National Institute for Health and Care Excellence

Draft

Chronic pain: assessment and management

Cost-effectiveness analysis: Cost effectiveness of exercise in people with chronic primary pain

NICE guideline

Economic analysis report

August 2020

Draft for Consultation

This guideline was developed by the National Guideline Centre

Disclaimer

The recommendations in this guideline represent the view of NICE, arrived at after careful consideration of the evidence available. When exercising their judgement, professionals are expected to take this guideline fully into account, alongside the individual needs, preferences and values of their patients or service users. The recommendations in this guideline are not mandatory and the guideline does not override the responsibility of healthcare professionals to make decisions appropriate to the circumstances of the individual patient, in consultation with the patient and, where appropriate, their carer or guardian.

Local commissioners and providers have a responsibility to enable the guideline to be applied when individual health professionals and their patients or service users wish to use it. They should do so in the context of local and national priorities for funding and developing services, and in light of their duties to have due regard to the need to eliminate unlawful discrimination, to advance equality of opportunity and to reduce health inequalities. Nothing in this guideline should be interpreted in a way that would be inconsistent with compliance with those duties.

NICE guidelines cover health and care in England. Decisions on how they apply in other UK countries are made by ministers in the Welsh Government, Scottish Government, and Northern Ireland Executive. All NICE guidance is subject to regular review and may be updated or withdrawn.

ISBN
[to be added on publication]

Contents

1 Introduction

A systematic review of the published clinical and economic evidence was undertaken as part of the guideline, comparing different forms of exercise (the majority were supervised exercise) to usual care, and also comparing types of exercise to each other (full details including the committee's discussion are in evidence report E). This showed a benefit of exercise compared to usual care in reducing pain and improving quality of life. When comparing types of exercise compared to each other, there was less evidence and it was difficult to draw conclusions about a hierarchy of types of exercise.

One UK economic evaluation was identified for this review comparing exercise to treatment as usual.[6] This was a within-trial analysis with the intervention being a gym-based exercise program (gym membership provided), and 6 fitness instructor-led monthly sessions, for a duration of 6 months. The committee view was that this study was quite different to most of the other studies in the clinical review, which tended to be structured class-based interventions, generally group based, with varying frequency/intensity. The economic evaluation found that at follow up (30 months) exercise was not cost effective in the base case analysis using complete case data, but it was cost effective when using imputed data. A second Spanish economic evaluation was identified, which was a within trial analysis comparing 8 months of group pool-based exercised to usual care. This found exercise to be cost effective, although the staff costs were very low compared to UK costs so cost effectiveness was uncertain from this study. Pool-based exercises are not considered to be current practice in the UK because they have higher costs. Both studies had limitations regarding their generalisability because of the types of interventions analysed, and uncertainty remained around cost effectiveness.

The committee consensus was that, currently, exercise is sometimes offered as part of the management for chronic pain. At present, promotion of exercise to treat chronic pain, functional impairment and co-morbidities varies widely in different settings of care. Variability in the uptake of exercise may also vary because this could be a difficult topic for people with chronic pain and their clinicians to discuss. Therefore, a recommendation could have a resource impact given the large size of the population living with chronic pain.

For the above reasons, this area was prioritised for new economic modelling.

2 Methods

2.1 Model overview

A cost-utility analysis was undertaken where lifetime quality-adjusted life years (QALYs) and costs from a current UK NHS and personal social services perspective were considered. Discounting was applied in line with NICE methodological guidance; this specifies a rate of 3.5% per annum for costs and QALYs (although note that costs were not incurred in this analysis beyond 1 year and so did not require discounting).[9] An incremental analysis was undertaken.

2.1.1 Comparators

The comparators selected for the model were:

1. Exercise
2. No exercise

It was assumed that both groups receive the same other care.

In the clinical review, different types of exercise were analysed separately. Evidence was sought comparing different types of exercise with each other. However, there was relatively little evidence comparing different types of exercise, and the committee decided there was insufficient evidence to draw conclusions about whether one type of exercise was better than another. Given this, the committee agreed it was not appropriate to compare different types of exercise to each other in the economic analysis, and the analysis should consider exercise versus no exercise based on pooled data from all types of exercise.

The committee discussed the many differences between the interventions in the studies in terms of the type of exercise, intensity (i.e. frequency, duration, and total number of sessions), the staff delivering the exercise, and the varying descriptions of usual care between studies. However, noting all the complexities, the committee agreed that pooling the data would give a more reliable overall estimate of the likely cost effectiveness of exercise. Clearly, the results would need to be interpreted with caution given the heterogeneity in the data created by pooling different interventions that might have different costs. In general, assessing complex interventions or programmes is difficult because every study is likely to define things differently, which increases uncertainty in the results because of heterogeneity. However, pooling data can also decrease uncertainty in the results. See the approach to modelling section for more discussion.

2.1.2 Population

The population for the cost-effectiveness analysis was people with chronic primary pain aged 16 or over.

The specific populations included in individual trials identified in the clinical review varied, but were predominantly either fibromyalgia or chronic neck pain. The populations were pooled in the clinical review, and this approach was also taken for the economic analysis. The committee agreed that these populations are likely to be generalisable to the wider chronic primary pain population, as the general approach throughout the guideline has been that the response to treatment would be sufficiently similar across conditions to allow generalisability of evidence across all chronic primary pain conditions, even when evidence was available for only 1 condition.

2.2 Approach to modelling

Incremental lifetime costs and QALYs per person for exercise compared to no exercise were calculated based on data from randomised controlled studies identified by the systematic review of the clinical evidence that reported appropriate quality of life (QoL) data.

The clinical evidence showed that exercise reduced pain and improved quality of life. Mortality is not impacted by treatment. Although exercise can have an effect on mortality through the wider benefits of exercise, the focus in this model is on the impact of the interventions on the symptoms of the condition itself. The differences in QALYs between exercise and no exercise in the model would be driven by differences in QoL alone. In economic evaluation, a particular measure of QoL is required known as a utility. The analysis is therefore based on studies from the clinical review that reported utilities (EQ-5D), or the SF-36 that could be mapped to utilities (see section 2.3.2.1 for more detail). The available data on the difference in utility between exercise and no exercise were combined with assumptions about what was likely to happen to treatment effect beyond the follow-up in the trials, to calculate the average QALY gain with exercise compared to no exercise. This is described in detail in section 2.3.2. An alternate base case did not extrapolate beyond the trial data.

