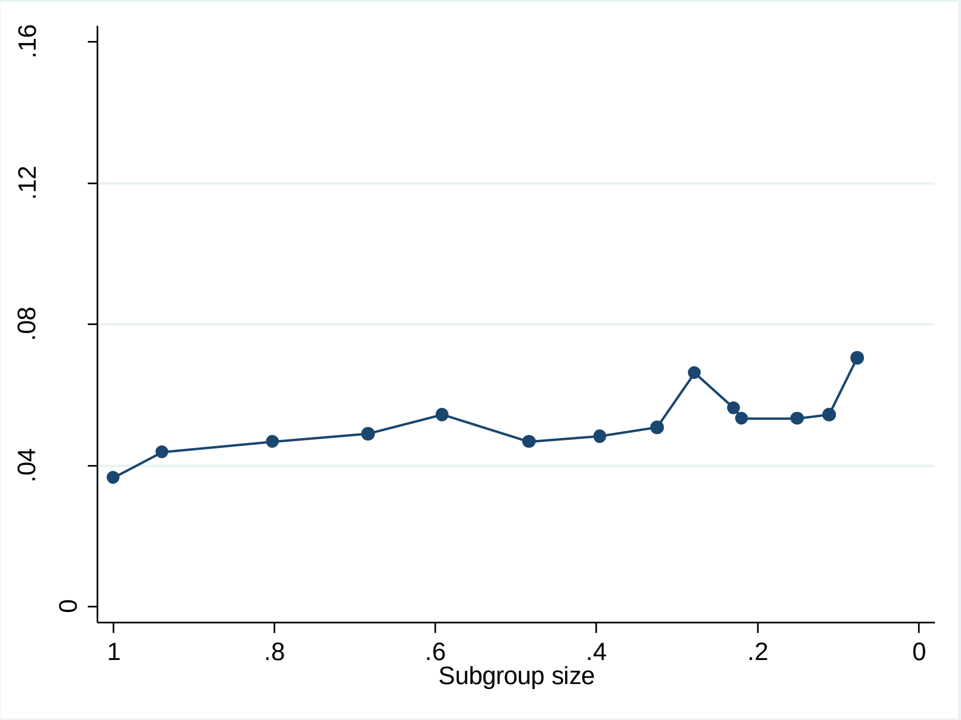


Figure 45 Mean treatment effect in subgroup

9.3.5 Passive physical intervention vs control. Moderators: age and PCS

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

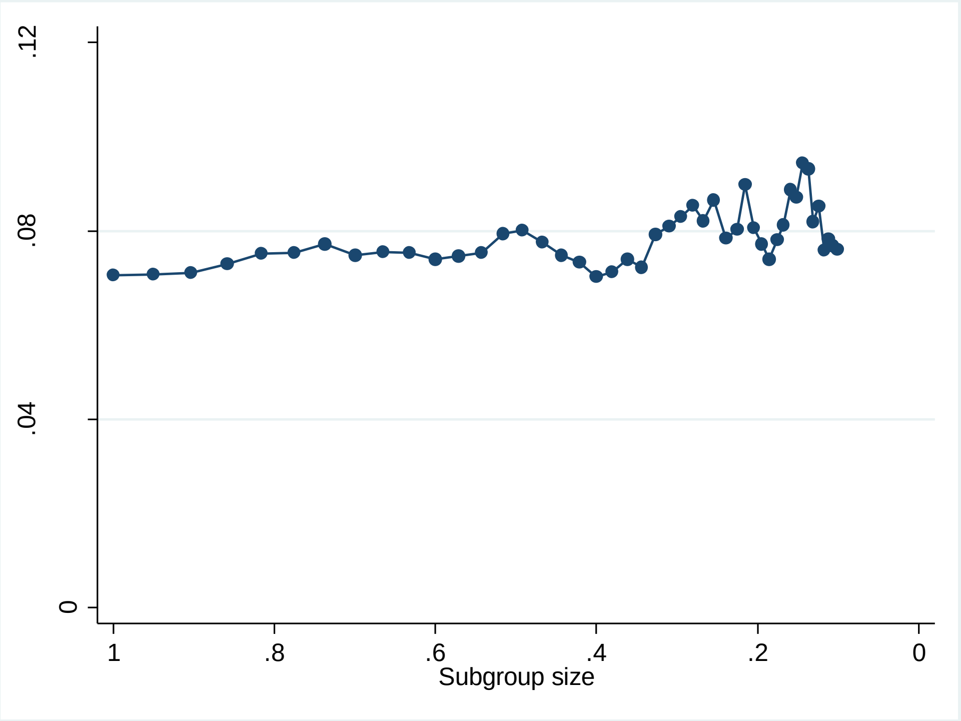


Figure 46 Mean treatment effect in subgroup

9.3.6 ARDP-MA DIRECTED PEEL. ACTIVE PHYSICAL VS PASSIVE PHYSICAL INTERVENTIONS. MODERATORS: AGE AND RMDQ

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 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.

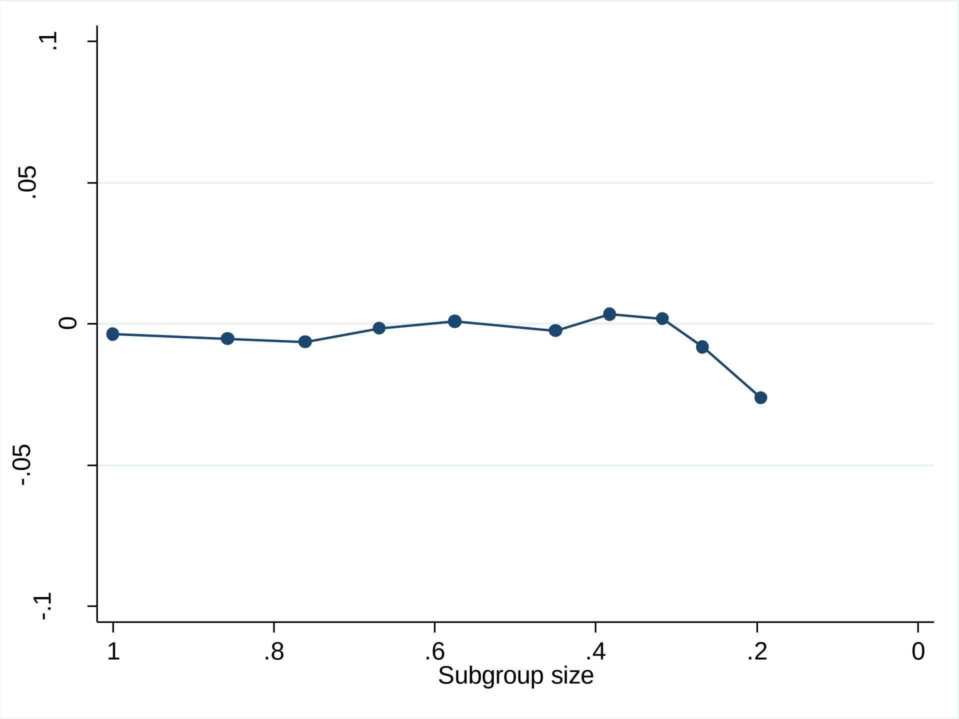


Figure 47 Mean treatment effect in subgroup

9.4 DISCUSSION

The application of the peeling algorithm was successful in identifying potentially interesting subgroups for the interventions vs 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 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 to 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 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 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 area under the curve for a sequence of follow up points. However, it is also possible the results are susceptible to missing data bias. Patients with missing EQ-5D data at one or more follow up points were on average four 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 1,774 patients (1,271 Intervention; 503 Control) for which there was complete data. Of the 1,774 patient, 1,467 (1,093 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.

CHAPTER 10 – METHODOLOGY AND STATISTICAL DEVELOPMENTS 4: SUBGROUP IDENTIFICATION WITH INDIVIDUAL PARTICIPANT DATA INDIRECT NETWORK META-ANALYSIS

10.1 BACKGROUND

The recursive partitioning and adaptive peeling approaches described in our analysis plan, whilst 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 to identify groups who may gain the greatest benefit from different treatment choices from a Bayesian rather than a frequentist perspective.

10.2 METHODS

We carried out network meta-analyses of the repository trials to explore how the optimal choice of treatment for low back pain might vary across subgroups. Network meta-analysis (NWMA) is an extension of standard pair-wise meta-analysis applicable in situations where we have multiple treatments and an evidence base of trials which individually provide evidence on different subsets of all possible pairwise treatment combinations.[148] NWMA involves analysing this network as a whole, by assuming consistency across treatment effects, so that a given pairwise comparison B vs C can be derived from trials against a common comparator (A vs B and A vs C trials) even if no B vs C trials exist.[149] NWMA has become increasing 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[150, 151]). Given their widespread use in Health Technology Assessment, NWMA commonly uses aggregate data, although there are examples illustrating the value of this approach when individual participant data (IPD) is available, particularly in understanding participant-level effect modification.[152, 153]

The standard model for pairwise meta-analysis involving a continuous normally-distributed outcome with linear effect modification can be written as equations (3) & (4).

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

$$\Delta_{it} = I_{it}\left(d_t + \beta_t\left(X_{it} - \overline{X}\right)\right) \quad (4)$$

where y_{it} is the outcome for participant *i* in trial *t*, μ_{it} is the expected outcome for participant *i* if they 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*, $\overline{X}$ 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)

$$\mu_{it} = \mu_t + b\left(X_{it} - \overline{X}\right) \quad (5)$$

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)

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

where $d_{1,j}$ is defined as the treatment effect of any treatment *j* in the network compared to 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)

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

We carried out three separate NWMAs for the outcomes of interest – short-term change in Roland Morris disability questionnaire (RMDQ), short-term change in physical component scale of SF-12/36 (PCS), and short-term change in mental component scale of SF-12/36 (MCS). 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 for both adjustments. Model estimation involved Bayesian Markov Chain Monte Carlo simulation carried out using WinBUGS 1.4.3, using NMWA models adapted for IPD analysis from aggregate-data NWMA models developed for the UK National Institute of Health and Care Excellence.[154]

10.3 RESULTS

10.3.1 SHORT-TERM RMDQ OUTCOME

Thirteen trials (n = 3447) in the repository reported this outcome. The resulting network of evidence is illustrated in *Figure 48*.

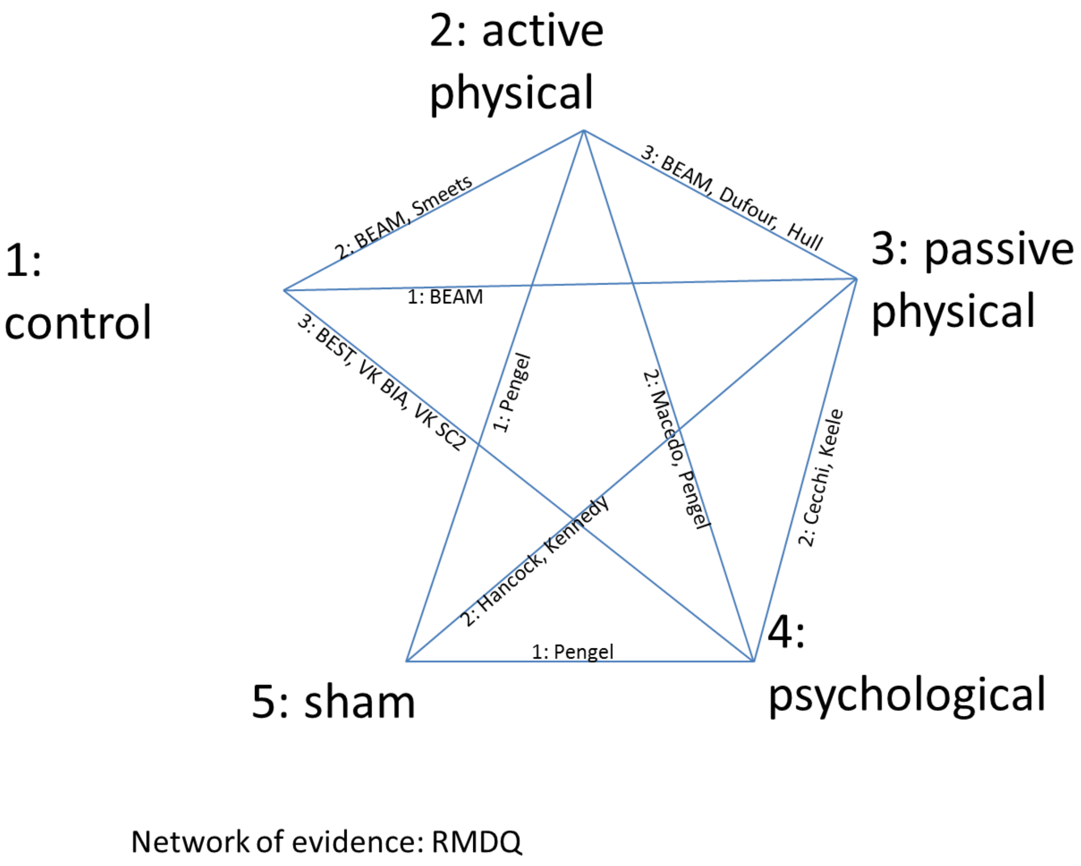


Figure 48 Network of evidence for short-term RMDQ. Each line denotes the existence of head-of-trials of the two treatments being connected, and the accompanying information denotes the number and names of trials making the comparison.

Table 44 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, male, and baseline values of RMQ=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. Also 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 44 Treatment effect with modification (absolute reduction in short term RMDQ, mean and 95% credible interval). Coefficients given for individual aged 50, male, RMQ=10, PCS=40 and MCS=40 at baseline^{a}.

		Comparator			
Intervention		**Control**	**Active Physical**	**Passive Physical**	**Psychological**
	Active Physical	1.94 (1.17, 2.72)			
	Passive Physical	2.17 (1.39, 1.95)	0.23 (-0.61, 1.07)		
	Psychological	1.45 (0.74, 2.15)	-0.49 (-1.31, 0.32)	-0.72 (-1.52, 0.08)	
	Sham	1.60 (-1.07, 4.11)	-0.34 (-2.95, 2.1)	-0.57 (-3.2, 1.9)	0.15 (-2.47, 2.63)

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

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 verum interventions exclude zero; for sham treatment it does include zero. This analysis suggests that for each one point increase in baseline RMDQ an additional 0.17 to 0.26 benefit from active treatments and a 0.43 point benefit from sham treatment will be achieved. However, the 95% credible intervals suggest 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, the *BP*s are all above 0.99 (except for sham, with a *BP* of 0.92) - overwhelming evidence that the effect of treatment depends on baseline scores.

The Bayesian Probabilities 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.

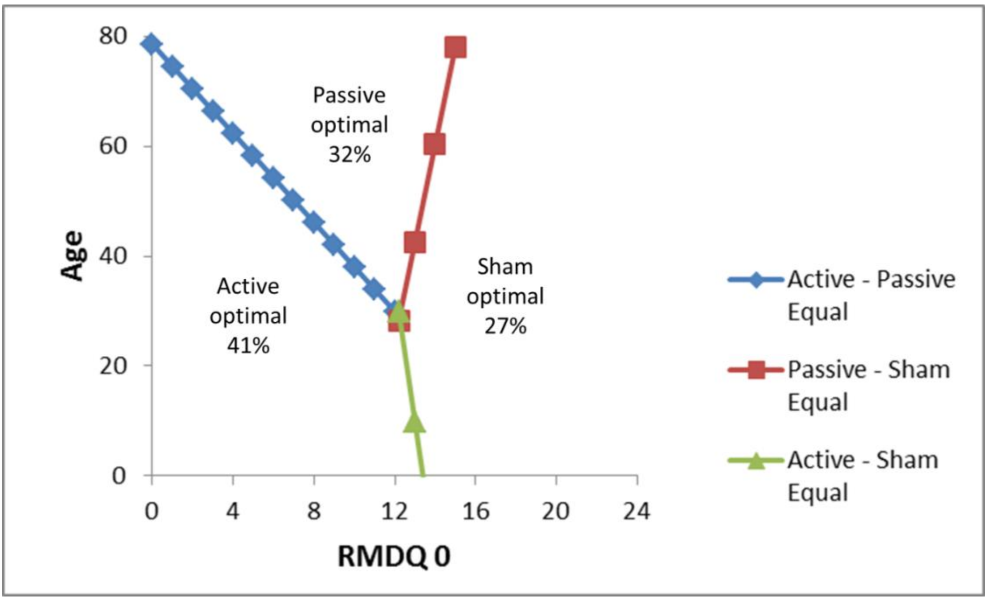
Table 45 Mean, 95% credible intervals and Bayesian Probabilities (*BP*) for impact of participant characteristics on effect of treatments (Vs. Control).

	Active Physical	Passive Physical	Psychological	Sham
Age[a]	-0.02 (-0.05, 0.02) *BP* = 0.83	0.00 (-0.03, 0.03) *BP* = 0.60	-0.02 (-0.05, 0.01) *BP* = 0.91	-0.01 (-0.08, 0.07) *BP* = 0.56
Sex	-0.22 (-1, 0.56) *BP* = 0.71	-0.38 (-1.16, 0.4) *BP* = 0.83	-0.01 (-0.78, 0.77) *BP* = 0.51	-1.12 (-2.74, 0.49) *BP* = 0.91
RMDQ[a]	0.18 (0.06, 0.31) *BP* > 0.99	0.26 (0.14, 0.39) *BP* > 0.99	0.17 (0.05, 0.29) *BP* > 0.99	0.43 (-0.11, 0.93) *BP* = 0.92
MCS[a]	-0.01 (-0.06, 0.05) *BP* = 0.59	0 (-0.05, 0.05) *BP* = 0.51	0.03 (-0.03, 0.08) *BP* = 0.85	-0.06 (-0.35, 0.24) *BP* = 0.59
PCS[a]	0.05 (-0.03, 0.13) *BP* = 0.89	0.04 (-0.04, 0.12) *BP* = 0.84	0.03 (-0.04, 0.11) *BP* = 0.81	-0.04 (-0.53, 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.

