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

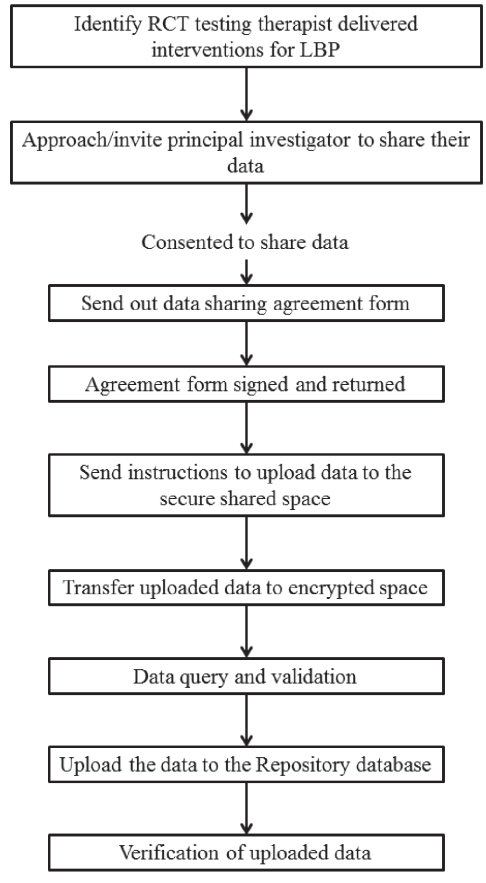
A systematic review was performed to search for RCT of therapist delivered interventions for LBP that identified patient characteristics at baseline that might predict the response to treatments. Variables that were identified from this review are entered into the pool of potential moderators to inform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



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).

FFbHR score = (sum of all items)/24 × 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

ODI score = Total score from all sections/Total possible score × 100.

For example, if all 10 sections were completed and the total score was 16, then ODI score was 16/50×100=32. However, if one section was missing or not applicable and the total score was also 16 then ODI score was 16/45×100=35.5. The range of the score is from 0 (no disability) to 100 (greatest disability).

Pain Disability Index

The Pain Disability Index (PDI) is a measurement of the degree to which pain interferes with functioning in family/home responsibilities, recreation, social activity, occupation, sexual behaviour, self-care, and life-support activities (Tait *et al*., 1990). Each item score ranges from 0 (no disability) to 10 (worst disability).

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

Modified Zung Depression Index

The Modified Zung Depression Index (MZDI) is an instrument that could recognise depressive features and has been highly associated with participant's level of disability (Main *et al*., 1992). It consists of 23 items and participant is to rate how frequent they experience each of the statement recently. The scoring for each item ranges from 0 (less than 1 day per week) to 3 (5 to 7 days per week). The scoring for items 2, 6, 7, 12, 14, 16, 18, 20, 21 and 23 is reversed.

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.

ALBPSQ score = 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).

$FABQ_{physical}$ = sum(of items 2, 3, 4, and 5).

$FABQ_{work}$ = 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:

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

Two comparators, treatment A vs. B

In the simple case, one treatment of interest (B) will be compared to a control of usual care (or best current practice) (A). Let P_j denote the proportion of the total population P treated with intervention j (j = A, B), ranging from 0 to 1. The treatment options are considered exhaustive and mutually exclusive. Therefore, the subsets of the population given each treatment can be defined in terms of one another; $P_B = P - P_A$. There will be a minimum sample size equal to 10% of P, denoted by $P_{10\%}$.

Let us consider the peeling algorithm to maximize expected NMB across the total population P. The starting case is that the maximum number patients receive treatment B. Based on the moderators of interest, the peeling algorithm will iteratively reduce the sample receiving treatment B provided a higher expected NMB across the whole population (P) can be achieved. This process will continue until the expected NMB can no longer increase, or the minimum sample size of $P_B = P_{10\%}$ is reached.

As the algorithm reduces the size of the subgroup (P_B) for treatment B by 10%, the subgroup (P_A) for treatment A will be increased in size by 10%. The 10% will be made up of patients with the same characteristics as those removed from B, defined by the treatment modifier criteria. By weighting the E(NMB) by P_j for each treatment a representative total E(NMB) across the total population is estimated.

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

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

where P_A and P_B satisfied these conditions; $P_A \geq P_{10\%}$, $P_B \geq P_{10\%}$, $P_C = 0$, and $P_A + P_B + P_C = 1$.

At this point the optimal subgroup between P_A and P_B has been determined excluding treatment C. This has determined the starting subgroups for the next round of optimization.