The key difference in costs was agreed to be those related to delivering an exercise programme. No other costs were incorporated in the analysis. The committee discussed how other resource use, and therefore costs, could be reduced by an effective intervention, from their own experience, as this could reduce healthcare visits for example, however there was limited evidence on this. Only one study in the clinical review reported use of healthcare services. In this study, GP and specialist visits increased, but the confidence limits crossed the line of no difference, whereas physiotherapy visits reduced in the short and longer term, albeit with some uncertainty about the effect size. The included UK economic evaluation also reported other resource use at the follow up timepoint, and this showed a decrease for all resource use in the exercise group except for inpatient admission days, which increased with exercise. There remains uncertainty particularly about whether any change in resource use is related to chronic primary pain, and (on the available data) whether exercise increases or reduces resource use. Due to this uncertainty, no costs other than the cost of exercise itself have been included in the model, as this would have required assumptions in one direction or the other as to whether exercise increases or decreases other resource use. Threshold analyses have however been undertaken on cost. The average resource use from the interventions in each study was identified and costed, and an overall weighted average cost calculated, weighting by the number of participants analysed in each study. This is described in detail in section 2.3.3.

Costs and QALYs were combined to derive the overall cost effectiveness of exercise in a chronic primary pain population.

Pooling different types of exercise

It was acknowledged that different interventions may have different costs, and it was agreed that using pooled costs based on the interventions in the clinical studies in combination with the pooled treatment effects was the most appropriate approach.

The committee discussed whether the analysis should try and account for the potential for a relationship between intervention intensity (and so treatment cost) and treatment effect. But it was agreed that as the clinical review hadn't established the existence and nature of that relationship, (e.g. if it is the intensity or the frequency of exercise, or the fact that people meet with other people that have the same condition that has an effect), it is not known what it is specifically about exercise that improves outcomes. On that basis, it was not considered appropriate to explore this only in the economic analysis.

The committee discussed the limitations of pooling the studies given the differences between them and considered whether analysis of individual studies would be useful given potentially different costs and benefits. However, the committee agreed that analysis at individual study level would not be helpful as it may lead to over interpretation of individual studies.

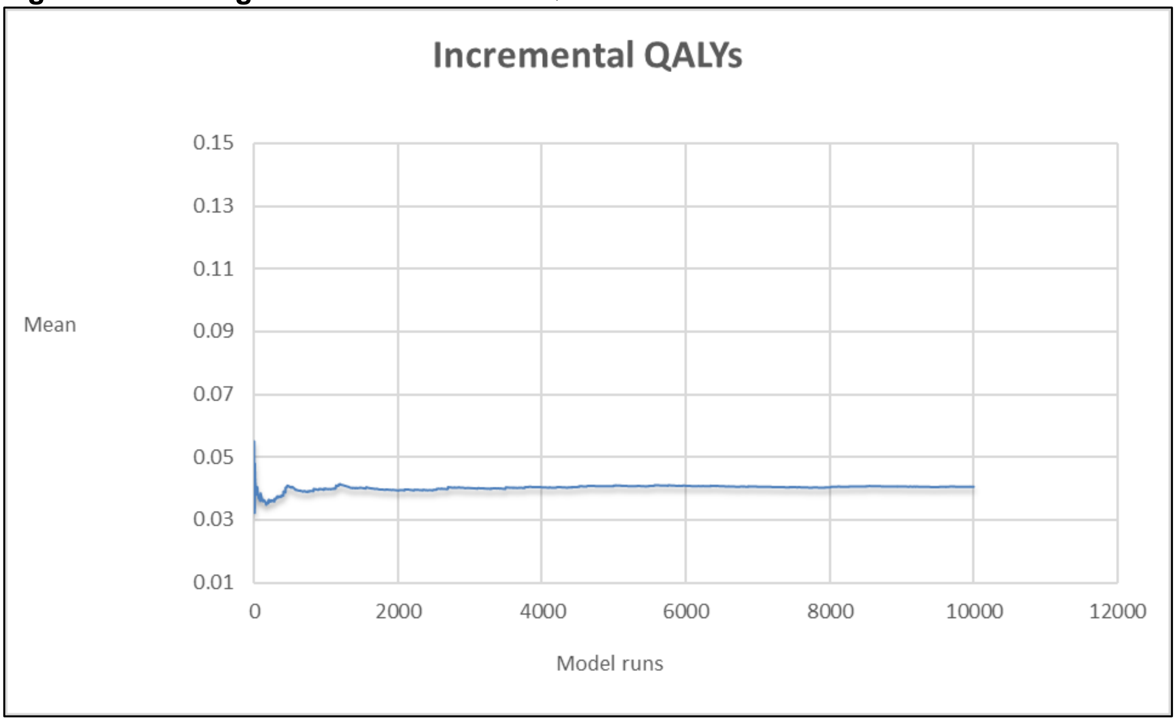
The approach taken aims to give an indication about whether exercise is likely to be cost effective to the NHS based on the currently available evidence. However, if exercise is found to be cost effective, uncertainties will remain due to the heterogeneity in the underlying evidence base. The greatest heterogeneity is in what is meant by exercise, but all of these considerations should be taken into account when interpreting the results of the analysis.

2.2 Uncertainty

A probabilistic model was built to take account of the uncertainty around input parameter point estimates. A probability distribution was defined for each model input parameter. When the model was run, a value for each input was randomly selected simultaneously from its probability distribution; mean costs and mean QALYs were calculated using these values. The model was run repeatedly – 10,000 times for the base case and each sensitivity analysis – and results were summarised in terms of mean costs and QALYs, and the percentage of runs where exercise was the most cost-effective strategy at a threshold of £20,000/£30,000 per QALY gained. Probability distributions were selected to reflect the nature of the data and were parameterised using error estimates from data sources.