All treatment effects increase but at different rate, so that the optimal treatment changes as RMDQ varies. Passive physical therapy is the optimal therapy for the participant as described in *Table 45*, whose RMDQ is 10. However, sham therapy becomes the optimal treatment if RMDQ increases beyond 14 points, while active physical therapy becomes optimal if RMDQ decreases beyond seven points.

These thresholds depend on values for other effect modifiers, although their influence is less certain. The only other characteristics with a *BP* above 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. *Figure 49* and *Figure 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 mild-to-moderate RMDQ at baseline, active physical therapy is optimal for participants with low RMDQ at baseline, and sham therapy is optimal for participants with high RMDQ 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.



.

Figure 49 RMDQ outcome; optimal treatment as a function of RMDQ at baseline and age for men with MCS=PCS=40, with proportion of male trial participants whose baseline RMDQ and age fit into each zone (*n* = 721)

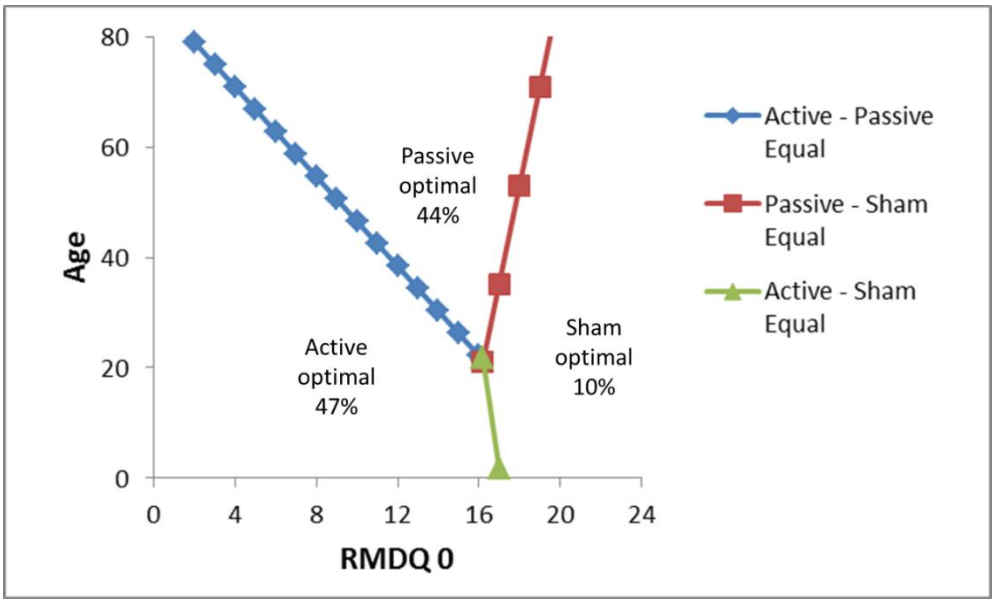


Figure 50 RMDQ outcome; optimal treatment as a function of RMDQ at baseline and age for women with MCS=PCS=40, with proportion of female trial participants whose baseline RMDQ and age fit into each zone (*n* = 1,054)

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 (see *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 high RMDQ at baseline (i.e. participant profiles 1, 3, 4 and 6).

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

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

10.3.2 SHORT-TERM SF-12/36 PCS OUTCOME

Nine trials ($n = 5574$) in the repository reported this outcome. The resulting network of evidence is illustrated in *Figure 51*.

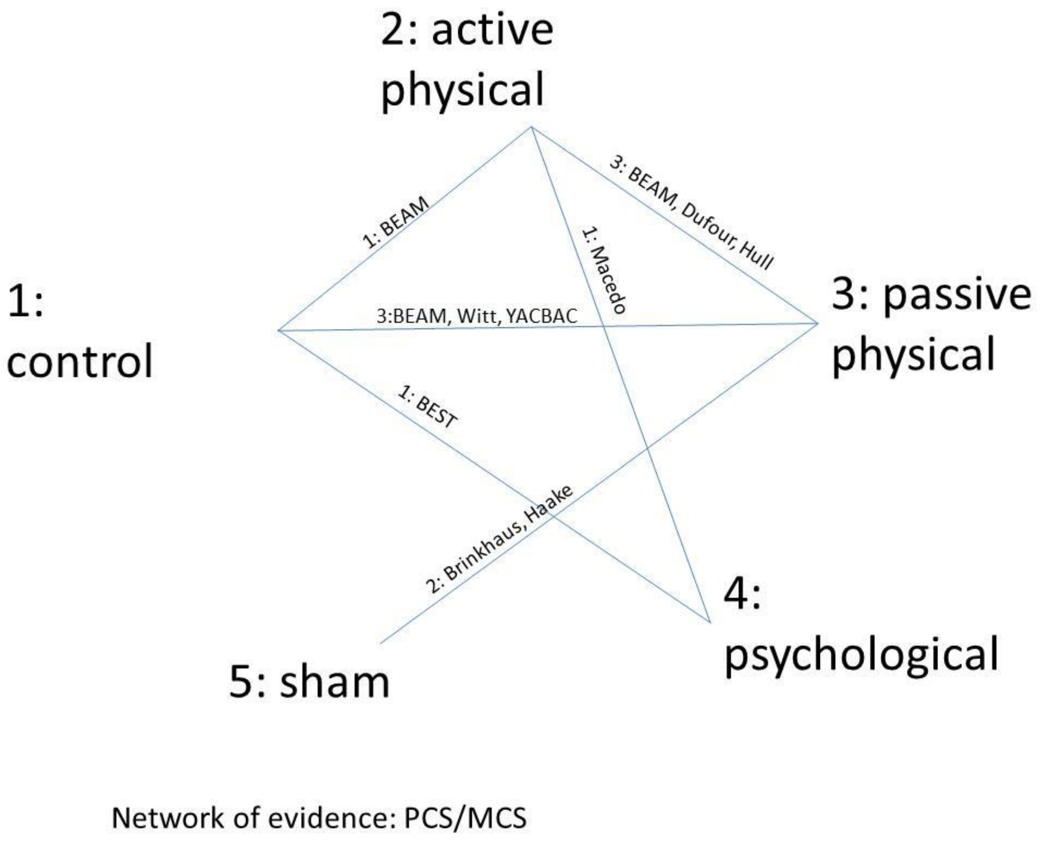


Figure 51 Network of evidence for short-term PCS. Each line denotes the existence of head-of-trials of the two treatments being connected, and the accompanying information denotes the number and names of trials making the comparison.

Table 47 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 Bayesian Probability above 0.95.

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

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

Table 48 Mean, 95% credible intervals and Bayesian Probability for impact of participant characteristics on effect of treatments in the network

	Active Physical	Passive Physical	Psychological	Sham
Age[a]	0.02 (-0.05, 0.08) *BP* = 0.68	-0.01 (-0.04, 0.03) *BP* = 0.71	-0.04 (-0.1, 0.03) *BP* = 0.87	(-0.06, 0.06) *BP* = 0.52
Sex[b]	0.25 (-1.25, 1.75) *BP* = 0.63	0.95 (0.04, 1.87) *BP* = 0.98	0.29 (-1.43, 2.01) *BP* = 0.63	1.55 (-0.15, 3.23) *BP* = 0.96
MCS0[a]	-0.01 (-0.07, 0.06) *BP* = 0.59	0.01 (-0.02, 0.05) *BP* = 0.76	0.03 (-0.04, 0.11) *BP* = 0.80	-0.07 (-0.14, 0.00) *BP* = 0.97
PCS0[a]	-0.05 (-0.15, 0.05) *BP* = 0.85	-0.07 (-0.13, -0.02) *BP* > 0.99	-0.03 (-0.13, 0.06) *BP* = 0.76	-0.10 (-0.22, 0.02) *BP* = 0.95

a Positive value indicates greater increase in PCS from treatment (Vs. Control) as covariate increases; b Positive value indicates greater increase in PCS from treatment (Vs. Control) for females Vs. males.

Figure 52 and *Figure 53* show how effect modification can be used to define PCS/MCS zones in which each treatment is optimal with short-term PCS as outcome of interest. Broadly speaking, passive physical therapy is optimal for participants with low PCS scores and high MCS scores, while active physical therapy is optimal for participants with high PCS scores and low MCS scores. Sham appears optimal for participants with low PCS and MCS scores 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.

The results (see *Table 49*) show that, as with RMDQ, there is greater certainty around which treatments are sub-optimal than around which treatments are optimal. For the paradigmatic cases in *Figure 52* and *Figure 53*, it is unlikely that psychological treatments would be the best choice for either gender, but 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.

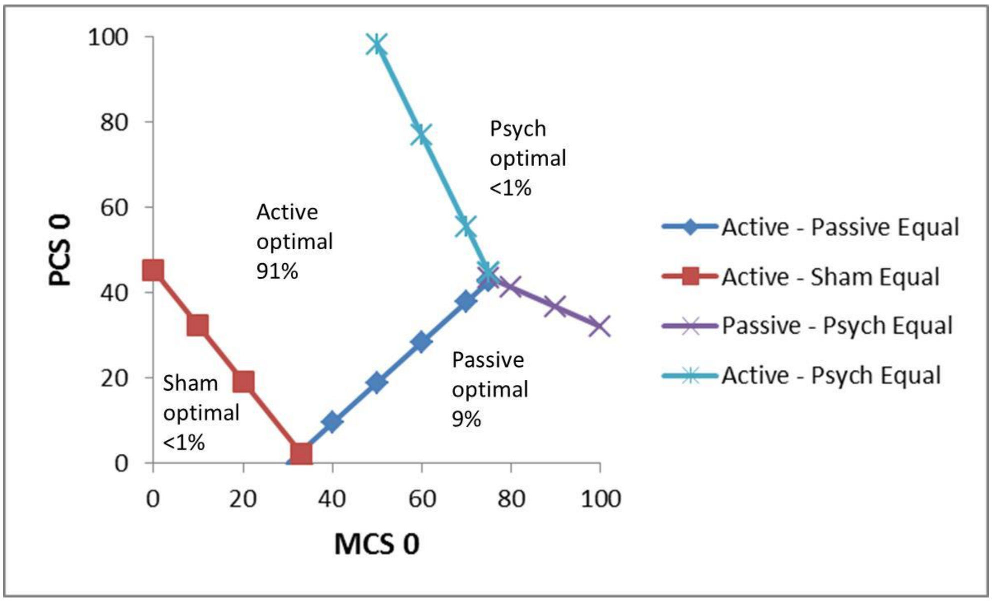


Figure 52 PCS outcome; optimal treatment as a function of MCS and PCS at baseline for men aged 50, with proportion of male participants whose MCS and PCS baseline scores fit into each zone (*n* = 2,296).

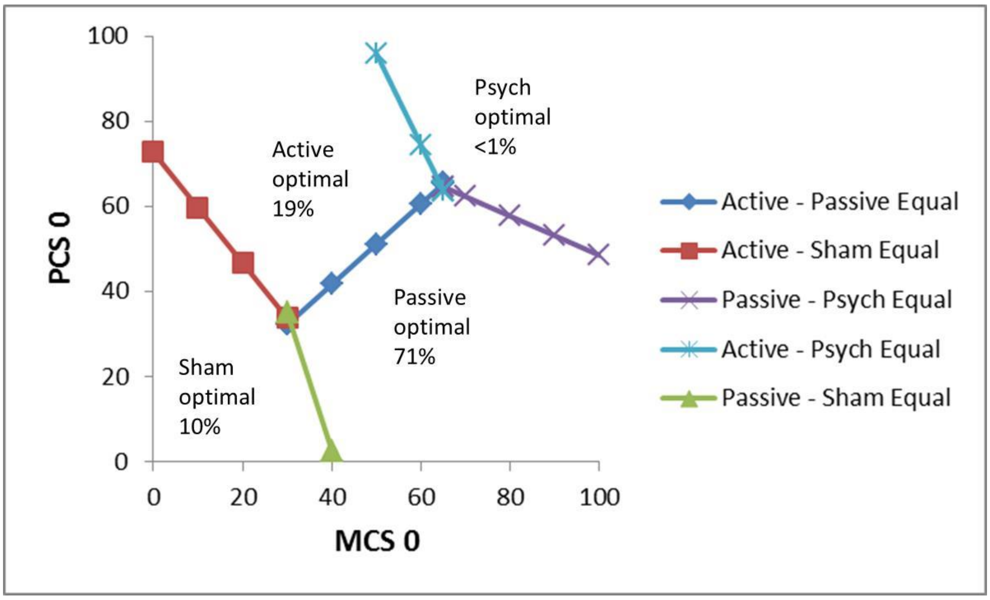


Figure 53 PCS outcome; optimal l treatment as a function of MCS and PCS at baseline for women aged 50, with proportion of female participants whose MCS and PCS baseline scores fit into each zone (*n* = 3,278).

Table 49 Probability that any given treatment is optimal for a range of participant profiles with PCS as outcome of interest.

	Probability that treatment is optimal for this participant profile			
	Active Physical	Passive Physical	Psychological	Sham
Participant profile 1: Male, MCS 40 and PCS 40	81%	11%	7%	<1%
Participant profile 2: Male, MCS 70 and PCS 20	42%	43%	15%	<1%
Participant profile 3: Female, MCS 30 and PCS 50	55%	18%	6%	21%
Participant 4: Female, MCS 60 and PCS 30	23%	68%	9%	<1%
Participant 5: Female, MCS 20 and PCS 20	20%	11%	1%	68%

10.3.3 SHORT-TERM SF-12/36 MCS OUTCOME

The network of evidence for this outcome is the same as for 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 Bayesian Probability above 0.95. It is, perhaps, worth noting here that for short-term MCS as an outcome that for our paradigmatic case it is passive physical therapy that has the largest effect size. At least for the comparison with active physical the 95% credibility interval does not cross zero.

Table 50 Treatment effect with modification (absolute change in short term MCS, mean and 95% credible interval). Coefficients given for individual aged 50, male, PCS and MCS=40. Predicted change in condition without treatment adjusted for age, sex, baseline values of SF-12/36 PCS and MCS.

		Comparator			
		Control	Active Physical	Passive Physical	Psychological
Intervention	Active Physical	1.53 (0.04, 3.02)			
	Passive Physical	3.04 (2.23, 3.85)	1.50 (0.05, 2.96)		
	Psychological	2.59 (0.80, 4.39)	1.06 (-1.04, 3.17)	-0.44 (-2.26, 1.39)	
	Sham	2.13 (0.44, 3.82)	0.60 (-1.53, 2.73)	-0.90 (-2.59, 0.79)	-0.46 (-2.83, 1.90)

Table 51 Mean, 95% credible intervals and Bayesian Probabilities (BP) for impact of participant characteristics on effect of treatments in the network.

	Active Physical	Passive Physical	Psychological	Sham
Age[a]	-0.02 (-0.09, 0.05) *BP* = 74%	-0.03 (-0.07, 0.01) *BP* = 93%	0.00 (-0.06, 0.07) *BP* = 53%	-0.09 (-0.15, -0.03) *BP* > 99%
Sex[b]	0.36 (-1.23, 1.96) *BP* = 67%	-0.20 (-1.18, 0.78) *BP* = 66%	-0.47 (-2.26, 1.34) *BP* = 70%	0.73 (-0.99, 2.44) *BP* = 63%
MCS[a]	-0.06 (-0.13, 0.01) *BP* = 97%	-0.10 (-0.14, -0.06) *BP* > 99%	-0.05 (-0.13, 0.03) *BP* > 90%	-0.17 (-0.24, -0.09) *BP* > 99%
PCS[a]	-0.03 (-0.13, 0.08) *BP* = 68%	-0.08 (-0.14, -0.02) *BP* > 99%	0.05 (-0.04, 0.15) *BP* > 86%	-0.15 (-0.27, -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.

Figure 54 and *Figure 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 PCS scores (low levels of disability) and moderate-to-high MCS scores (low levels of psychological distress). Passive physical therapy is optimal for participants with low PCS scores and high MCS scores, and sham therapy is optimal for participants with low PCS and MCS scores (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 (see *Table 52*) show that, as with RMDQ, there is greater certainty around which treatments are sub-optimal than around which treatments are optimal. However, the evidence for effect modification appears strongest on this outcome. It is 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.

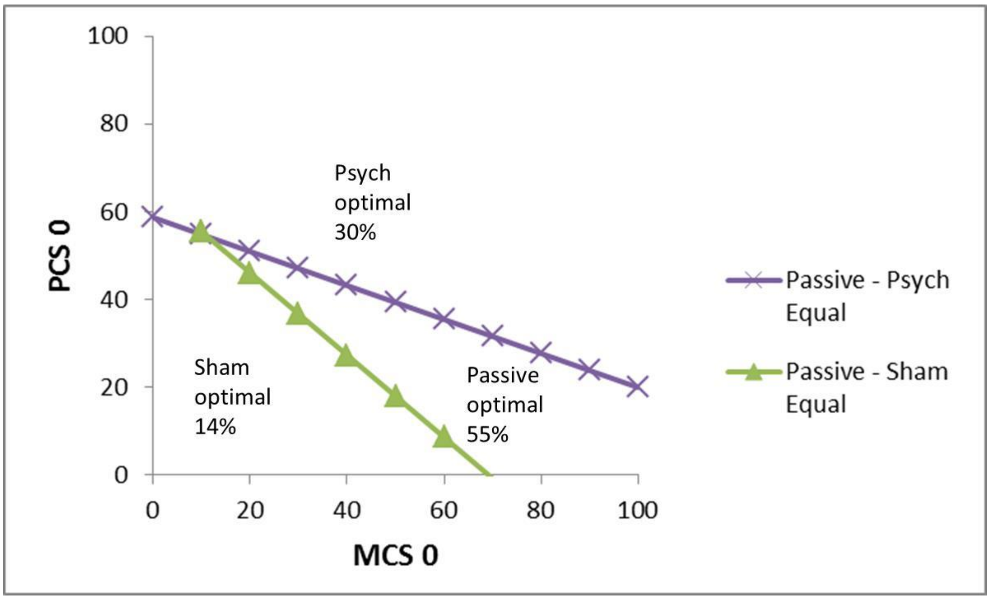


Figure 54 MCS outcome; optimal l treatment as a function of MCS and PCS at baseline for men aged 50, with proportion of male participants whose MCS and PCS baseline scores fit into each zone (*n* = 2,296).