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

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

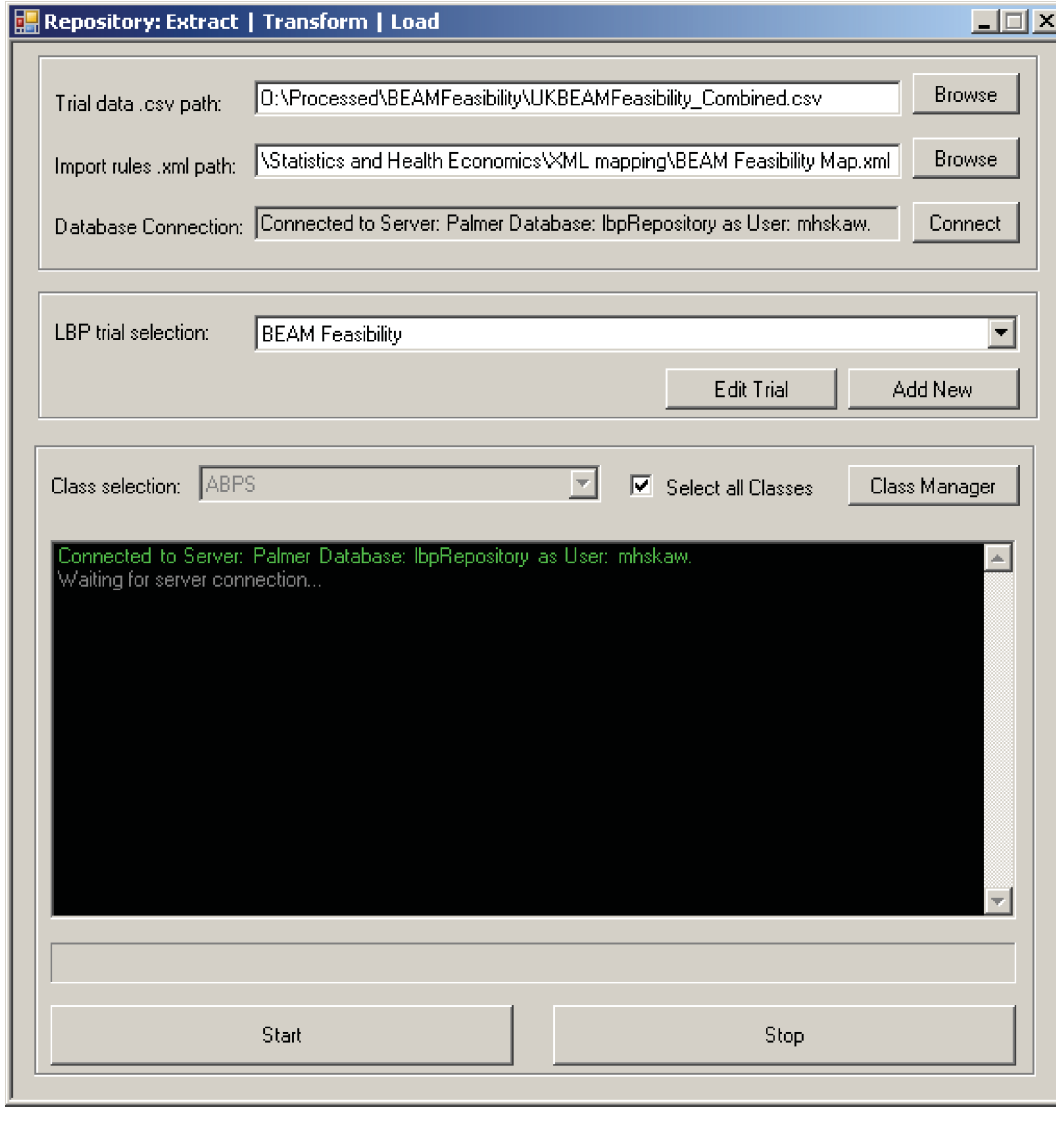


Figure A.1 The screenshot of the ETL program.

B. Screenshots of SPSS

Figure B.1 Screenshot of steps (a) – (b) to access Repository data with SPSS as given in Section 7.4. This page has been left intentionally blank. For a copy of the screenshots, please contact the corresponding author.

C. Screenshots of STATA

Figure C.1 Screenshots of step (a) – (c) to access Repository data with STATA given in Section 7.5. This page has been left intentionally blank. For a copy of the screenshots, please contact the corresponding author.