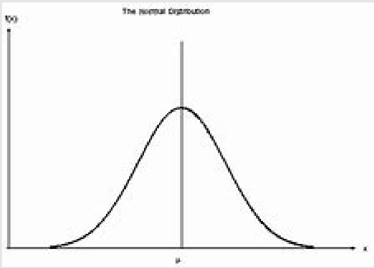
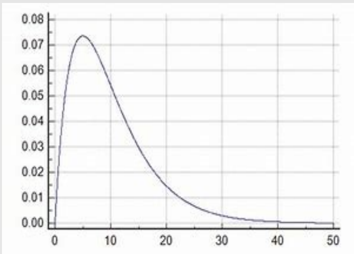
When running the probabilistic analysis, multiple runs are required to take into account random variation in sampling. To ensure the number of model runs were sufficient in the probabilistic analysis, the model was checked for convergence in the incremental costs, QALYs and net monetary benefit at a threshold of £20,000 per QALY gained for exercise versus no exercise. This was done by plotting the number of runs against the mean outcome at that point (see example in Figure 1) for the base-case analysis. Convergence was assessed visually and all had stabilised well before 10,000 runs.

Figure 1: Convergence of incremental QALYs



The way in which distributions are defined reflects the nature of the data. All of the variables that were probabilistic in the model and their distributional parameters are detailed in Table 1 and in the relevant input sections below. Probability distributions in the analysis were parameterised using error estimates from data sources.

Table 1: Description of the type and properties of distributions used in the probabilistic sensitivity analysis

Parameter	Type of distribution	Properties of distribution
Mean difference in EQ-5D between exercise no exercise groups	Normal 	The normal distribution is symmetric. Derived from mean and its standard error.
Intervention costs	Gamma 	Bounded at 0, positively skewed. Derived from mean and its standard error. Alpha and Beta values were calculated as follows: Alpha = $(mean/SE)^2$ Beta = $SE^2/Mean$ Note: SE determined based on the standard deviation across the studies.

The following variables were left deterministic (that is, they were not varied in the probabilistic analysis):

- the cost-effectiveness threshold (which was deemed to be fixed by NICE),
- the resources, including time and cost of staff, required to implement each exercise intervention from each study. Note that intervention costs are modelled probabilistically based on the variation in total costs between studies, but assuming the resource use in each study is fixed,
- the average age,
- the distribution of gender,
- the average life expectancy,
- the regression weights.

In addition, various sensitivity analyses were undertaken to test the robustness of model assumptions. In these, one or more inputs were changed, and the analysis rerun to evaluate the impact on results and whether conclusions on the cost effectiveness of the intervention would change. Details of the sensitivity analyses undertaken can be found in methods section 2.5 Sensitivity analyses.

2.3 Model inputs

Model inputs were based on clinical evidence identified in the systematic review undertaken for the guideline, supplemented by additional data sources as required. Model inputs were validated with clinical members of the guideline committee. More details about sources, calculations and rationale for selection can be found in the sections below.

2.3.1 Clinical studies used in analysis

In economic evaluation, a particular measure of QoL is required known as a utility in order to be able to calculate QALYs. The analysis is therefore based on studies from the clinical review that reported utilities: EQ-5D, or SF-36 that could be mapped to EQ-5D.

Seventeen studies out of the eighty seven included in the clinical review reported data from either the utility instrument EQ-5D-3L (4 studies) or the quality of life instrument SF-36 (13 studies) that can be mapped to EQ-5D-3L. Individual domain data for the SF-36 are required for mapping to EQ-5D-3L. Authors were contacted for those studies that did not report this, however five of the seventeen studies did not provide the subscale data that was needed to map to the EQ-5D[3, 11, 15, 17, 19] and so were not used in this analysis.

This left twelve studies (see Table 2 for references), of which three were three arm trials where both active intervention were exercise. In the three arm trials, the two active exercise arms were combined to create a single pairwise comparison from each of these three studies, as suggested in the Cochrane Handbook [9](see Appendix C: for how these were combined).

A summary of the twelve clinical studies that reported quality of life data that was usable for the economic evaluation are shown in Table 2. The studies were all supervised exercise, and most were group based.

Note some terms being used that should be defined are: Post intervention – outcomes measured at the end of the intervention period (e.g. for a 12 week intervention this would be outcomes measured at 12 weeks); Follow-up – outcomes measured at a future time point beyond when the intervention had ended (e.g. a 12 week intervention following up patients at 24 weeks).

There are other scales that could map to utilities, like mapping from pain scales, which might have allowed for more studies to be used. However pain is only one domain on the EQ-5D, and although this may correlate with QoL, other Qol measures like the SF-36 capture many more components of QoL than just pain. Also, as there was felt to be a sufficient quantity of studies using EQ-5D and QoL measures that could be mapped to EQ-5D, then mapping of pain was not explored further in this analysis.

Table 2: Clinical studies overview

Study	Population	Duration of pain	Level of pain	QoL measure	Intervention (clinical review classification)	Intervention detail (b)	Intervention length (weeks)	Intervention intensity detail	Follow up detail	Number of participants
Sanudo (2011)[26]	Fibromyalgia	NR	SF-36 pain domain =23	SF-36	Aerobic and strength	Combined aerobic and muscle strength training. Assumed group based.	24	2 times a week. 60 min sessions	NA - Post intervention only	42
Tomas-carus (2007)[28] (a)	Fibromyalgia	19-24 yrs depending on group	SF-36 pain domain =21 to 23	SF-36	Aerobic and strength	Pool based aerobics and limb strengthening exercises. Assumed group based.	12	3 times a week. 60 min sessions.	Post intervention, and follow up at 24 weeks	34
Gusi 2008[13]	Fibromyalgia	Approximately 20 years	SF-36 pain domain =20 to 28. Number of tender points = approx. 17	EQ-5D	Aerobic and strength	Pool based aerobics and limb strengthening exercises. Assumed group based.	32	3 times a week. 60 min sessions.	At 12 weeks, then post intervention at 32 weeks	33
Baptista (2012)[4]	Fibromyalgia	NR	VAS = 75 to77mm	SF-36	Mind body	Belly dancing. Group based.	16	2 times a week. 60 min sessions.	Post intervention, and follow up at 32 weeks	80
Von trott (2009)[31]	Neck pain	17-20 yrs depending on group	VAS = 50 to 56 mm	SF-36	Strength and flexibility	Neck strengthening exercises. Group based.	12	2 times a week. 45 min sessions.	Post intervention, and follow up at 24 weeks	117
					Mind body	Qigong. Group based.	12	2 times a week. 45 min sessions.		
Garcia-martinez (2012)[12]	Fibromyalgia	NR	SF-36 pain domain =26 to 34	SF-36	Aerobic strength + flexibility	Individualised protocol including aerobics, walking and stretching.	12	3 times a week. 60 min sessions.	NA - Post intervention only	28