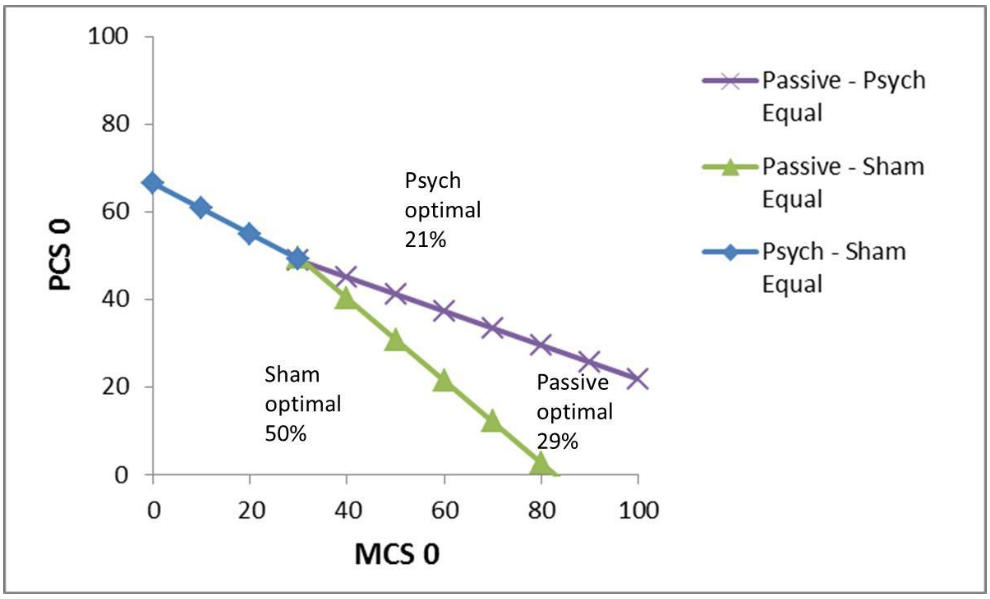


Figure 55 MCS outcome; optimal treatment as a function of MCS and PCS at baseline for women aged 50, with proportion of female participants whose MCS and PCS baseline scores fit into each zone (*n* = 3,278).

Table 52 Probability that any given treatment is optimal for a range of participant profiles.

	Probability that treatment is optimal for this participant profile			
	Active Physical	Passive Physical	Psychological	Sham
Participant profile 1: Male, MCS 60 and PCS 60	6%	<1%	91%	<1%
Participant profile 2: Male, MCS 70 and PCS 20	11%	65%	13%	10%
Participant profile 3: Male, MCS 30 and PCS 30	<1%	31%	<1%	68%
Participant profile 4: Female, MCS 60 and PCS 60	12%	<1%	82%	<1%
Participant profile 5: Female, MCS 80 and PCS 20	26%	32%	15%	11%
Participant profile 6: Female, MCS 80 and PCS 20	<1%	13%	<1%	87%

It is perhaps of note that for some patient 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.

CHAPTER 11 - DISCUSSION

11.1 INTRODUCTION

This work is grounded in the pressing need to improve the outcomes for people living with low back pain. The targeting of treatments of proven but modest average effectiveness at those likely to gain the greatest benefit holds promise. It is the considerable uncertainty over which patients are most likely to benefit from which treatment strategy that was the driver for this research. 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 low back pain; for example the psychosocial 'yellow flags'[155] or the StartBack tool[156]. 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 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 sub-groups of patients with low back pain, within randomised controlled trials 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 sub-groups. A large focus of this work has been very technical on how best to address the challenge of pooling very complex datasets and how best to define sub-groups 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 network meta-analysis.

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. Whilst 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.

11.2 SUMMARY KEY FINDINGS

11.2.1 SYSTEMATIC REVIEWS (CHAPTER 2)

Notwithstanding the perceived importance of performing research to identify subgroups of people living with chronic low back 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. Whilst 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 means 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.

11.2.2 ANCOVA ANALYSES (CHAPTER 6)

Our ANCOVA analyses replicate the conventional approach to moderator identification in a pooled dataset. The main purpose of this analysis was to inform selection of potential moderators for our main analysis based on identifying variables significant at the 20% level.

As with our analyses we were restricted in these analyses by the pool of trials using a common set of baseline covariates and outcomes. In this analysis comparing all intervention groups to 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 physical component score, PCS) have the most to gain from treatment on physical outcomes and those who are worse on the SF-12/36 mental component score (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 that 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 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 where 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 that they should have been identified.

Whilst these observations are of some interest the main purpose of this analyses was to select potential moderators 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 single in one or more analysis.

11.2.3 RECURSIVE PARTITIONING (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 whilst the IPD-SIDES method is seeking to detect groups with the largest treatment effects.[141] 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 where maximising any moderation is the outcome of interest. We have presented both analyses here to explore how they perform on a real dataset. The IPD-SIDES approach appears to be more sensitive as it has successfully identified some subgroups within our data whilst the IPD-IT method did not (see *Table 53, Table 55, Table 56, Table 57 and Chapter 7*). Our overall analysis of all interventions vs 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 vs. control (sham plus usual care) using the full dataset there are some clusters of characteristics with different treatment outcomes. For example, for the outcome FFbHR (range of the score 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 an FFbHR score ≤54.2 and aged ≤60 or for the SF-12/36 Physical Component Score (range 0-100 best) the overall treatments 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 physical component score ≤40.0 and an SF-12/36 mental component score >54.2. It is, however, the pairwise comparisons, with usual care control, that might be useable to inform clinical practice.

11.2.3.1 Passive physical therapy

For passive physical therapy we identified subgroups for the outcomes of FFbHR, plus SF-12/36 – mental and physical component scores. The results for FFbHR, which represent just acupuncture trials, find a maximal effect of 16.67 (95% CI 13.16 to 20.18) when compared to an overall treatment effect of 9.95 (95% CI 8.80 to 11.11) in those aged ≤53 and with an FFbHR ≤54.2. Thus acupuncture is likely to 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 whom the treatment was ineffective and only 17% of participants (571/3272) were in this group with the largest effect. For the SF-12/36 mental component score the maximal effect is seen in those with a low score on both physical and mental component score. In the group with an MCS≤ 54.3 and PCS≤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 (2,171/3,898) fall into this group. Again none of the splits identified a group where the treatment was not effective suggesting it would not be helpful in clinical practice. This could, in any event, only be plausibly clinically important if the outcome of interest was mental health.

For the SF-12/36 physical component score IPD-SIDES found nine candidate models, including one with three splits; baseline physical and mental component scores and gender. The final split on gender did not, however, achieve conventional statistical significance at the 5% level. Several candidate models were identified. All included severity on the physical component score as the first split with either age of mental component score as the second split. Treatment most effective in those with more severe problems and who were younger or had better mental health. There was little to choose 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.

11.2.3.2 Active physical therapy

We did not find any subgroups with an enhanced response to active physical therapy

11.2.3.3 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 an RMDQ 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 an RMDQ score of ≤4 the 95% confidence interval 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, unlike the other outcome measures reported here, there is an established minimally importance change for an individual; 5.0 points.[30] The size of the interaction can be interpreted as a small difference; i.e. 0.21 of the minimally important change.[157] It is nevertheless comparable with the overall effect size at three months identified in the Back Skills Training (BeST) trial (1.1 points on RMDQ 95% CI 0.38 to 1.71) that did not have a lower limit of the RMDQ for study entry.[31] These data can reasonably be used to indicate that psychological treatments should be reserved those with an RMDQ of >4. Interpreting 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 treatment tend to have greater effects in those with worse baseline scores on the outcome of interest

11.2.3.4 Sham treatment

Interpreting the findings for sham treatments, on this occasion sham acupuncture from two trials [102, 117] on the SF-12/36 mental component score is quite challenging. The results of the IPD-SIDES analysis appear to show that for those aged over 65 and for those with an SF-12/36 PCS of greater than 42.0 that sham acupuncture is substantially less effective and that in rest of the population the effect size is enhanced. Whilst the point estimates indicate harm the 95% confidence intervals include zero and, at least for SF-12/36 the interaction effect is of borderline statistical significance (*p*=0.043). It may well be, for age, that we are observing the same phenomena seen for other interventions where 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 over 65 and those with a better physical component score > 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.

11.2.4 ADAPTIVE REFINEMENT BY DIRECTED PEELING IN IPD META-ANALYSIS (CHAPTER 8)

We have successfully extended an adaptive risk group refinement method for use for identifying subgroups of patient who may respond better to different treatments. In contrast to 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 that the additional benefit for selecting subgroups would be clinically worthwhile. This is achieved by repeatedly searching within the dataset to identify successively smaller subgroups with larger effects. This approach does not produce the monotonic changes in subgroup specification seen when a peeling approach *Chapter* 9 is used, but may give a better representation of effect for a pre-specified size of subgroup.

We were limited, by lack of computational power, to just exploring the effect of four co-variates; 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 vs. 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. In contrast no such pattern is seen for the RMDQ outcomes *Figure 25* suggesting that there is not potential for subgroup identification for this group of studies. For the SF-12/36 mental and physical component score outcomes the high variability as subgroup size decreases suggest that it is not possible to define subgroups reliably for these outcomes. Thus for our interpretation of all intervention vs. control (non-active usual care/placebo) is that for the FFbHR outcome younger people with a worse FFbHR and worse PCS may gain more from treatment and that for the SF-12/36 MCS outcome that 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.

11.2.4.1 Passive physical therapy

We found a similar pattern to the overall comparison, for the FFbHR result when passive physical (acupuncture) was compared to non-active usual care; i.e. 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 that those who were younger with worse PCS and MCS gained a greater benefit.

11.2.4.2 Active physical therapy

We did not find any subgroups with an enhanced response to active physical therapy. In particular we did not find that baseline RMDQ consistently identified subgroups with a better treatment effect.

11.2.4.3 Psychological therapy

We did not find any subgroups with an enhanced response to psychological therapy.

11.2.4.4 Sham

We were again able to identify a group who might do better with sham treatment. Its definition was again driven by age and baseline severity. Curiously, a worse baseline mental component score appears to predict who responds better to sham acupuncture but not who responds to true acupuncture.

11.2.5 Identification of cost-effective subgroups by direct peeling (Chapter 9)

The application of the peeling algorithm was successful in identifying potentially interesting subgroups for the interventions vs 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 to the point estimates used in the clinical analysis. This reduced the power of statistical analyses. For the QALY analyses the group who had sham treatment were excluded. Whilst 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) 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 vs. 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, the age 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 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 area under the curve for a sequence of follow up points. However, it is also possible the results are susceptible to missing data bias. Patients with missing EQ-5D data at one or more follow up points were on average four 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 1,774 patients (1271 Intervention; 503 Control) for which there was complete data. Of the 1,774 patient, 1,467 (1,093 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 physical component score 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 the subgroup are likely, on average, to benefit from treatment. Treating only the 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

11.2.6 NETWORK META-ANALYSIS (CHAPTER 10)

In a further methodological development we successfully adapted a network meta-analysis 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 very different format to our other approaches to subgroup identification. Analysing the trials as a single network of evidence allow us to detect subgroup effects with greater precision, and the use of Bayesian methods allows to quantify 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*, that shows that for a paradigmatic case (male, age 50, baseline RMDQ=10, baseline PCS and MCS both equal 40) active physical, passive physical and psychological treatments are all likely to be effective in reducing RMDQ 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 pre-planned 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 physical and mental component scores are weaker and are not consistent across modalities. It is this variability that allows tables of probability for a particular treatment choice being 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 all exclude 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.

Table 53 Overview of results: Intervention versus control (usual care or sham)

METHOD **(section)**		**OUTCOME**[a]						
		physical health			**pain**	**mental health**	**quality of life**	
		FFbHR	RMDQ	PCS	average pain	MCS	EQ-5D	QALY[e]
ANCOVA **(6.6.3)**	positive moderator	none found	moderate catastrophizing[2]; positive fear avoidance[b]	none found	pain[b]; MCS[b]; moderate fear avoidance[b]	none found	female[b]; RMDQ[c] pain[c]; moderate fear avoidance[c]	age[b]; RMDQ[b]
	negative moderators	age[b]; FFbHR[c]; PCS[c]	none found	age[b]; PCS[c]; MCS<50[b]	PCS[b]; low anxiety[b]; positive coping[b];	FFbHR[c]; MCS[c]	PCS[c]; MCS[b]; low/mod anxiety[c]	PCS[b]
recursive partitioning IPD-SIDES **(7.5.1.2)**	subgroups	younger with worse FFbHR	none found	1. better MCS & worse PCS 2 female with worse PCS	none found	worse MCS	none found	none found
directed search[d] **(8.4.1; 9.3)**	subgroups	1. younger with worse FFbHR/ 2. younger with worse PCS	none found	none found	none found	younger with worse MCS	none found	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 variables with *p*<0.2 and >0.05for interactions with treatment effect (FFbHR ANCOVA Age *p*=0.2018); c variables with *p*<0.05 for interactions with treatment effect; d 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; e sham interventions not included in QALY analyses

Table 54 Overview of results: active physical versus control (usual care)

METHOD (section)		**OUTCOME[a]**						
		physical health			**pain**	**mental health**	**quality of life**	
		FFbHR	RMDQ	PCS	average pain	MCS	EQ-5D	QALY[d]
NWMA (10.3)[b]	positive moderators	Not conducted	RMDQ; PCS	none found	Not conducted	none found	Not conducted	Not conducted
	negative moderators	Not conducted	age	PCS	Not conducted	MCS	Not conducted	Not conducted
recursive partitioning IPD-SIDES (7.5.1.2)	subgroups	none found	none found	none found	none found	none found	none found	none found
directed search [c] (8.4.1 & 9.3)	subgroups	none found	none found	none found	none found	none found	none found	none found

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 variables with BP>0.8 for interactions with treatment effect; 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 sham interventions not included in QALY analyses1

Table 55 Overview of results: passive physical versus usual care control

METHOD (section)		OUTCOME[a]						
		physical health			pain	mental health	quality of life	
		FFbHR	RMDQ	PCS	average pain	MCS	EQ-5D	QALY[d]
NWMA (10.3)[b]	positive moderators	Not conducted	men; RMDQ; PCS	women	Not conducted	none found	Not conducted	Not conducted
	negative moderators	Not conducted	none found	PCS	Not conducted	age; PCS; MCS	Not conducted	Not conducted
recursive partitioning IPD-SIDES (7.5.1.2)	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	none found
directed search [c] (8.4.1 & 9.3)	subgroups	younger with worse FFbHR	Not conducted	none found	Not conducted	younger with worse PCS and Worse MCS	Not conducted	none found

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 variables with BP>0.8 for interactions with treatment effect; 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 sham interventions not include in QALY analyses

Table 56 Overview of results: psychological versus usual care control

METHOD (section)		**OUTCOME[a]**						
		physical health			**pain**	**mental health**	**quality of life**	
		FFbHR	RMDQ	PCS	average pain	MCS	EQ-5D	QALY[d]
NWMA (10.3)[b]	positive moderators	Not conducted	RMDQ; PCS; MCS	MCS	Not conducted	PCS	Not conducted	Not conducted
	negative moderators	Not conducted	age	age	Not conducted	MCS	Not conducted	Not conducted
recursive partitioning IPD-SIDES (7.5.1.2)	subgroups	none found	worse RMDQ	none found	none found	none found	none found	none found
directed search [c] (8.4.1 & 9.3)	subgroups	Not conducted	none found	Not conducted	Not conducted	Not conducted	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 variables with BP>0.8 for interactions with treatment effect; 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 sham interventions not include in QALY analyses

Table 57 Overview of results: sham versus control

METHOD (section)		**OUTCOME[a]**						
		physical health			**pain**	**mental health**	**quality of life**	
		FFbHR	RMDQ	PCS	average pain	MCS	EQ-5D	QALY
NWMA (10.3)[b]		Not conducted	men RMDQ	women	Not conducted	none found	Not conducted	Not conducted
	negative moderator	Not conducted	none found	MCS; PCS	Not conducted	age; PCS; MCS	Not conducted	Not conducted
recursive partitioning IPD-SIDES (7.5.1.2)	subgroups	none found	none found	none found	none found	younger with worse PCS	none found	none found
directed search [c] (8.4.1 & 9.3)	subgroups	Younger with either worse FFbHR or PCS	Not conducted	Not conducted	Not conducted	any age; worse PCS; worse 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 variables with BP>0.8 for interactions with treatment effect; 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.

11.2.7 INTERPRETATION

11.2.7.1 Clinical relevance

In our overall analyses (all interventions vs. 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 physical component score of the SF-12/36. For psychological outcomes, as measured by the SF-12/36 mental component score 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 ANCOVA analyses with percentage change from baseline as the dependant variable [Data not shown]. The apparent significance of any moderator effects was substantially reduced; For example, the significance of any effect of any moderation of effect of baseline FFbHR on FFbHR 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 pre-specified approaches, recursive partitioning and ARDP, did produce identifiable subgroups whose parameter definitions were grounded in the data. The differences in effect sizes were, however, generally 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. This overall picture is, however, potentially misleading as the choice is not typically between treatment vs. no treatment; rather it is how to select particular treatments for individuals.

Our pre-specified 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 effects sizes are unlikely to be clinically important. Defining what is clinically important is a challenge for LBP research researchers exploring treatment moderation. The authors of the published protocol for an IPD meta-analysis of studies of exercise treatment for LBP have set minimally clinically important difference for moderation, where p-value is <0.05, to be 20 points on a 100 point scale for pain, ten points on a 100 point scale for disability, or 'another magnitude deemed clinically important by experts'.[158] Others have argued that, for exercise interventions for LBP that a worthwhile between group differences in pain may be as ten points.[159] 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 where the direction of the treatment effect was in favour of the intervention arm in both subgroups. It is open to debate whether a differential subgroup effect that is smaller than a main treatment effect is worthwhile. Where the choice is between treatment, or no treatment, one might expect that to be clinically meaningful any moderator effect should 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 & 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 individual 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 for would benefit from treatment.

A randomised controlled trial of stratified care based on patient’s prognosis using the StartBack 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, whilst a promising overall approach to targeting back pain treatments it does not help us to identify differential subgroup effects in this population.

Currently, treatment choice 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.[160] An alternative approach of using the output from network meta-analysis to help physiotherapists and their patients choose treatment options could be tested empirically.

11.2.7.2 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.[161-166] Our data suggest that, for the interventions assessed here, that 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 where people with more positive attitude (low scores on catastrophizing or low scores on fear avoidance) had greater treatment effect than those with a more negative attitude. In our dataset, 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 that treatment with conventional therapist delivered interventions should be focussed on those with higher levels of psychological distress is not sustainable. What these data cannot tell us whether 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 randomised controlled trial 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 current NICE guidance.

11.2.8 METHODOLOGICAL DEVELOPMENT

A substantial part of the programme grant was around the development of new approaches to identifying subgroups. From our review of literature on subgroups we concluded the existing methods have a number of problems including being severely underpowered, only able to provide exploratory or insufficient findings and have 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 *Chapter 8 & 9*)
3. Individual participant data indirect network meta-analysis (see *Chapter 10*)

These new methods challenge the current paradigm for subgroup identification in which single moderator variables are sought. Whilst 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 multi-parametric approach to subgroup identification that give 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 dataset. We think that this reflects both the limitations of the dataset 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 that we anticipate that have potential across a wide range of clinical areas. Importantly they both use an approach that examines both the effect of variables and provide cut-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 or cost-effectiveness) that would be worthwhile and identify the parameters that would define such a group. For our adaptive approaches we have here just presented 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 network meta-analysis to provide individualised advice on which treatment has the highest probability of being optimal for a particular patient profile is extremely exciting. Whilst no more than exploratory here, as it was not pre-specified 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. Where 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 dataset. This has presented substantial challenges (not fully appreciated at the start of the project) in terms of data management and coding. In contrast to some other areas where individual patient data 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 we have had to do further methodological development in order to develop a new EAV approach to managing such datasets that is far more flexible and simple for non-specialist IT staff to adapt as needed. We think this approach is more robust and flexible than the approach of using an Access database used by others doing IPD meta-analysis of back pain trials.[158] This is an important methodological development that we consider has utility beyond the scope of this project.

Whilst not exactly a methodological development we have examined carefully how one might map between different back pain outcome measures. The important 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 an important finding. This may not be entirely surprising if one examines the time windows over which different measures are considering outcome and the exact content of the measures. We are aware the NIH taskforce on back pain research identified producing cross-walk values for these 'legacy measures' as priority.[130] Our findings demonstrate that this exercise is not worth pursuing further. These findings also mean that existing meta-analyses of back interventions where 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 standardised mean differences of 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 it 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 of field of back pain.

It may well be that the lasting legacy and impact of the programme of work resides in the methodological developments needed to do the analyses rather than the outputs of the analyses.

11.2.9 Strengths

This pooled dataset of randomised control trials of therapist delivered interventions for LBP is a valuable resource for academics and researchers in the field for the future. Such a large dataset provides the statistical power needed for subgroup analyses, something which is 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 3,000 participants to do our analyses. That we have a pooled data set of 9,328 means that we have substantially more statistical power than anticipated. This means that even though for many analyses we were only able to use relatively small sub-sets of the data where the same outcomes had been used that we were still able to perform robust analyses. Whilst 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 datasets 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 we have conducted. In particular we have, wherever possible, included 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 in the future 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, therefore we would encourage academics in the field to approach us with datasets 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 dataset 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 network meta-analysis 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 that gain a greater benefit). Therefore as a strength we can be confident that our analyses are robust yielding the same overall outcome.

11.2.10 LIMITATIONS

Our exploratory work on mapping between outcome measures which measure the same domain, to a common scale led us to conclude that it is not possible to do this and therefore we would be unable to pool outcomes measuring the same domain (see *Chapter 5*). For this reason, despite having a large dataset, for some comparisons we had rather fewer data. As the programme was originally developed we had anticipated using individual item data to help define subgroups. 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. Further, as the work developed, we selected moderators for our analyses grounded in existing data. There is a hazard we would falsely identify moderators as data from three of the fours studies that informed our choice of potential moderators were included in our analyses here. In the event the results were have not identified large subgroup effects and this need not be of great concern. We were only able to explore some of the domains identified in our literature review because in many cases only one study had measured that particular variable and there would be no added value from for running an analysis in the pooled dataset.

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 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 we are evaluating here. We did not have enough detail to include these in our pooled analyses.

From our experience of the BeST[31] and BEAM[34] trials where we know these were measured we have found therapist effects to be negligible and therefore unlikely to be a source of bias.

All of these findings need to be interpreted with some caution. We have done many analyses meaning that some positive findings might have been observed by chance. Also several of the datasets we included in our analyses were also datasets that were used, in other studies, to identify our possible moderators; and were the same dataset we used for our ANCOVA analyses. This again increases the possibility that we might have found a spurious positive result. That, in 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 that with increased psychological distress, as measured by the SF-12/36 Mental Component Score may gain less benefit, thus become a very strong finding.

Our exploratory analytical approach to identifying subgroups who may do best with different treatment approaches using a Bayesian network meta-analysis 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 that 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. Since 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 many argue is inherently a sham treatment; being no more than a theatrical placebo.[167] Even if it is truly a sham treatment it is one that many have belief in that could be offered rather than something no-one has belief in such a de-tuned ultrasound. Whilst 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 the option of sham treatment could be removed and the second choice approach recommended.

11.3 MEANING OF THE RESULTS AND CLINICAL IMPLICATIONS

The important clinical implication of the results is that there is very little clinical or cost-effectiveness justification for using 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 that might help improve treatment selection; this will need empirical testing before it can be recommended. Most importantly we have developed statistical methods for subgroups analysis that move beyond simply looking for interaction effects with single moderator variables. These approaches may have quite wide applicability.

11.4 RECOMMENDATIONS FOR FUTURE RESEARCH

We have made a number of suggestion for further research however these are not necessarily in order of priority.

1. Making the dataset 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 datasets to the repository
   We are aware of two other groups working on intervention specific individual patient data 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 dataset 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 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 low back pain in light of existing developments in the area.

11.5 CONCLUSIONS

The lasting legacy of this work is likely to be the methodological developments need to do our analyses. We have; developed improved systems for storing large, complex datasets; developed methods for assessing comparability of outcome measures that have demonstrated different back pain outcome measures cannot be safely pooled for meta-analyses; we have developed three different approaches to the identification of differential subgroup effects that provide considerable added values compared to conventional analyses that simply test for interactions.

Using frequentist approaches (recursive partitioning or adaptive approaches) has not allowed the identification of subgroups who might have worthwhile additional benefits from different treatment approaches beyond the potential benefits being greater in those with more disability at baseline. Importantly increased psychological distress, as measured using the SF-12/36 mental component score 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 network meta-analysis offers a potential approach to deciding on optimal therapies. We would suggest that these methods are applied in other clinical areas where subgroup identification and targeting of treatment may be advantageous.

Our findings do challenge conventional wisdom on who should be prioritised for back pain treatments; i.e. 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

Funder

This project was funded by the NIHR Programme Grants for Applied Research and will be published in full in Programme Grants for Applied Research Journal; Vol. X, No. X. See the (PGfAR programme website) for further project information.

This project benefitted from facilities funded through Birmingham Science City Translational Medicine Clinical Research and Infrastructure Trials Platform, with support from Advantage West Midlands (AWM) and the Wolfson Foundation.

Contribution of authors

Dr Shilpa Patel (Senior Research Fellow, Health Psychology, Co-applicant) made substantial contribution to the design, organisation and conduct of the programme grant, was responsible for negotiating trial datasets and assisted with interpreting data. She substantially contributed 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 (funded through this programme grant) and made a substantial contribution to the application and interpretation of this method to the datasets obtained in the repository. He drafted the literature review section on subgroups and the results chapter for recursive partitioning as well as commented 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 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 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 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 dataset, interpreted these results and wrote the corresponding chapter for this report.

Dr Tom Morris (Research Fellow, Statistics) substantially contributed to the work on cross walking 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) developing the programming to enable pooling of large datasets. He contributed to 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.

Chief Investigators and data custodians

Our thanks goes to all the Chief Investigators and data custodians who agreed to share their trial data with us for this project. This includes:

- 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 YorkBP)
- Professor Claudia Witt (Witt, Brinkhaus)

Participants

All 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 Trialist Collaboration for assisting with access to trial datasets.
- 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 (Chair), Professor Claudia Witt, Professor Andrea Manca, Dr Richard Riley, Dr Mindy Cairns, Mr Mike Andrews, Mr Chris Phillips.

REFERENCES

1. Andersson GB. Epidemiology of low back pain. *Acta orthopaedica Scandinavica Supplementum* 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. *Annual review of public health* 1991;12:141-56.

3. Dionne CE, Dunn KM, Croft PR. Does back pain prevalence really decrease with increasing age? A systematic review. *Age and ageing* 2006;35(3):229-34.

4. Rapoport J, Jacobs P, Bell NR, Klarenbach S. Refining the measurement of the economic burden of chronic diseases in Canada. *Chronic diseases in Canada* 2004;25(1):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 (Clinical research ed)* 2000;320(7249):1577-8.

6. Raspe H. Back pain. In: Silman A, Hochberg M, editors. Epidemiology of the rheumatic diseases. Oxford: Oxford University Press; 2001. p. 309–38.

7. Raspe H, Hueppe A, Neuhauser H. Back pain, a communicable disease? *International journal of epidemiology* 2008;37(1):69-74.

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(9859):2163-96.

9. Pengel LH, Herbert RD, Maher CG, Refshauge KM. Acute low back pain: systematic review of its prognosis. *BMJ (Clinical research ed)* 2003;327(7410):323.

10. Walker BF. The prevalence of low back pain: a systematic review of the literature from 1966 to 1998. *Journal of spinal disorders* 2000;13(3):205-17.

11. Jeffries LJ, Milanese SF, Grimmer-Somers KA. Epidemiology of adolescent spinal pain: a systematic overview of the research literature. *Spine* 2007;32(23):2630-7.

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 and rheumatism* 2012;64(6):2028-37.

13. IASP. IASP Taxonomy: Pain terms Washington D.C.: International Association for the Study of Pain; 2014 [updated October 06, 2014; cited 2014 13 October 2014]. Available from: http://www.iasp-pain.org/Taxonomy?navItemNumber=576#Pain.

14. BPS. FAQs London: The British Pain Society; 2008 [updated 2008; cited 2014 October 13 2014]. Available from: http://www.britishpainsociety.org/media_faq.htm.

15. Savigny P, Watson P, Underwood M. Early management of persistent non-specific low back pain: summary of NICE guidance. *BMJ (Clinical research ed)* 2009;338: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 (Clinical research ed)* 2013;347: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(9871):997-1020.

18. Waddell G. The back pain revolution. 2nd ed. London: Churchill; 2004.

19. Kent PM, Keating JL. The epidemiology of low back pain in primary care. *Chiropractic & osteopathy* 2005;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. *Occupational and environmental medicine* 2005;62(12):851-60.

21. Thelin A, Holmberg S, Thelin N. Functioning in neck and low back pain from a 12-year perspective: a prospective population-based study. *Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine* 2008;40(7):555-61.

22. Maniadakis N, Gray A. The economic burden of back pain in the UK. *Pain* 2000;84(1):95-103.

23. Dagenais S, Caro J, Haldeman S. A systematic review of low back pain cost of illness studies in the United States and internationally. *The spine journal : official journal of the North American Spine Society* 2008;8(1):8-20.

24. Dunn KM, Croft PR. Epidemiology and natural history of low back pain. *Europa medicophysica* 2004;40(1):9-13.

25. ONS. Full Report: Sickness Absence in the Labour Market, February 2014. 2014 25 February 2014. Report No.

26. Ehrlich G, Khaltaev N. Low back pain initiative. Geneva: World Health Organization; 1999.

27. NICE. Low back pain: Early management of persistent non-specific low back pain. Manchester: National Institute for Health and care Excellence, 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(2):141-4.

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 medical research methodology* 2009;9:35.

30. Ostelo RW, Deyo RA, Stratford P, Waddell G, Croft P, Von Korff 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(1):90-4.

31. 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(9718):916-23.

32. Moore A, Derry S, Eccleston C, Kalso E. Expect analgesic failure; pursue analgesic success. *BMJ (Clinical research ed)* 2013;346:f2690.

33. Lin CW, Haas M, Maher CG, Machado LA, van Tulder MW. Cost-effectiveness of guideline-endorsed treatments for low back pain: a systematic review. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society* 2011;20(7):1024-38.

34. United Kingdom back pain exercise and manipulation (UK BEAM) randomised trial: effectiveness of physical treatments for back pain in primary care. *BMJ (Clinical research ed)* 2004;329(7479):1377.

35. 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 (Clinical research ed)* 2008;337:a884.

36. 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 technology assessment (Winchester, England)* 2010;14(41):1-253, iii-iv.

37. Tilbrook HE, Cox H, Hewitt CE, Kang'ombe AR, Chuang LH, Jayakody S, et al. Yoga for chronic low back pain: a randomized trial. *Annals of internal medicine* 2011;155(9):569-78.

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(9802):1560-71.

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(3):243-51.

40. Turner JA, Holtzman S, Mancl L. Mediators, moderators, and predictors of therapeutic change in cognitive-behavioral therapy for chronic pain. *Pain* 2007;127(3):276-86.

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 practice & research Clinical rheumatology* 2010;24(2):181-91.

42. Lachenbruch PA. A note on sample size computation for testing interactions. *Statistics in medicine* 1988;7(4):467-9.

43. Cochrane Handbook of Systematic Reviews of Interventions2011.

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 medical research methodology* 2011;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 [ISRCTN32683578]. *Rheumatology (Oxford, England)* 2007;46(8):1297-302.

46. Underwood M, Mistry D, Lall R, Lamb S. Predicting response to a cognitive-behavioral approach to treating low back pain: Secondary analysis of the BeST data set. *Arthritis Care & Research* 2011;63(9):1271-9.

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? *The Clinical journal of pain* 2011;27(6):550-5.

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 musculoskeletal disorders* 2009;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. *Archives of internal medicine* 2009;169(9):858-66.

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. *American Journal of Epidemiology* 2006;164(5):487-96.

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(7):618-29.

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. *The American journal of psychiatry* 2001;158(6):848-56.

53. Kent P, Keating JL, Leboeuf-Yde C. Research methods for subgrouping low back pain. *BMC medical research methodology* 2010;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. Reexamining priorities. *Spine* 1998;23(18):1992-6.

55. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA : the journal of the American Medical Association* 1991;266(1):93-8.

56. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005;365(9454):176-86.

57. Lagakos SW. The challenge of subgroup analyses--reporting without distorting. *The New England journal of medicine* 2006;354(16):1667-9.

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. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society* 2012;21(7):1250-6.

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 and rheumatism* 2009;61(9):1202-9.

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(5):473-80.

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. *European journal of physical and rehabilitation medicine* 2012;48(3):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. *The New England journal of medicine* 1998;339(15):1021-9.

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. *Archives of internal medicine* 2001;161(8):1081-8.

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(1):98-108.

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(9476):2024-30.

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. *Annals of the rheumatic diseases* 2009;68(9):1420-7.

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(10):1069-76.

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. *Journal of consulting and clinical psychology* 1999;67(6):931-44.

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. *Archives of physical medicine and rehabilitation* 2007;88(10):1229-35.

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 [ISRCTN22714229]. *BMC musculoskeletal disorders* 2006;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(3):263-76.

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. *The Clinical journal of pain* 2008;24(5):421-30.

73. Bendix AF, Bendix T, Haestrup 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(16):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(8990):1596-600.

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. *The Journal of manual & manipulative therapy* 2011;19(1):20-5.

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. *Disability and rehabilitation* 2005;27(16):929-37.

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 therapy* 2009;14(1):61-7.

78. Glazov G, Schattner P, Lopez D, Shandley K. Laser acupuncture for chronic non-specific low back pain: a controlled clinical trial. *Acupuncture in medicine : journal of the British Medical Acupuncture Society* 2009;27(3):94-100.

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. *European spine journal : official publication of the*

European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society 2006;15(7):1070-82.

80. Hsieh LL, Kuo CH, Yen MF, Chen TH. A randomized controlled clinical trial for low back pain treated by acupressure and physical therapy. *Preventive medicine* 2004;39(1):168-76.

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(3):350-9.

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(15):1578-85.

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(13):1418-24.

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. *Scandinavian journal of rehabilitation medicine* 1989;21(2):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(11):1167-72; discussion 73.