Study	Population	Duration of pain	Level of pain	QoL measure	Intervention (clinical review classification)	Intervention detail (b)	Intervention length (weeks)	Intervention intensity detail	Follow up detail	Number of participants
						Assumed group based.				
Rendant (2011)[25]	Neck pain	Approximately 3 years	VAS = 57 mm	SF-36	Strength + flexibility	Standard program for chronic neck pain. Assumed group based.	24	18 sessions over 6 months (weekly for 3 months then bi-weekly for 3 months). 90 min sessions (assumed to be same length as Qigong arm of trial).	At 12 weeks (halfway through intervention, and post intervention at 24 weeks	122
					Mind body	Qigong. Assumed group based.	24	18 sessions over 6 months (weekly for 3 months then bi-weekly for 3 months). 90 min sessions.		
Lauche (2016)[16]	Neck pain	NR	Unclear pain scale, had to report pain >45mm on VAS to enter trial.	SF-36	Strength, proprioception and flexibility	Neck exercises. Assumed group based.	12	Once a week. 60-75 min sessions.	Post intervention, and follow up at 24 weeks	114
					Mind body	Tai Chi. Assumed group based.	12	Once a week. 75-90 min sessions.		
Gusi (2006)[14] (a)	Fibromyalgia	19-24 yrs depending on group	63 out of 100	EQ-5D-3L; UK value set	Aerobic + strength	Pool based aerobics and limb strengthening exercises. Assumed group based.	12	3 times a week. 60 min sessions.	Post intervention, and follow up at 24 weeks	34
Beasley (2015)[6]	Chronic widespread pain	NR	Categorised into one of 4	EQ-5D-3L; UK value set	Aerobic	Leisure-facility–and gym based	24	Fitness instructor led monthly	Post intervention, and follow up at 36	218

Study	Population	Duration of pain	Level of pain	QoL measure	Intervention (clinical review classification)	Intervention detail (b)	Intervention length (weeks)	Intervention intensity detail	Follow up detail	Number of participants
			chronic pain grades (majority fell into grade 2)			exercise program. Individual based.		appointments, and encouraged to attend the gym at least twice a week.	weeks (9 months) and 120 weeks (30 months).	
Andrade (2019)[1] (a)	Fibromyalgia	NR	VAS = 5.5-5.8	SF-36	Aerobic	Aquatic physical training. Group based.	16	2 times a week. 45 min sessions.	Post intervention, and follow up at 32 weeks	54
Van Eijk-Hustings (2013)[30]	Fibromyalgia	Approximately 7 years	NR	EQ-5D-3L; value set unclear	Aerobic	Gym-based aerobics. Group based.	12	2 times a week. 60 min sessions.	Post intervention, and follow up at 84 weeks (21 months).	95

(a) Participants in these studies were instructed to stop exercising after the intervention ended. This was referred to as 'de-training' to most likely assess the impact of a short term intervention on the follow up outcomes. Outcomes following a de-training period are only included in a sensitivity analysis.
(b) Studies were reported as group based if the study specifically stated the number of people per group. Otherwise group based was assumed from the type of exercise and the way the intervention was described. The Beasley study stated it was an individual based intervention.

2.3.2 Calculating the difference in QALYs

2.3.2.1 EQ-5D and SF-36 data extraction from clinical studies

Most of the studies measured quality of life at more than one time point (not including baseline), generally after the intervention had ended (post-treatment), and later in time (follow-up).

In the clinical review, outcomes from a study were only extracted at the time point closest to 3 months, and the longest time point after 3 months that was closest to 12 months. This meant there were some outcomes in the studies that were not included in the clinical review. For the economic analysis, EQ-5D and SF-36 data was extracted for all time points at which quality of life outcomes were reported in the studies. The different approach taken to the data in the economic analysis was because the EQ-5D was the outcome of interest in the modelling so all the data available was used, and also the committee was interested to understand the effect of exercise over time after the intervention had ended.

Another decision made in the clinical review was to exclude outcomes that were measured after a ‘de-training’ period. This is where some studies told people not to engage in any physical activity after they had undertaken an intervention, and outcomes were measured again following this period of inactivity. From a meta-analysis perspective, outcomes in studies where people were told to stop exercising do not provide information about the intervention as this does not reflect a real-life scenario. These trials were seen as different to studies where follow-up outcomes were based on people either not being given any advice after the intervention, or being encouraged to continue exercise. Although it is known that some people will not continue exercising after an intervention, some might. Even for those who do stop spontaneously, discontinuation might be gradual over time. Therefore, outcomes in the clinical review measured at later follow-ups were intended to see if the interventions affected longer-term engagement in exercise, rather than to see if a short period of exercise itself had any long-term benefits. This latter point could be seen as a different question and wasn’t the question the protocol was designed to answer. This clinical rationale was followed in the base case analysis; outcomes following a de-training period were excluded, but were included in sensitivity analyses. Table 2 highlights in a footnote which studies were the ones where follow-up outcomes were following a de-training period.

Both baseline QoL data from each arm, and follow up outcomes at each time point, as well as confidence intervals, were extracted.

All SF-36 data were reported as mean scores (baseline and follow-up), and three EQ-5D studies reported mean scores (baseline and follow-up) and one reported change from baseline scores so the mean at follow-up was calculated using the baseline and change score.

2.3.2.2 Mapping SF-36 data to EQ-5D

For studies that reported SF-36 data, the mean scores for each of the subscales were extracted for the baseline and any follow-up (post intervention or later follow-up), for both the intervention and control groups.