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. *Journal of general internal medicine* 2008;23(2):148-53.

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. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society* 1998;7(6):461-70.

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 (Clinical research ed)* 2006;333(7569):623.

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. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society* 2008;17(9):1193-200.

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 multiaxial assessment instruments: the multidimensional pain inventory and lumbar dynamometry. *Clinical rehabilitation* 2004;18(5):566-79.

91. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine--reporting of subgroup analyses in clinical trials. *The New England journal of medicine* 2007;357(21):2189-94.

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 technology assessment (Winchester, England)* 2001;5(33):1-56.

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(9):762-9.

94. Su X, Tsai C-L, Wang H, Nickerson DM, Li B. Subgroup Analysis via Recursive Partitioning. *Journal of Machine Learning Research* 2009;10:141-58.

95. Dusseldorp E, Meulman J. The regression trunk approach to discover treatment covariate interaction. *Psychometrika* 2004;69(3):355-74.

96. 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. *Statistics in medicine* 2011;30(21):2601-21.

97. 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. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society* 2007;16(11):1812-21.

98. 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(2):65-77.

99. Long A, Donelson R, Fung T. Does it matter which exercise? A randomized control trial of exercise for low back pain. *Spine* 2004;29(23):2593-602.

100. 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(23):2608-15.

101. Underwood MR, Harding G, Klaber Moffett J. Patient perceptions of physical therapy within a trial for back pain treatments (UK BEAM) [ISRCTN32683578]. *Rheumatology (Oxford, England)* 2006;45(6):751-6.

102. 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. *Archives of internal medicine* 2007;167(17):1892-8.

103. 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 randomized clinical trial. *Arthritis and rheumatism* 2007;57(3):466-73.

104. 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. *Archives of internal medicine* 2006;166(4):450-7.

105. 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. *Spine* 2010;35(5):469-76.

106. 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. *Annals of internal medicine* 2007;146(11):787-96.

107. Ratcliffe J, Thomas KJ, MacPherson H, Brazier J. A randomised controlled trial of acupuncture care for persistent low back pain: cost effectiveness analysis. *BMJ (Clinical research ed)* 2006;333(7569):626.

108. 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. *European spine journal : official publication of the European Spine Society, the European*

Spinal Deformity Society, and the European Section of the Cervical Spine Research Society 2008;17(7):936-43.

109. 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(9599):1638-43.

110. 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. *European journal of pain (London, England)* 2009;13(1):51-5.

111. 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(3):323-30.

112. Moore JE, Von Korff M, Cherkin D, Saunders K, Lorig K. A randomized trial of a cognitive-behavioral program for enhancing back pain self care in a primary care setting. *Pain* 2000;88(2):145-53.

113. 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. *Clinical rehabilitation* 2010;24(1):26-36.

114. Moffett JK, Torgerson D, Bell-Syer S, Jackson D, Llewlyn-Phillips H, Farrin A, et al. Randomised controlled trial of exercise for low back pain: clinical outcomes, costs, and preferences. *BMJ (Clinical research ed)* 1999;319(7205):279-83.

115. 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. *Physical therapy* 2012;92(3):363-77.

116. Carlsson CP, Sjolund BH. Acupuncture for chronic low back pain: a randomized placebo-controlled study with long-term follow-up. *The Clinical journal of pain* 2001;17(4):296-305.

117. 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. *Complementary therapies in medicine* 2008;16(3):139-46.

118. Vickers AJ, Cronin AM, Maschino AC, Lewith G, MacPherson H, Foster NE, et al. Acupuncture for chronic pain: individual patient data meta-analysis. *Archives of internal medicine* 2012;172(19):1444-53.

119. Codd EF. The relational model for database management: version 2: Addison-Wesley Longman Publishing Co., Inc.; 1990. 567 p.

120. Marenco L, Tosches N, Crasto C, Shepherd G, Miller PL, Nadkarni PM. Achieving evolvable Web-database bioscience applications using the EAV/CR framework: recent advances. *Journal of the American Medical Informatics Association : JAMIA* 2003;10(5):444-53.

121. XML Technology: World Wide Web Consortium (W3C); 2010 [29 August 2014]. Available from: http://www.w3.org/standards/xml/.

122. 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(10):734-9.

123. Von Korff M, Ormel J, Keefe FJ, Dworkin SF. Grading the severity of chronic pain. *Pain* 1992;50(2):133-49.

124. Kohlmann T, Raspe H. [Hannover Functional Questionnaire in ambulatory diagnosis of functional disability caused by backache]. *Die Rehabilitation* 1996;35(1):I-viii.

125. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy* 1980;66(8):271-3.

126. Tait RC, Chibnall JT, Krause S. The Pain Disability Index: psychometric properties. *Pain* 1990;40(2):171-82.

127. Stratford P. Assessing Disability and Change on Individual Patients: A Report of a Patient Specific Measure. *Physiotherapy Canada* 1995;47(4):258-63.

128. 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: QualityMetric Incorporated; 2002.

129. Ware J, Kosinski M, Dewey J. How to score version 2 of the SF-36 health survey. Lincoln: QualityMetric Incorporated; 2000.

130. 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. The Spine Journal [Internet]. 2014 Aug; 14(8):[1375-91 pp.]. Available from: http://www.thespinejournalonline.com/article/S1529-9430(14)00463-X/abstract.

131. Puhan MA, Soesilo I, Guyatt GH, Schunemann HJ. Combining scores from different patient reported outcome measures in meta-analyses: when is it justified? *Health and quality of life outcomes* 2006;4:94.

132. Agresti A. Categorical Data Analysis. 2nd ed. Hoboken: Wiley; 2002.

133. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159-74.

134. 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 technology assessment (Winchester, England)* 2005;9(32):iii-iv, ix-x, 1-109.

135. Dolan P, Gudex C, Kind P. A social tariff for EuroQol: results from a UK general population survey. Center for Health Economics, 1995.

136. Gray AM, Rivero-Arias O, Clarke PM. Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. *Medical decision making : an international journal of the Society for Medical Decision Making* 2006;26(1):18-29.

137. Rowen D, Brazier J, Roberts J. Mapping SF-36 onto the EQ-5D index: how reliable is the relationship? *Health and quality of life outcomes* 2009;7:27.

138. Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. *Health economics* 2005;14(5):487-96.

139. Melzack R. The short-form McGill Pain Questionnaire. *Pain* 1987;30(2):191-7.

140. Whitehead A. Meta-Analysis Of Controlled Clinical Trials. Senn S, Barnett V, editors. Chichester: John Wiley & Sons; 2003.

141. Mistry D. Recursive partitioning based approaches for low back pain subgroup identification in individual participant data meta-analyses. Coventry: The University of Warwick; 2014.

142. Zhang Z, Singer B. Recursive Partitioning and Applications. 2nd ed. New York: Springer; 2010.

143. LeBlanc M, Crowley J. Survival Trees by Goodness of Split. *Journal of the American Statistical Association* 1993;88(422):457-67.

144. Doyle P. The Use of Automatic Interaction Detector and Similar Search Procedures. *Operational Research Quarterly (1970-1977)* 1973;24(3):465-7.

145. Shih Y-S, Tsai H-W. Variable selection bias in regression trees with constant fits. *Computational Statistics & Data Analysis* 2004;45:595-607.

146. LeBlanc M, Moon J, Crowley J. Adaptive risk group refinement. *Biometrics* 2005;61(2):370-8.

147. NICE. Guide to the methods of technology appraisal. London: National Institute for Health and Care Excellence, 2013.

148. Lumley T. Network meta-analysis for indirect treatment comparisons. *Statistics in medicine* 2002;21(16):2313-24.

149. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ (Clinical research ed)* 2005;331(7521):897-900.

150. 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 in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2011;14(2):371-80.

151. Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. *PharmacoEconomics* 2008;26(9):753-67.

152. Jansen JP. Network meta-analysis of individual and aggregate level data. *Research Synthesis Methods* 2012;3(2):177-90.

153. 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. *Statistics in medicine* 2012;31(29):3840-57.

154. 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. 2011; last updated April 2014.

155. Bouter LM, Pennick V, Bombardier C. Cochrane back review group. *Spine* 2003;28(12):1215-8.

156. 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 and rheumatism* 2008;59(5):632-41.

157. 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. *Journal of clinical epidemiology* 2013;66(2):173-83.

158. Hayden JA, Cartwright JL, Riley RD, Vantulder MW. Exercise therapy for chronic low back pain: protocol for an individual participant data meta-analysis. *Systematic reviews* 2012;1:64.

159. Ferreira ML, Herbert RD, Crowther MJ, Verhagen A, Sutton AJ. When is a further clinical trial justified? *BMJ (Clinical research ed)* 2012;345:e5913.

160. 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 disorders* 2014;15:282.

161. 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. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society* 2007;16(11):1755-75.

162. Kent PM, Keating JL. Can we predict poor recovery from recent-onset nonspecific low back pain? A systematic review. *Manual therapy* 2008;13(1):12-28.

163. 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. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society* 2006;15(11):1633-44.

164. 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(3):245-52.

165. Grotle M, Brox JI, Veierod MB, Glomsrod B, Lonn JH, Vollestad NK. Clinical course and prognostic factors in acute low back pain: patients consulting primary care for the first time. *Spine* 2005;30(8):976-82.

166. 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 (Clinical research ed)* 2008;337:a171.

167. Colquhoun D, Novella SP. Acupuncture is theatrical placebo. *Anesthesia and analgesia* 2013;116(6):1360-3.

APPENDIX 1 – REVIEW 2: SUMMARY OF EXCLUDED PAPERS

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. *Manual Therapy* 2006;11:316-20	Testing a clinical prediction rule
Costa LO, Maher CG, Latimer J, Hodges PW, Herbert RD, Refshauge KM *et al*. Motor control exercise for chronic low back pain: a randomized placebo-controlled trial. *Physical Therapy* 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. *Spine* 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. *Spine* 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. *Journal of Orthopaedic & Sports Physical Therapy* 2006;36:354-63.	Pooled datasets of similar trials
George SZ, Zeppieri G, Jr., Cere AL, Cere MR, Borut MS, Hodges MJ *et al*. A randomized trial of behavioral physical therapy interventions for acute and sub-acute low back pain (NCT00373867). *Pain* 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, Groupp E, Muench J, Kraemer D, Brummel-Smith K, Sharma R *et al*. Chronic disease self-management program for low back pain in the elderly. *Journal of Manipulative & Physiological Therapeutics* 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. *Spine* 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. *European Spine Journal* 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. *BMJ* 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 *et al*. Low back pain in general practice: cost-effectiveness of a minimal psychosocial intervention versus usual care. *European Spine Journal* 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 *et al*. Increasing days at work using function-centered rehabilitation in nonacute nonspecific low back pain: a randomized controlled trial. *Archives of Physical Medicine & Rehabilitation* 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 *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 Technology Assessment (Winchester, England)* /20;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. *Spine* 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. *Spine* 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. *Spine* 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? *Arthritis & Rheumatism* 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. *Spine* 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 *et al*. Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain. *Health Technology Assessment (Winchester, England)* /1/10;9:iii-iiv.	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. *Journal of Orthopaedic Science* 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. *JAMA* 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 is 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: 02476 574664
Email: M.Underwood@warwick.ac.uk or repository@warwick.ac.uk

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 INSTIUTION 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 individual patient data on potential moderators, health outcomes, and health care resource use & costs, from RCTs testing therapist delivered interventions for low back pain. We will conduct statistical and health economic analyses on this pooled dataset.

We will not reanalyse any trial data already published.

Data access is restricted to those named in Table 1 of this agreement. Any changes will be notified to [INSERT INSITUTION NAME].

Table 1 - 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.

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 is held under the guardianship of 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.

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 INSITUTION NAME] and the University of Warwick

Anonymous data will be obtained from [INSERT INSITUTION NAME]. Data will be encrypted and sent to the University of Warwick by [INSERT INSITUTION NAME] via the University's file transfer application.

Once the data has 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 INSITUTION NAME] will provide a detailed description of the variables.

8.0 Storage of Data

The originally 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 onto a failover server and routinely backed up to a Storage Area Network (SAN).

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

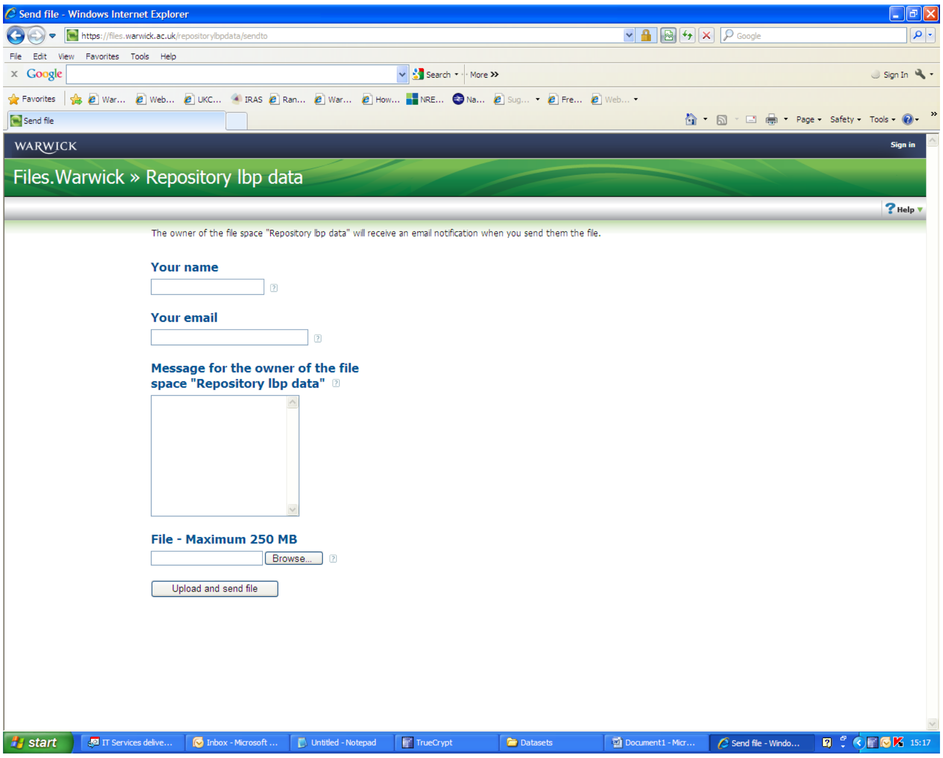
For and on behalf of:	For and on behalf of:
Warwick Clinical Trials Unit	[INSERT INSITUTION NAME]
Signed:	Signed:
Print Name: Professor Martin Underwood	Print Name:
Post/Title: Head of Division of Health Sciences, Warwick Medical School	Post/Title:
Date:	Date:

APPENDIX 5 – INSTRUCTION ON SECURE DATA TRANSFER

Repository Programme

Instructions for transferring datasets to the University of Warwick

- Please ensure your datasets are anonymised.
- Compress/encrypt your dataset using an open-source compression software programme (e.g. 7Zip)
- Follow this link: https://files.warwick.ac.uk/repositorylbpdata/sendto



Send file - Windows Internet Explorer

https://files.warwick.ac.uk/repositorylbpdata/sendto

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 email; 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 email confirming that the dataset have 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, et al. Interventions and physician characteristics in a randomized multicenter trial of acupuncture in patients with low-back pain. J Altern Complement Med 2006;12(7):649-57.	Brinkhaus	301
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(1):51-5.	Hancock	240
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(9599):1638-43.	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. Scand J Rehabil Med. 1989;21(2):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. Scand J Rehabil Med. 1990;22(4):181-8.	Härkäpää	476
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. American Journal of Public Health 2005;95(10):817-1824.	Hurwitz	681

Paper	**Trial**	**Number of participants**
Hurwitz EL, Morgenstern H, Harber P, Kominski GF, Belin TR, Yu F, et al. 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. Spine 2002;27(20):2193-204.	Hurwitz	681
Hurwitz EL, Morgenstern H, Harber P, Kominski GF, Belin TR, Yu F, et al. The effectiveness of physical modalities among patients with low back pain randomized to chiropractic care: Findings from the UCLA low back pain study. Journal of Manipulative & Physiological Therapeutics 2002;25(1):10-20.	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. Spine (Phila Pa 1976) 2006;31(6):611-21; discussion 22.	Hurwitz	681
Koes BW, Bouter LM, van Mameren H, Essers AH, Verstegen GM, Hofhuizen DM, et al. Randomised clinical trial of manipulative therapy and physiotherapy for persistent back and neck complaints: results of one year follow up. Bmj 1992;304(6827):601-5.	Koes	256
Koes BW, Bouter LM, van Mameren H, Essers AH, Verstegen GM, Hofhuizen DM, et al. A blinded randomized clinical trial of manual therapy and physiotherapy for chronic back and neck complaints: physical outcome measures. J Manipulative Physiol Ther 1992;15(1):16-23.	Koes	256

Paper	Trial	Number of participants
Koes BW, Bouter LM, van Mameren H, Essers AH, Verstegen GM, Hofhuizen DM, et al. The effectiveness of manual therapy, physiotherapy, and treatment by the general practitioner for nonspecific back and neck complaints. A randomized clinical trial. Spine 1992;17(1):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. Medical care. 2005;43(5):428-35.	Hurwitz	681
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):1-253, iii-iv.	BeST	701
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. Scandinavian journal of rehabilitation medicine 1989;21(2):91-5.	Härkäpää	459
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 Intern Med 2008;23(2):148-53.	Myers	444
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.	Sherman	638

Paper	Trial	Number of participants
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. Pain 1998;77(2):201-7.	Skargren	323
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. Spine 1997;22(18):2167-77.	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? Arthritis Rheum 2009;61(9):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. Spine (Phila Pa 1976) 2009;34(12):1243-9.	Steenstra	196
Rivero-Arias O, Gray A, Frost H, Lamb SE, Stewart-Brown S. Cost-utility analysis of physiotherapy treatment compared with physiotherapy advice in low back pain. Spine (Phila Pa 1976) 2006;31(12):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 [ISRCTN32683578]. Rheumatology (Oxford) 2007;46(8):1297-302.	BEAM	1,334

Paper	Trial	Number of participants
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 randomized clinical trial. Arthritis Rheum 2007;57(3):466-73.	Whitehurst	402

APPENDIX 7 – TRIALS UNAVAILABLE

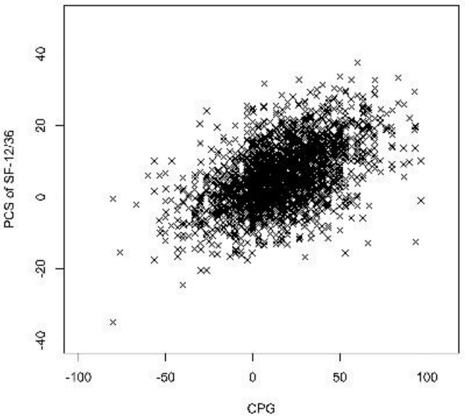
Full reference	Number of participants
Alaranta H, Rytokoski U, Rissanen A, Talo S, Ronnemaa T, Puukka P, et al. Intensive physical and psychosocial training program for patients with chronic low back pain. A controlled clinical trial. Spine 1994;19(12):1339-49.	193
Albaladejo C, Kovacs FM, Royuela A, 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. Spine (Phila Pa 1976) 2010;35(5):483-96.	348
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(3):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. Spine 1989;14(3):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. The New England Journal of Medicine 1998;339(15):1021-29.	321
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 Intern Med 2001;161(8):1081-8.	262
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(9):858-66.	638

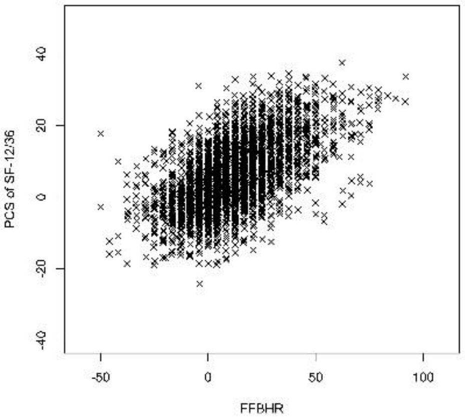
Full reference	Number of participants
Damush TM, Weinberger M, Perkins SM, Rao JK, Tierney WM, Qi R, et al. The long-term effects of a self-management program for inner-city primary care patients with acute low back pain. Archives of internal medicine. 2003;163(21):2632-8.	211
Eisenberg DM, Post DE, 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(2):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. Bmj 2004;329(7468):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. Spine (Phila Pa 1976) 2006;31(10):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. Spine J 2002;2(6):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? Spine (Phila Pa 1976). 2000 Aug 1;25(15):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. Scand J Public Health 2010;38(7):731-8.	246

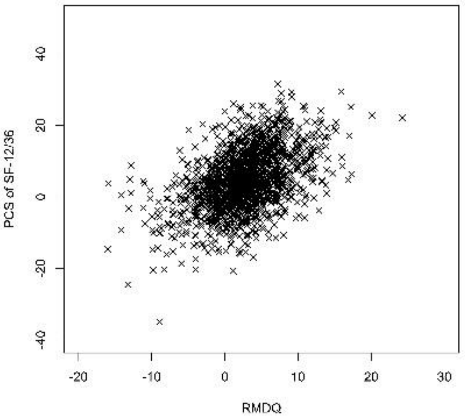
Full reference	**Number of participants**
Heymans MW, de Vet HC, Bongers PM, Knol DL, Koes BW, van Mechelen W. The effectiveness of high-intensity versus low-intensity back schools in an occupational setting: a pragmatic randomized controlled trial. Spine (Phila Pa 1976) 2006;31(10):1075-82.	299
Hondras 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. J Manipulative Physiol Ther 2009;32(5):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. Spine. 2004;29(20):2207-16.	240
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 (Phila Pa 1976) 2007;32(15):1578-85.	196
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: sub-group analysis and relationship between outcome measures. J Manipulative Physiol Ther 1993;16(4):211-9.	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. Spine 2000;25(21):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. Scandinavian journal of rehabilitation medicine 1990;22(4):189-94.	459

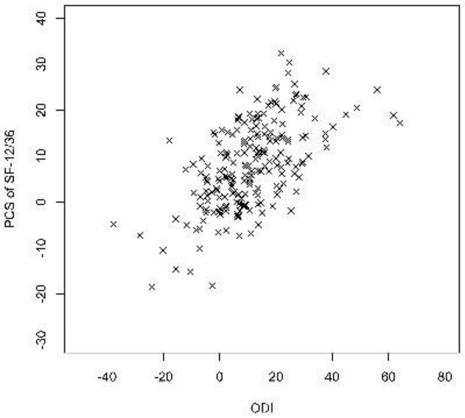
Full reference	Number of participants
Niemisto L, Lahtinen-Suopanki T, Rissanen P, Lindgren KA, Sarna S, Hurri H. A randomized trial of combined manipulation, stabilizing exercises, and physician consultation compared to physician consultation alone for chronic low back pain. Spine 2003;28(19):2185-91.	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. Spine (Phila Pa 1976) 2007;32(26):2948-56.	260
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(8):878-87.	243
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. Spine 1998;23(17):1875-83; discussion 84.	323
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(15):1471-7.	447
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(17):E811-9.	201
Triano JJ, McGregor M, Hondras MA, Brennan PC. Manipulative therapy versus education programs in chronic low back pain. Spine 1995;20(8):948-55.	209

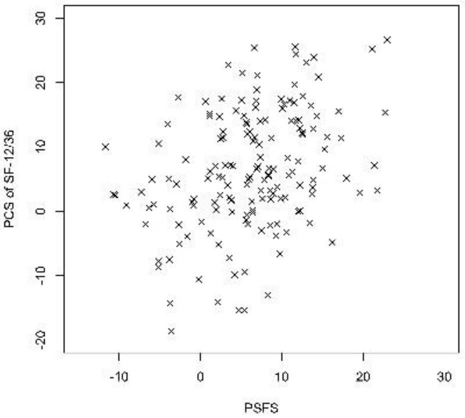
APPENDIX 8 – SCATTER PLOTS OF RAW CHANGE SCORES OF OUTCOME MEASURES



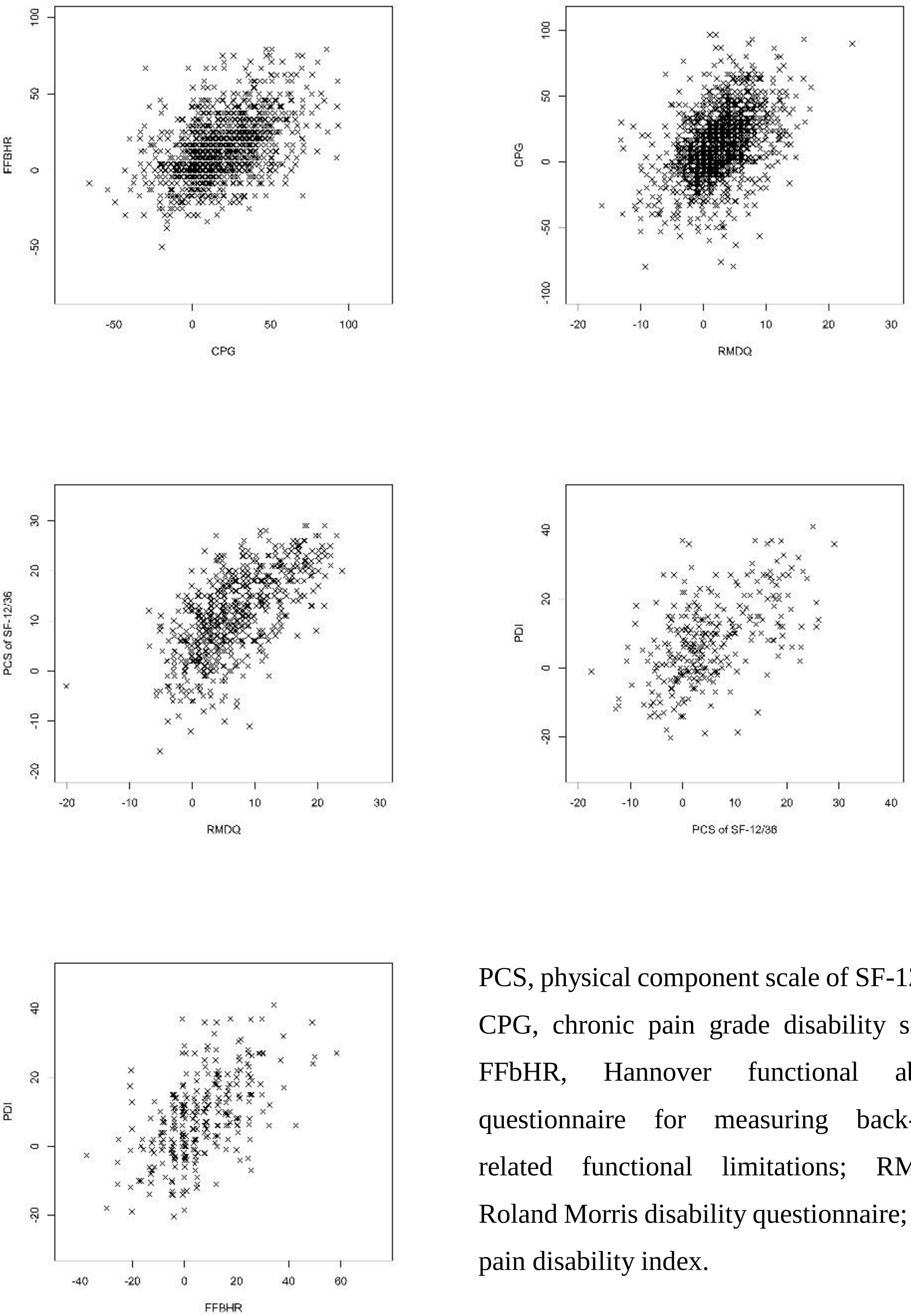








PCS, physical component scale of SF-12/36; CPG, chronic pain grade disability score; FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; RMDQ, Roland Morris disability questionnaire; ODI, Oswestry disability index; PSFS, patient specific functional scale.



PCS, physical component scale of SF-12/36; CPG, chronic pain grade disability score; FFbHR, Hannover functional ability questionnaire for measuring back-pain related functional limitations; RMDQ, Roland Morris disability questionnaire; PDI, pain disability index.

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

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 (Funktionsbeeintrachtigung durch Ruckenschmerzen)
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
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

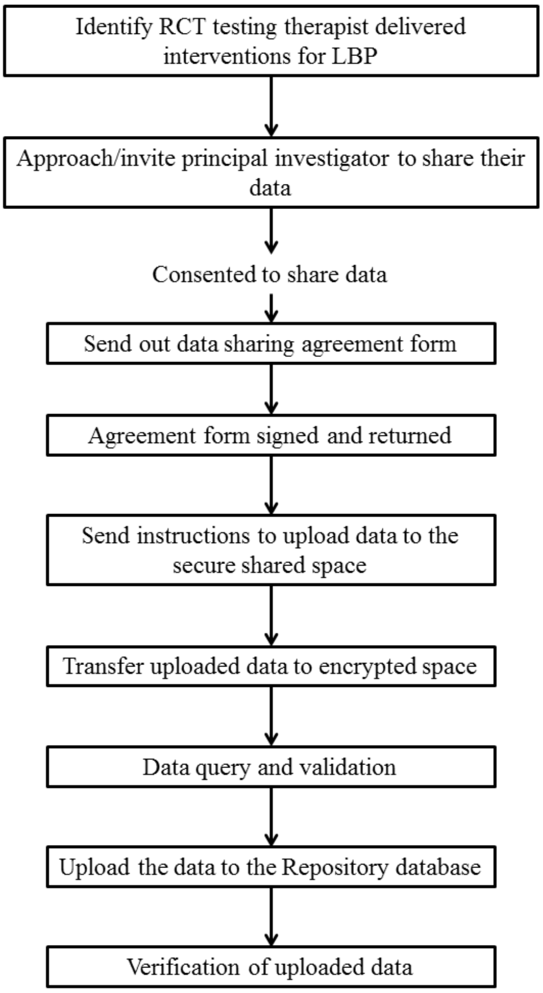
A systematic review was performed to search for RCT of therapist delivered interventions for LBP that identified patient characteristics at baseline that might predict the response to treatments. Variables that were identified from this review are entered into the pool of potential moderators to inform 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	


Identify RCT testing therapist delivered interventions for LBP
Approach/invite principal investigator to share their data
Consented to share data
Send out data sharing agreement form
Agreement form signed and returned
Send instructions to upload data to the secure shared space
Transfer uploaded data to encrypted space
Data query and validation
Upload the data to the Repository database
Verification of uploaded data

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,

$$\text{Disability score} = \text{mean(of the three items)} \times 10.$$

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 (Funktionsbeeintrachtigung durch Ruckenschmerzen)

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

$$\text{ODI score} = \text{Total score from all sections/Total possible score} \times 100.$$

For example, if all 10 sections were completed and the total score was 16, then ODI score was $16/50\times100=32$. However, if one section was missing or not applicable and the total score was also 16 then ODI score was $16/45\times100=35.5$. 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 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,

$$\text{Pain score} = \text{mean(of the three items)} \times 10.$$

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
- 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.

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$_{\text{depression}}$, DRAM, EuroQol (Anxiety/Depression), HADS$_{\text{depression}}$, MZDI, MCS of SF-12/36
Anxiety	DASS-42/21$_{\text{anxiety}}$, EuroQol (Anxiety/Depression), HADS$_{\text{anxiety}}$, 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.

MZDI score = 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.

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).

$$\text{FABQ}_{\text{physical}} = \text{sum(of items 2, 3, 4, and 5).}$$

$$\text{FABQ}_{\text{work}} = \text{sum(of items 6, 7, 9, 10, 11, 12 and 15).}$$

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 catatrophizing (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 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:

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)

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.

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 endogeniety 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:

$$\mathrm{E}\left(NMB_j|P_j\right) = [\,\lambda \times \mathrm{E}\left(E_j|P_j\right) - \mathrm{E}\left(C_j|P_j\right)]$$

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.

The objective function being maximized can therefore be expressed as

$$\mathrm{E}(NMB|P) = (P_A) \times \mathrm{E}(NMB_A|P_A) + (1 - P_A) \times \mathrm{E}(NMB_B|P_B),$$

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

$$\mathrm{E}(NMB|P) = [(P_A) \times \mathrm{E}(NMB_A|P_A)] + [(P_B) \times \mathrm{E}(NMB_B|P_B)] + [(P_C) \times \mathrm{E}(NMB_C|P_C)], (1)$$

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.

$$Starting\ sample\ set\ of\ treatment\ A = P_A^1,$$

$$Starting\ sample\ set\ of\ treatment\ B = P_B^1.$$

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.

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 anaylsis 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**	

Contents

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:

a. Open the "LBP Repository ETL" program.

b. Select the CSV file and the corresponding XML file.

c. Click "Connect".

d. Select server "Palmer", and enter the username and password assigned by the programming team (Mr Ade Willis).

e. Under the field "LBP trial selection", select the name of the trial.

f. Choose either a specific "Class" of data to be uploaded or check "Select all Classes".

g. 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:

a. Open the STATA program

b. Increase memory size by typing in “set memory 1000m” in the command box

c. Click “Enter”

d. 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

e. 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 lick and select "Delete Attribute".

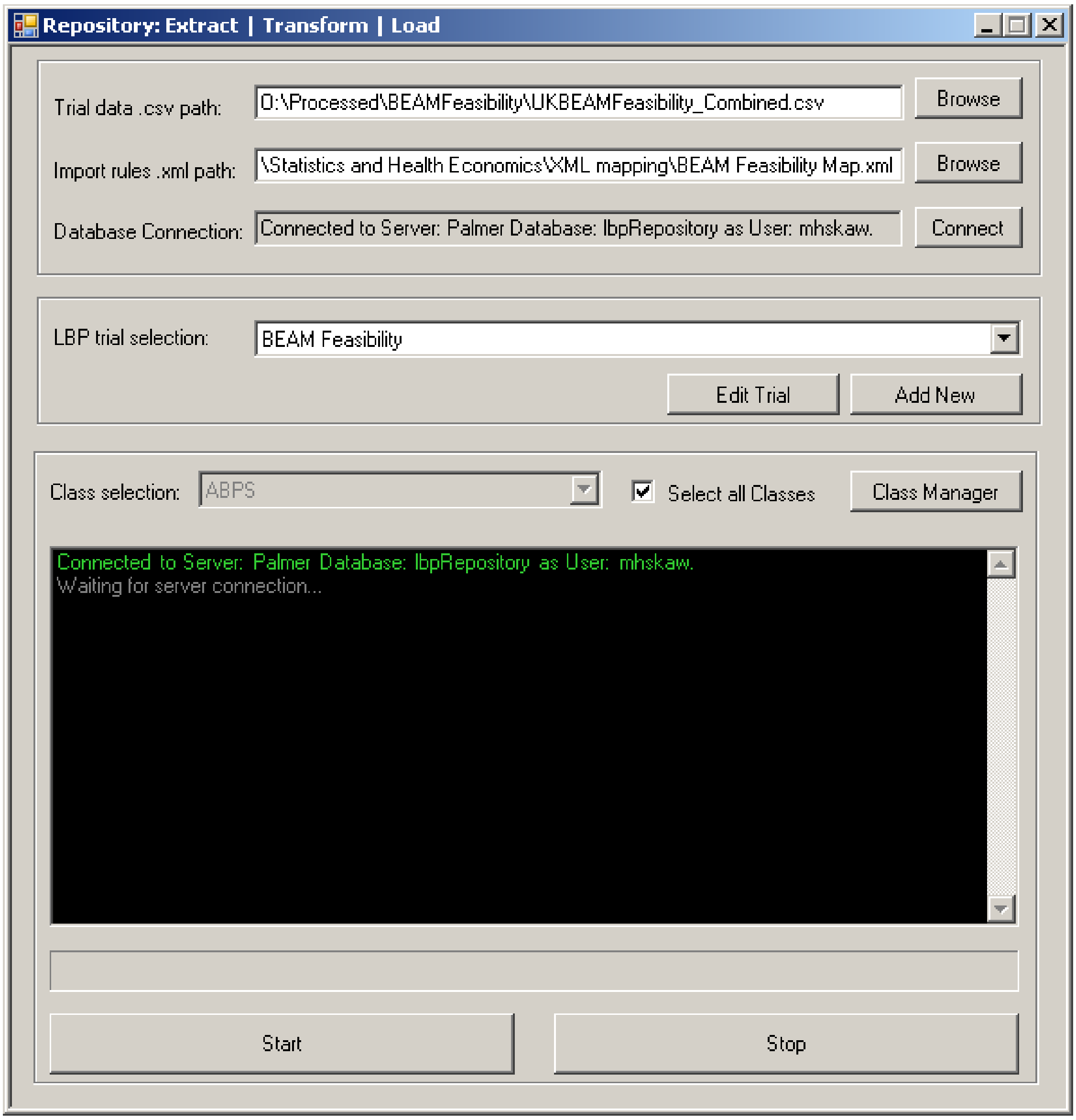
8.9. After all changes have been made, click “Refresh Stat Views”. Email the programming team (Mr Ade Willis) of all the changes that have been made so that they can subsequently update the “Bauer” database.