The standard deviation (SD) or confidence intervals of the SF-36 individual domain means were also extracted. Where only SD’s were reported, the confidence intervals were calculated in Revman software using: the number of participants in the study; the mean; and the SD.

The SF-36 scores and their confidence intervals were mapped onto the EQ-5D-3L (UK tariff) using regression model 4 from Ara &Brazier 2008.[2]

Full details on the data extracted (or calculated) from the studies, can be seen in the appendices A and B.

2.3.2.2.1 *Adjusting mapping for uncertainty in the regression*

Several studies have suggested that there is a problem with underestimation of uncertainties of utilities derived from mapping algorithms.[8],[5] This means that confidence intervals based on the derived utilities are tighter than the confidence intervals of the original actual utilities. This can have implications for utilities then used in cost effectiveness analyses, as uncertainty is being underestimated. The most obvious explanation for the variance underestimation of derived utilities is that there are important unmeasured predictors in most mapping algorithms. This leads to a relatively high degree of unexplained variance of utilities. In OLS based mapping algorithms, this is reflected as a relatively low R squared.

To account for this source of uncertainty in the mapping process, an additional variance component was included in the EQ-5D predictions. A mapping process involves additional sources of uncertainty – the uncertainty in the mapping function regression coefficients and the structure of the mapping model. These additional sources of uncertainty are not accounted for in this analysis.

Chan 2014[7] suggests methods that could be used to estimate the variance of mapped values, by accounting for a low R squared in OLS-based mapping algorithms. Multiple methods are suggested, but some are only possible if patient level data is available. One simple method however that could be used to account for an artificially low variance of utilities because of a low R squared, is to inflate the variance of the derived utilities by a factor of 1/R squared. This estimator helps account for a low R squared, but does not account for the uncertainty of the regression coefficients. This adjustment has also been used in other studies using mapping. [18]

This adjustment factor was applied to the variance of the mapped EQ-5D values for the utilities mapped from the SF-36. See Appendix section B.2 for details of the variance before and after the adjustment was made.

2.3.2.3 EQ-5D (original and mapped) over time by study

Table 3 and Figure 2 summarise the available EQ-5D data (original and mapped, by study).

Some studies measured QoL at a later point in time after the intervention ended. Some of these studies showed a continued improvement in QoL (Andrade,[1] Van Eijk-Hustings,[30] Beasley[6]), whereas other studies showed that QoL gain reduces at follow up possibly capturing the fact that people have reduced their activity over time after the intervention ended. It is difficult to explain why QoL in some studies would continue to improve, given the committee’s opinion that most people tend to discontinue continue exercise. This might be due to small numbers of people in the studies making their findings less likely to reflect the general chronic primary pain population.

There were two studies with longer outcomes (beyond a year) than other studies (Beasley,[6] Van Eijk-Hustings[30]). Both these studies showed that QoL continued to improve in the intervention arms at these follow-up points. The committee were not confident that quality of life continuing to improve after a course of exercise had finished was clinically plausible, especially so long after the interventions ended. For this reason, they decided to exclude these long-term outcomes from the base case, and to include them in a sensitivity analysis.

In the base case, only studies with very long follow-up were excluded as mentioned above. All the remaining studies’ outcomes were pooled regardless of; whether they were during or post intervention; or the direction of the QoL. As previously discussed, there were also three studies that had a de-training period. These outcomes following a de-training period were also excluded from the base case.

It is also important to note that because the pooled QoL values represent exercise treatment effect as the QoL *gain (or loss)* from exercise compared to usual care (taking into account the baselines), then an improvement could have many causes. For example: the usual care group may have had a reduction in QoL, but the exercise group remained stable, or: the exercise group had improved QoL, and the usual care group remained stable, or both groups improved similarly leading to small QoL gains from exercise. The baseline differences and direction of these QoL changes varied between the studies, as can be seen from Figure 2.

It is also important to point out that some studies had large baseline differences in QoL between arms. For example Andrade 2019, and Baptista 2012. Without taking these into account and using only final outcome differences between exercise and control groups, this may be over or underestimating the QoL difference between those who had exercise and those who did not. How baselines were accounted for in the meta-analyses where studies were pooled is discussed in the next section.