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



Trial data .csv path: O:\Processed\BEAMFeasibility\UKBEAMFeasibility_Combined.csv

Import rules .xml path: \Statistics and Health Economics\XML mapping\BEAM Feasibility Map.xml

Database Connection: Connected to Server: Palmer Database: lbpRepository as User: mhskaw.

LBP trial selection: BEAM Feasibility

Class selection: ABPS

☑ Select all Classes

```
Connected to Server: Palmer Database: lbpRepository as User: mhskaw.
Waiting for server connection...
```

Figure A.1 The screenshot of the ETL program.

B. Screenshots of SPSS

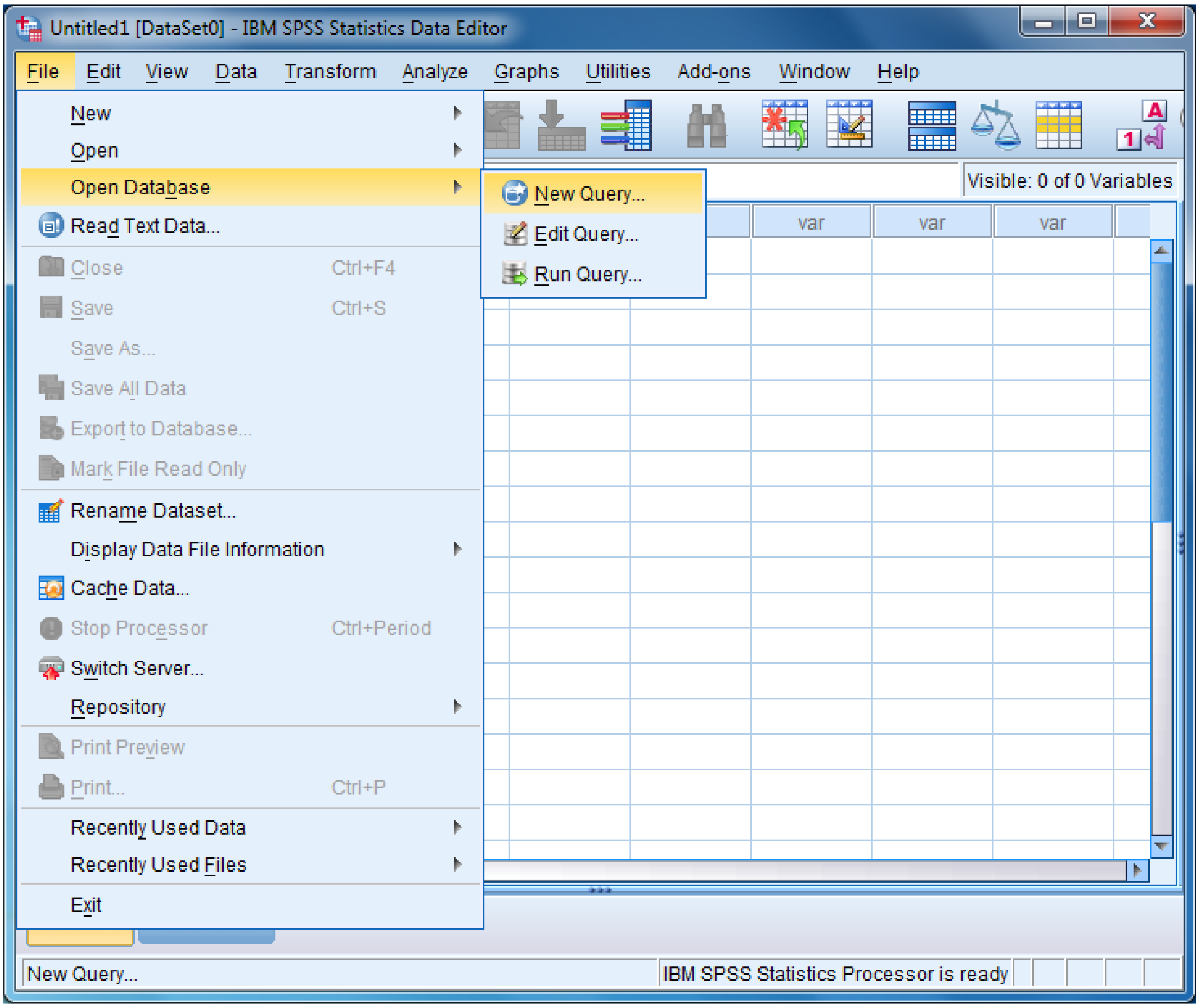
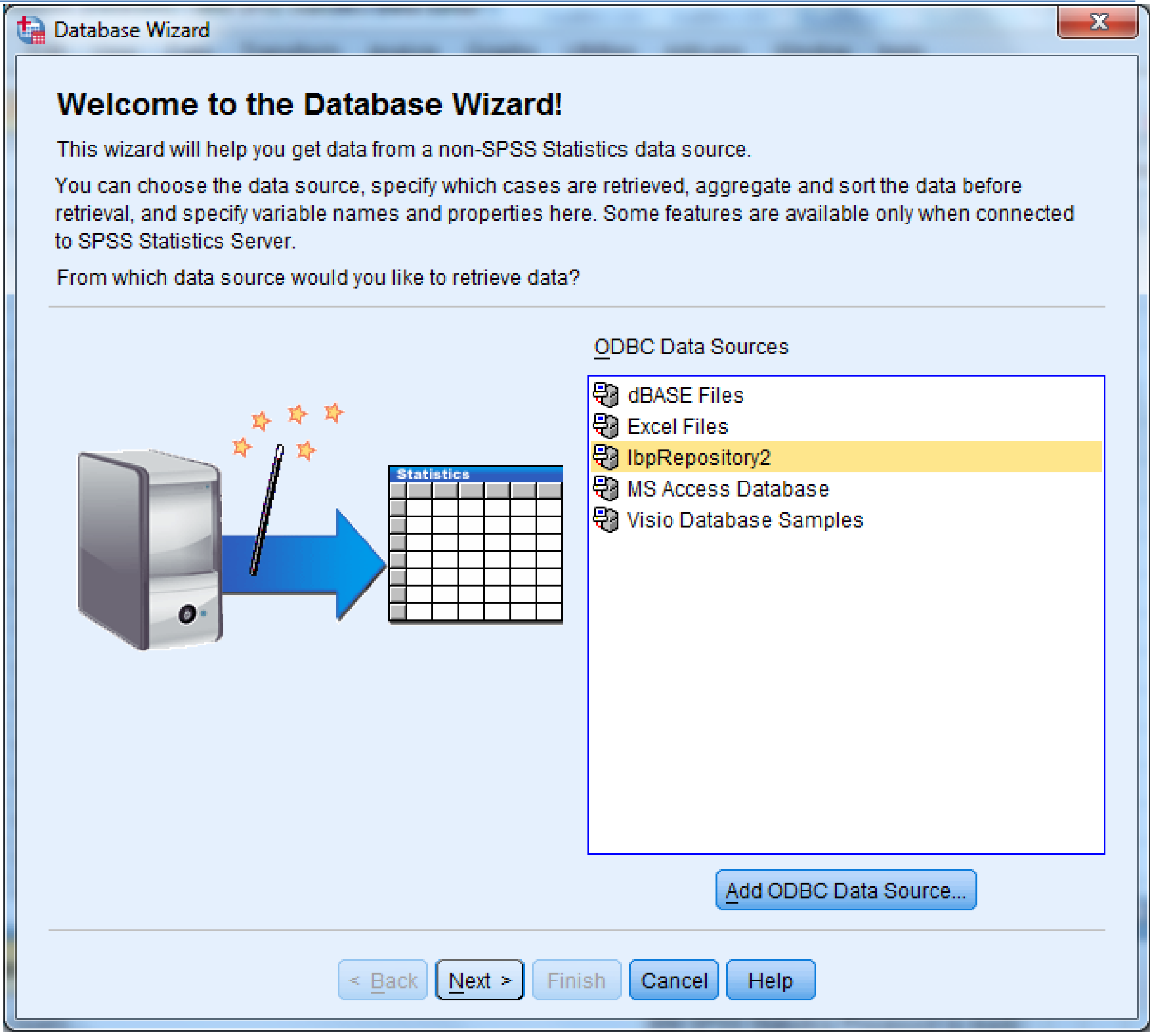


Figure B.1 Screenshot of steps (a) – (b) to access Repository data with SPSS as given in Section 7.4.



Database Wizard

Welcome to the Database Wizard!

This wizard will help you get data from a non-SPSS Statistics data source.

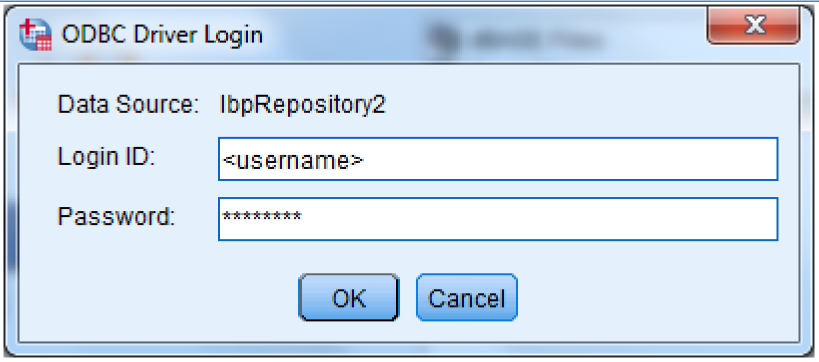
You can choose the data source, specify which cases are retrieved, aggregate and sort the data before retrieval, and specify variable names and properties here. Some features are available only when connected to SPSS Statistics Server.

From which data source would you like to retrieve data?

ODBC Data Sources

- dBASE Files
- Excel Files
- lbpRepository2
- MS Access Database
- Visio Database Samples

Figure B.2 Screenshot of steps (c) – (d) to access Repository data with SPSS as given in Section 7.4.

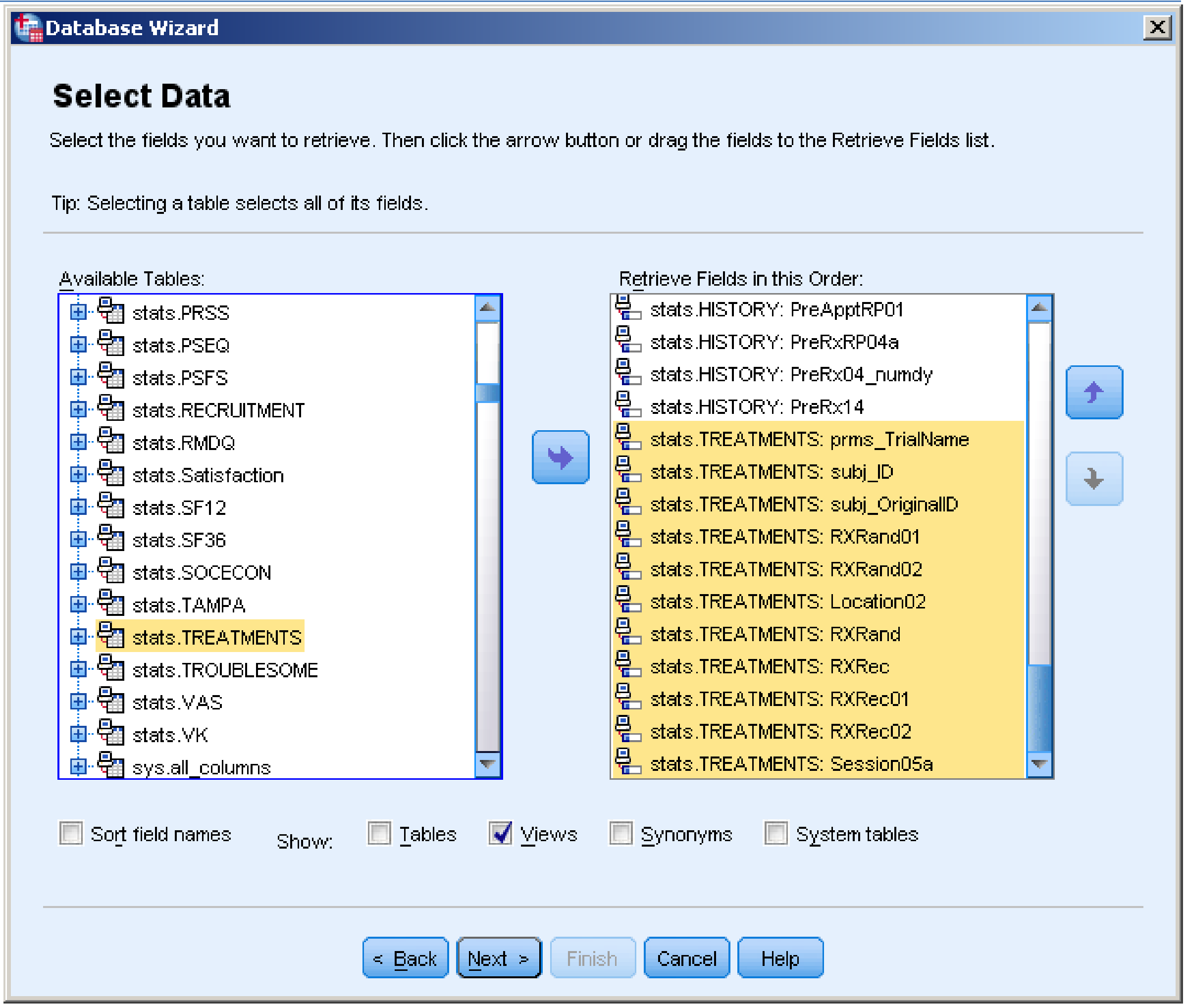


Data Source: lbpRepository2

Login ID: <username>

Password: ********

Figure B.3 Screenshot of steps (e) – (f) to access Repository data with SPSS as given in Section 7.4.



Database Wizard

Select Data

Select the fields you want to retrieve. Then click the arrow button or drag the fields to the Retrieve Fields list.

Tip: Selecting a table selects all of its fields.