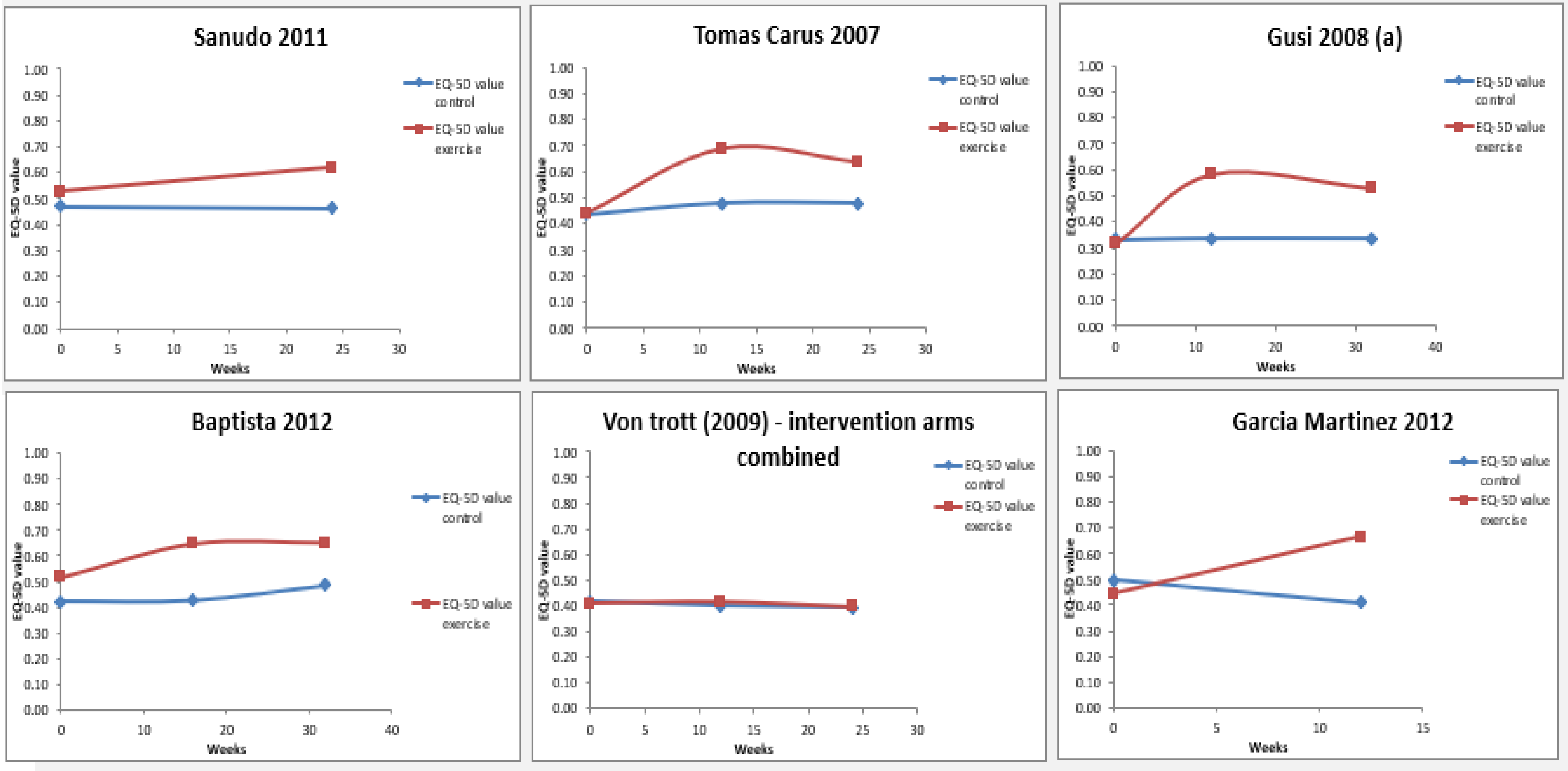
Table 3: EQ-5D (original and mapped) over time by study

Study	Timeframe (weeks) (b)	EQ-5D value control	EQ-5D value exercise
Sanudo (2011) [Intervention length (24 weeks)]	0	0.47	0.53
	24	0.46	0.62
Tomas-carus (2007) [Intervention length (12 weeks)]	0	0.44	0.44
	12	0.48	0.69
	24	0.48	0.64
Gusi (2008) [(a)] [Intervention length (32 weeks)]	0	0.33	0.32
	12	0.33	0.58
	32	0.33	0.53
Baptista (2012) [Intervention length (16 weeks)]	0	0.42	0.52
	16	0.42	0.65
	32	0.49	0.65
Von trott (2009) (c) [Intervention length (12 weeks)]	0	0.42	0.41
	12	0.40	0.41
	24	0.39	0.39
Garcia-martinez (2012) [Intervention length (12 weeks)]	0	0.50	0.44
	12	0.41	0.67
Rendant (2011) (c) [Intervention length (24 weeks)]	0	0.80	0.78
	12	0.79	0.85
	24	0.79	0.85
Lauche (2016) (c) [Intervention length (12 weeks)]	0	0.79	0.77
	12	0.78	0.82
	24	0.78	0.82
Gusi (2006) [(a)] [Intervention length (12 weeks)]	0	0.32	0.29
	12	0.30	0.56
	24	0.30	0.43
Beasley (2015) [(a)] [Intervention length (24 weeks)]	0	0.65	0.69
	24	0.69	0.72
	36	0.65	0.71
	120	0.63	0.71

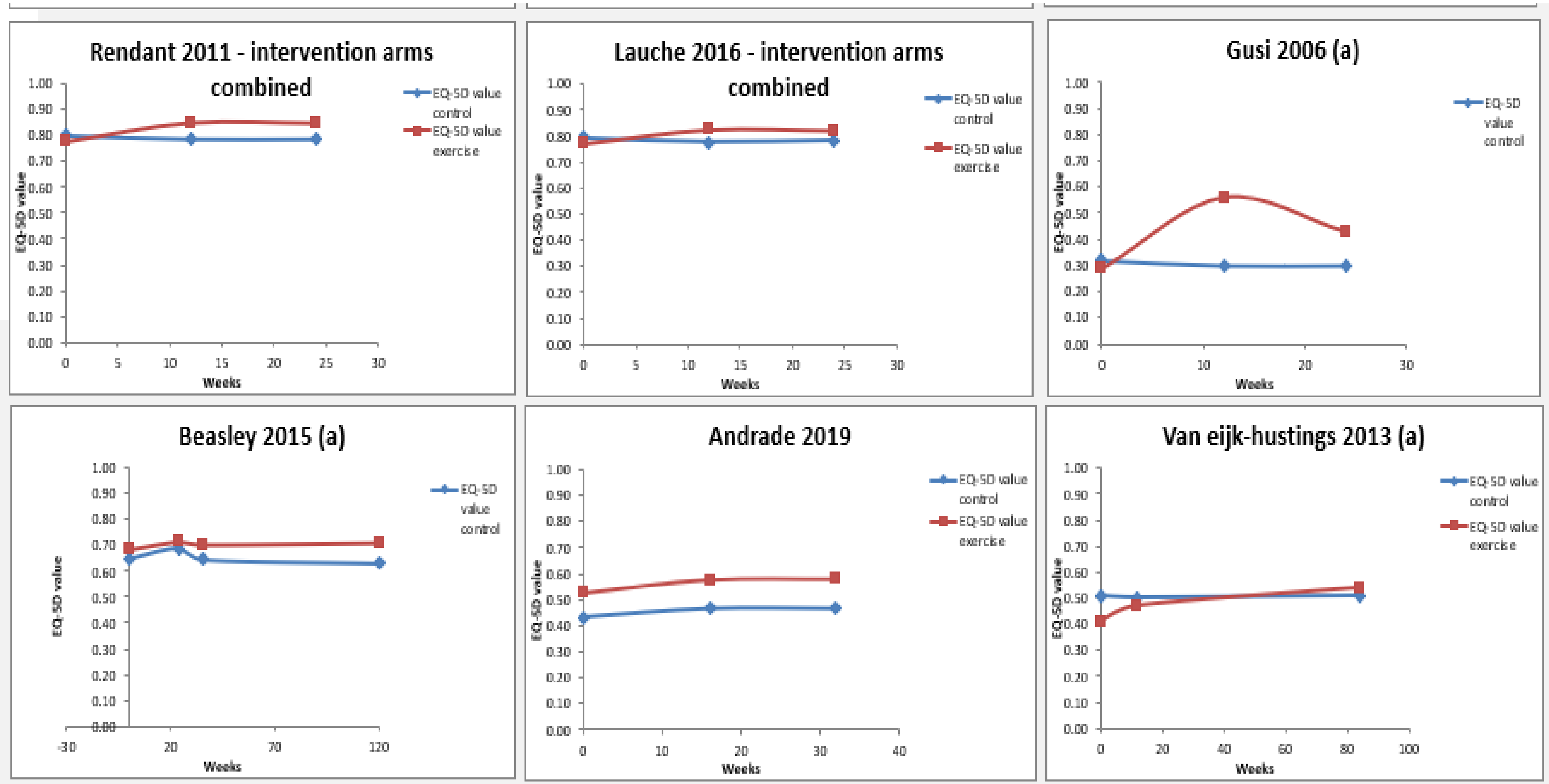
Andrade (2019) [Intervention length (16 weeks)]	0	0.43	0.53
	16	0.47	0.58
	32	0.47	0.58
Van eijk-hustings (2013) [(a)] [Intervention length (12 weeks)]	0	0.51	0.41
	12	0.50	0.47
	84	0.51	0.54

(a) These studies reported EQ-5D data.
(b) Timeframe 0 is the baseline.
(c) These three studies had three arms, but the two exercise arms have been combined in to a single arm following Cochrane methodology.[9] See Appendix C:

1 **Figure 2: EQ-5D (original and mapped) over time by study (b)**



2



(a) These studies report EQ-5D data
(b) Studies with only two dots per line had only a baseline and post intervention measurement. Studies with more than two dots per line usually had a baseline, post intervention, and later follow-up measurement. See Table 2 for more detail on the follow up detail of each trial.

2.3.2.4 Meta-analysing the EQ-5D data

As described in the ‘Approach to modelling’ section, the committee agreed the most informative approach would be to pool all available studies for exercise together in order to analyse the cost effectiveness of exercise versus no exercise. As quality of life benefits may change over time, it was agreed that pooling should be done by time point.

A meta-analysis of QoL values can be undertaken in various ways depending on the data available from the trials. For example, a meta-analysis could use only final (post intervention/follow-up) outcome EQ-5D data at each timepoint (not accounting for baseline differences), or if some studies report change from baseline EQ-5D and some report final outcome EQ-5D then these could also be combined in a meta-analysis based on Cochrane methodology using mean differences (as this mixture of outcomes means not everything would be on the same scale).[9] However the data is meta-analysed, standard deviations of the means are needed to undertake the meta-analysis. As most of the data was mapped from SF-36 to EQ-5D, then the uncertainty around these mapped values was in the form of confidence intervals (as the SF-36 confidence intervals were also mapped). Therefore, standard deviations around the baseline and follow up means were derived using the confidence intervals and number of participants in each arm. More detail can be found below on how the standard deviations around change from baseline scores was identified.

Calculating standard deviations of change scores

Given that there were baseline differences between studies, it was decided that meta-analysing EQ-5D change scores (i.e. change from baseline in the exercise and control groups from each study) would be a more precise way of using the data from the trials. However all the trials, except one, reported baseline and follow-up EQ-5D, not change scores, which meant that although change scores could be calculated by taking the difference between the baseline and follow-up QoL, there is no such simple method to calculate the SD around change scores if it is not reported in the studies.

The Cochrane handbook[9] suggests a method whereby standard deviations for changes from baseline can be imputed. This involves calculating a correlation coefficient from a study that is reported in considerable detail, and then using this coefficient to impute a change from standard deviation in another study. The correlation coefficient describes how similar the baseline and final measurements were across participants. See the equation below.

Equation 1: Correlation coefficient equation

$$\mathrm{Corr}_{\mathrm{E}} = \frac{\mathrm{SD}^2_{\mathrm{E,\ baseline}} + \mathrm{SD}^2_{\mathrm{E,\ final}} - \mathrm{SD}^2_{\mathrm{E,\ change}}}{2 * \mathrm{SD}_{\mathrm{E,\ baseline}} * \mathrm{SD}_{\mathrm{E,\ final}}}$$

Corr = correlation coefficient
E = experimental group (the correlation coefficient needs to be calculated per group)
SD = standard deviation

This information could be calculated from one study (Gusi 2006, an EQ-5D study), which reported QoL as baseline mean (and SD), and follow-up Qol was reported as change from baseline mean (with confidence intervals). Therefore, the confidence intervals could be used to derive a change from baseline SD. In addition, the change scores were used to calculate final values and confidence intervals around the final values, which also allowed calculation of SD's around the final values. Therefore, the three SD elements (per arm) needed in the above equation could be obtained. The correlation coefficients calculated can be seen below in Table 4.

Table 4: Correlation coefficient using Gusi 2006 study

Intervention	Baseline SD	SD - 12 wk FU	SD - 24 wk FU	change from baseline SD (12 wk FU)	change from baseline SD (24 wk FU)	correlation coefficient (12 wk FU)	correlation coefficient (24 wk FU)
Exercise	0.280	0.316	0.368	0.316	0.368	0.444	0.380
control	0.320	0.326	0.316	0.326	0.316	0.491	0.507
						Average: 0.467	Average: 0.444

Abbreviations: SD = standard deviation, FU follow-up, wk = week.

Correlation coefficients lie between –1 and 1. Cochrane methodology[9] states that a simple average across the interventions if the coefficients are similar will provide a reasonable measure of the similarity of baseline and final measurements across all individuals in the study. If a value less than 0.5 is obtained, then there is no value in using change from baseline, and an analysis of final values will be more precise. Although the average from the 12 week follow up is below 0.5 (albeit very close to), which if interpreted strictly, would imply that there is no value in using change from baseline scores in a meta-analysis, there is uncertainty around this coefficient because it was only possible to determine this from one study. Looking back at Table 3 also shows that there are other studies with larger baseline differences between groups than the study the correlation coefficient was calculated from. Therefore, it is possible that the intervention effect does depend on the baseline value, and additionally the sample size from Gusi 2006 was small, which can also affect the reliability of the correlation coefficient.