Available Tables:

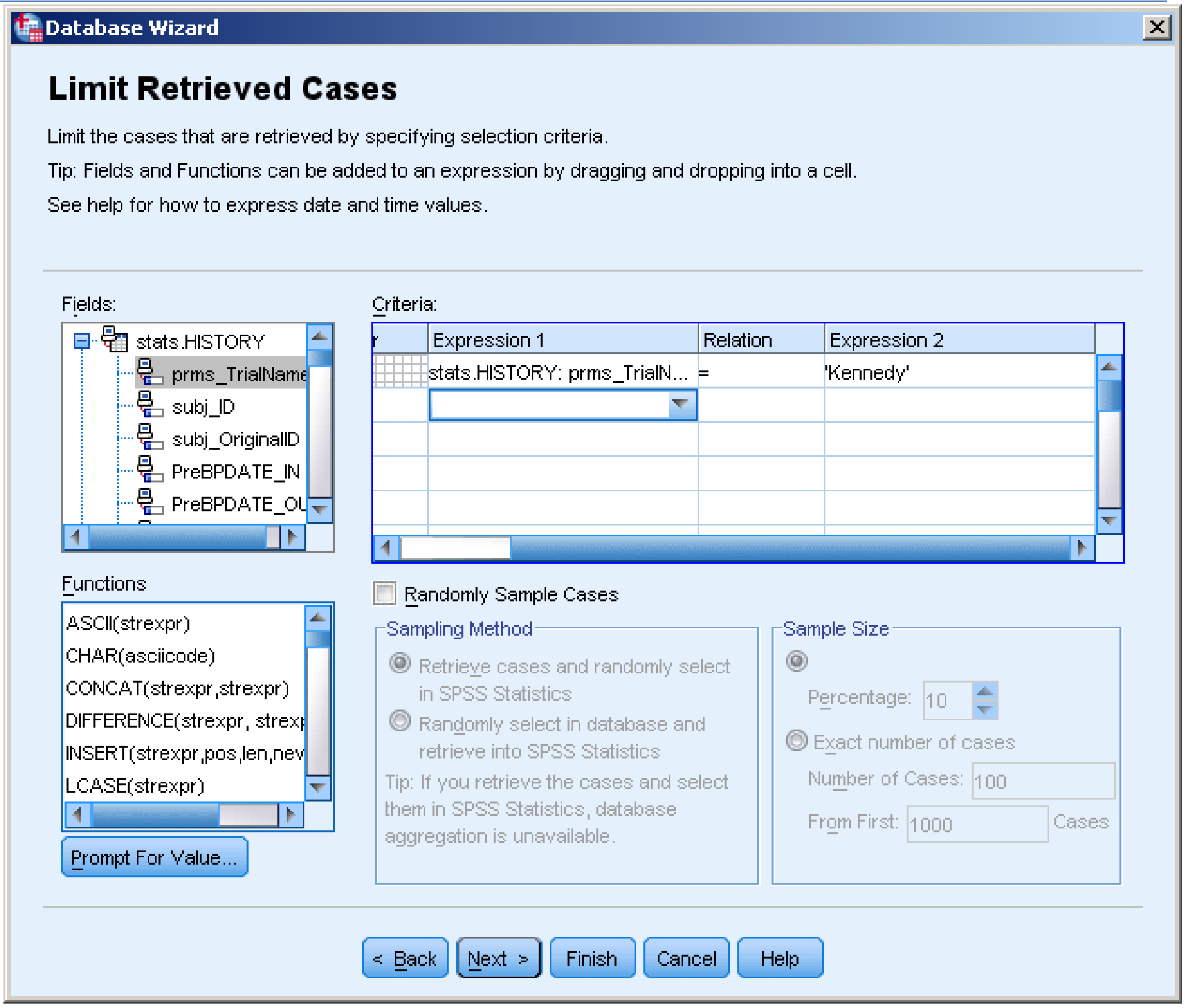
- stats.PRSS
- stats.PSEQ
- stats.PSFS
- stats.RECRUITMENT
- stats.RMDQ
- stats.Satisfaction
- stats.SF12
- stats.SF36
- stats.SOCECON
- stats.TAMPA
- stats.TREATMENTS
- stats.TROUBLESOME
- stats.VAS
- stats.VK
- sys.all_columns

Retrieve Fields in this Order:

- stats.HISTORY: PreApptRP01
- stats.HISTORY: PreRxRP04a
- stats.HISTORY: PreRx04_numdy
- stats.HISTORY: PreRx14
- stats.TREATMENTS: prms_TrialName
- stats.TREATMENTS: subj_ID
- stats.TREATMENTS: subj_OriginalID
- stats.TREATMENTS: RXRand01
- stats.TREATMENTS: RXRand02
- stats.TREATMENTS: Location02
- stats.TREATMENTS: RXRand
- stats.TREATMENTS: RXRec
- stats.TREATMENTS: RXRec01
- stats.TREATMENTS: RXRec02
- stats.TREATMENTS: Session05a

□ Sort field names Show: □ Tables ☑ Views □ Synonyms □ System tables

Figure B.4 Screenshot of steps (g) – (h) to access Repository data with SPSS as given in Section 7.4.



Limit Retrieved Cases

Limit the cases that are retrieved by specifying selection criteria.

Tip: Fields and Functions can be added to an expression by dragging and dropping into a cell.

See help for how to express date and time values.

Criteria:

Expression 1	Relation	Expression 2
stats.HISTORY: prms_TrialN...	=	'Kennedy'

□ Randomly Sample Cases

Sampling Method

● Retrieve cases and randomly select in SPSS Statistics

○ Randomly select in database and retrieve into SPSS Statistics

Tip: If you retrieve the cases and select them in SPSS Statistics, database aggregation is unavailable.

Sample Size

● Percentage: 10

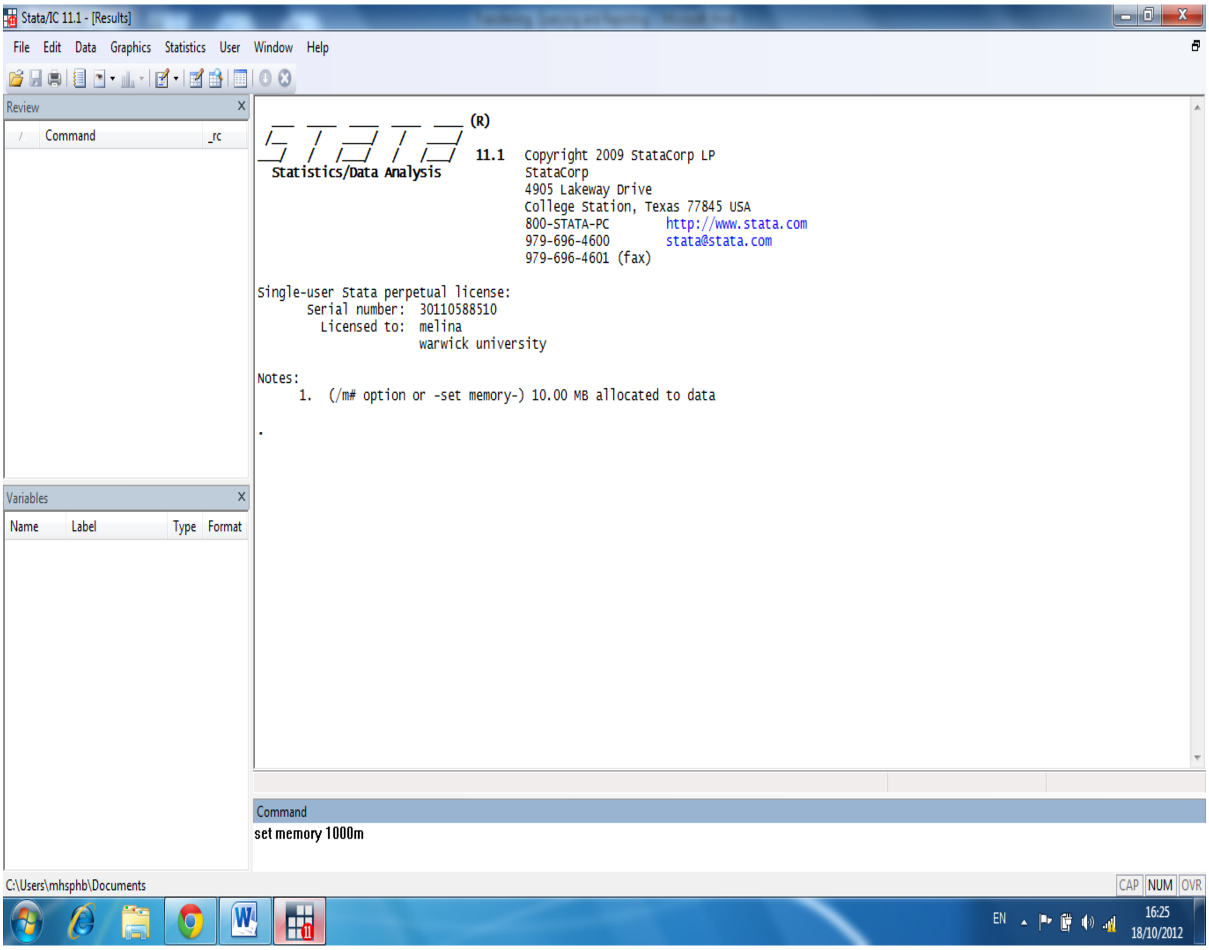
○ Exact number of cases

Number of Cases: 100

From First: 1000 Cases

Figure B.5 Screenshot of step (i) to access Repository data with SPSS as given in Section 7.4.

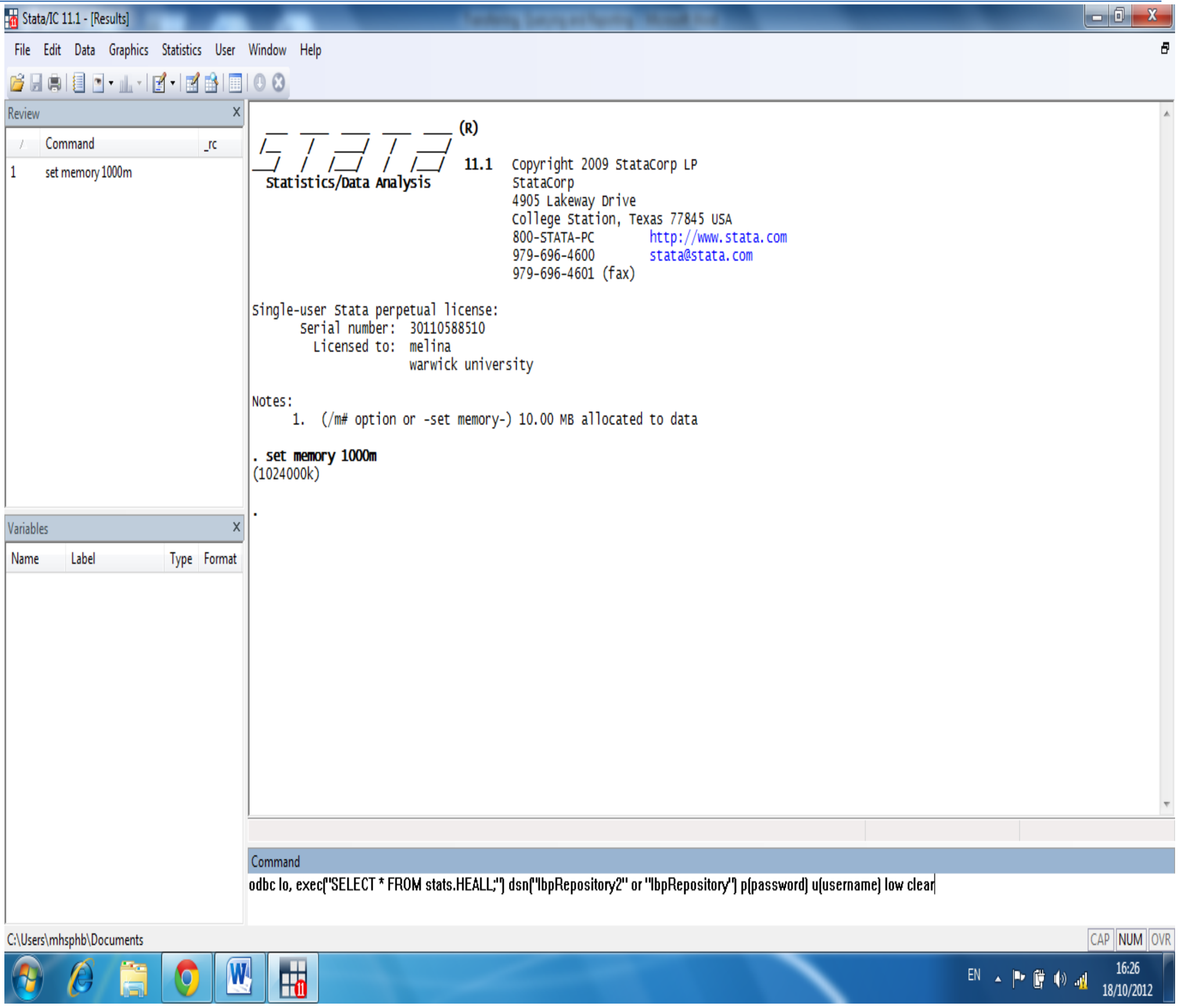
C. Screenshots of STATA



```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   11.1   Copyright 2009 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
                                      College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata perpetual license:
       Serial number:  30110588510
         Licensed to:  melina
                       warwick university

Notes:
      1.  (/m# option or -set memory-) 10.00 MB allocated to data

.
```

```
set memory 1000m
```

Figure C.1 Screenshots of step (a) – (c) to access Repository data with STATA given in Section 7.5.

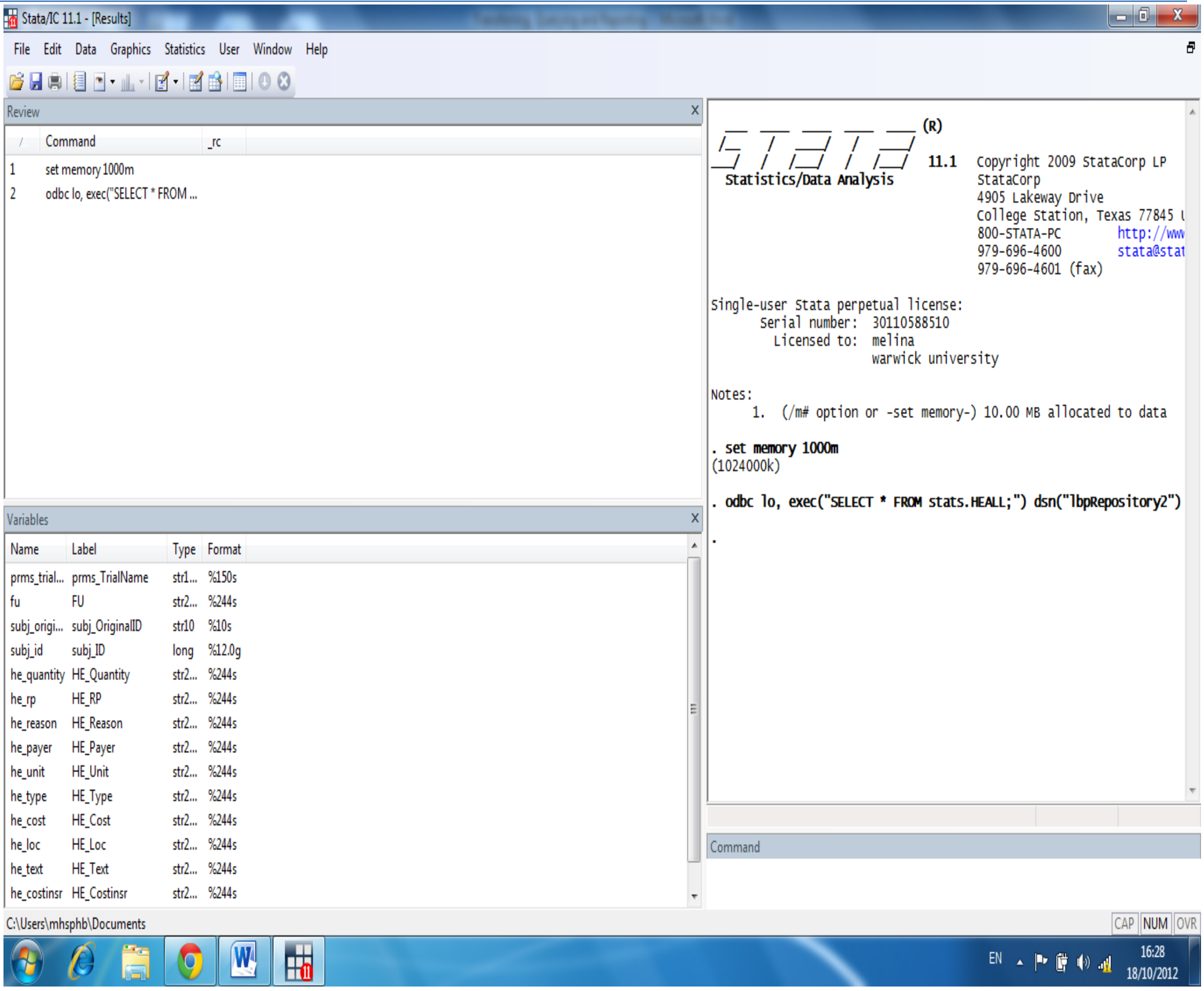

Stata/IC 11.1 - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Command
_rc
Variables
Name
Label
Type
Format
Command
C:\Users\mhsphb\Documents
CAP NUM OVR
EN
16:26
18/10/2012

```
1    set memory 1000m
```

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   11.1   Copyright 2009 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
                                      College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata perpetual license:
       Serial number:  30110588510
         Licensed to:  melina
                       warwick university

Notes:
      1.  (/m# option or -set memory-) 10.00 MB allocated to data

. set memory 1000m
(1024000k)

.
```

```
odbc lo, exec("SELECT * FROM stats.HEALL;") dsn("lbpRepository2" or "lbpRepository") p(password) u(username) low clear
```

Figure C.2 Screenshots of step (d) – (e) to access Repository data with STATA given in Section 7.5.



/	Command	_rc
1	set memory 1000m	
2	odbc lo, exec("SELECT * FROM ...	

Name	Label	Type	Format
prms_trial...	prms_TrialName	str1...	%150s
fu	FU	str2...	%244s
subj_origi...	subj_OriginalID	str10	%10s
subj_id	subj_ID	long	%12.0g
he_quantity	HE_Quantity	str2...	%244s
he_rp	HE_RP	str2...	%244s
he_reason	HE_Reason	str2...	%244s
he_payer	HE_Payer	str2...	%244s
he_unit	HE_Unit	str2...	%244s
he_type	HE_Type	str2...	%244s
he_cost	HE_Cost	str2...	%244s
he_loc	HE_Loc	str2...	%244s
he_text	HE_Text	str2...	%244s
he_costinsr	HE_Costinsr	str2...	%244s

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   11.1   Copyright 2009 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
                                      College Station, Texas 77845 U
                                      800-STATA-PC        http://www
                                      979-696-4600        stata@stat
                                      979-696-4601 (fax)

Single-user Stata perpetual license:
       Serial number:  30110588510
         Licensed to:  melina
                       warwick university

Notes:
      1.  (/m# option or -set memory-) 10.00 MB allocated to data

. set memory 1000m
(1024000k)

. odbc lo, exec("SELECT * FROM stats.HEALL;") dsn("lbpRepository2")

.
```

Figure C.3 Screenshot of Repository data in STATA .