In summary, only one study was available to compute the correlation coefficient, and although it implied that a change score meta-analysis would not add value beyond a meta-analysis of final values, the heterogeneity of the studies and the fact that this could only be computed from one small study led to the conclusion that a meta-analysis that accounts for baseline differences would be more appropriate. However, in a sensitivity analysis, treatment effects based on a meta-analysis of final QoL values was tested.

The equation showing how standard deviations were imputed using this correlation coefficient is shown below. The correlation coefficient from the 12 week outcomes was used as this was slightly higher and therefore more reflective of a correlation between baseline and final values.

Confidence intervals (around the mean baseline and mean follow up EQ-5D) and the number of participants in the study were used to derive the SD's of baseline and final values needed for the below equation.

Equation 2: Imputing standard deviations using correlation coefficient.

$$\mathrm{SD}_{\mathrm{E,\,change}} = \sqrt{\mathrm{SD}^2_{\mathrm{E,\,baseline}} + \mathrm{SD}^2_{\mathrm{E,\,final}} - (2 * \mathrm{Corr} * \mathrm{SD}_{\mathrm{E,\,baseline}} * \mathrm{SD}_{\mathrm{E,\,final}})}$$

Corr = correlation coefficient
E = experimental group (the correlation coefficient needs to be calculated per group)
SD = standard deviation

A summary of the meta-analysed data informing each timepoint can be seen in Table 5. The full data on the EQ-5D changes from baseline and their SD's from each study can be seen in the appendix.

The treatment effect reported here is the mean difference in changes from baseline QoL, between exercise and no exercise groups.

Table 5: EQ-5D mean difference between exercise and no exercise

	12 weeks	16 weeks	24 weeks	32 weeks	36 weeks (b)	84 weeks (b)	120 weeks (b)
Base case							
Pooled QoL difference	0.08	0.09	0.03	0.11	0.02		
Uncertainty	0.04 to 0.12	-0.04 to 0.21	0.01 to 0.07	-0.02 to 0.25	-0.05 to 0.09		
No. studies informing outcomes	8	2	5	2	1		
Including outcomes following a planned de-training period (a)							
Pooled QoL difference			0.04	0.08			
Uncertainty			0 to 0.07	-0.03 to 0.19			
No. studies informing outcomes			7	3			
Including long term outcomes (a)							
Pooled QoL difference						0.13	0.04
Uncertainty						-0.01 to 0.27	-0.04 to 0.13
No. studies informing outcomes						1	1

(a) Note that these are included in sensitivity analyses.
(b) Where there was only one study, this was still input into Revman software so that the confidence intervals around the mean difference (in change scores from exercise and no exercise) could be obtained.

It is noted that the some of the data points represent a measurement at the end of an intervention and some at later follow-up. In this analysis, all data from a particular time point have been pooled together. The committee agreed this was the best approach because pooling all this data will provide information on the average treatment effect over time from all the exercise programmes, taking into account the potential for discontinuation. This is a more conservative approach towards exercise than using only the post intervention outcomes, as follow up quality of life tended to be lower leading to a downward sloping trend line, whereas including only post intervention outcomes leads to an upward sloping trend line which would lead to a much higher QALY gain.

2.3.2 Using the EQ-5D data in the analysis

In the model, the EQ-5D data from different time points (meta-analysed where possible) were used to estimate QALY gain with exercise by plotting a linear trend line through the data points and calculating QALY gain as the area under the curve. The linear trend line was generated using weighted least squares regression so as to apply a higher weight to the treatment effect from timepoints that had smaller variance. Treatment effect was extrapolated beyond the trial data using the trajectory of the trend line until there was no additional quality of life benefit from exercise. A linear increase in EQ-5D from zero difference at time zero to the point estimated by the trend line at the first trial observation was also assumed. More discussion about the use of a linear trend line, the regression, and extrapolation beyond the data can be found below.

To make treatment effect probabilistic, a normal distribution was used around the mean difference in EQ-5D change scores, as this would not be bounded by zero, and it is possible for there to be a QoL loss from exercise compared to no exercise. The uncertainty around the treatment effect from the time points was varied independently: this means that the slope of the treatment effect line can change. It was considered whether the pooled QoL change at each time point could be correlated, but as not all the points were from the same study, it was decided to let the uncertainty around the pooled QoL from each time point be independent. Therefore this is a limitation in the model.

Use of a linear trend line in the analysis

Fitting a trend line to data allows you to predict the treatment effect for timeframes that go beyond those available, and also given that the points were from different studies and did not follow a tight trend, a trend line gives a smoothed estimate of the treatment effect trend over time which can make it easier to work out the area under the line (i.e. the QALYs).

A linear trend line was fitted to the QoL gain points over time. Different distributions were considered when fitting a trend line to the data, for example, exponential. On a practical level, the exponential distribution does not work with negative values, which were possible in probabilistic analysis in the model. Other properties of the exponential distribution, such as assuming independence between observations, were also not considered entirely appropriate, as this distribution is usually more suited to predicting time to the next event, where the time to the next event is independent of the time to the events that have gone before. This may not be the case in relation to the quality of life from exercise particularly because the interventions are short term, so a person’s quality of life after the intervention stopped could be dependent on whether they were benefitting during the intervention. Additionally, because an exponential distribution never reaches zero, a linear fit was considered more conservative because treatment benefit would reach zero sooner.

Weighted regression methods for generating a trend line

In order to better take account of uncertainty around the pooled treatment effects at each time point, then weighted regression was used to generate a trend line that would attach more importance to the time points where the treatment effect had higher certainty.

Weights that are used in weighted least squares regression typically involve using the reciprocal of the variance.

The standard error around the treatment effect from each timepoint was already calculated for making the treatment effect probabilistic. From this the variance could be calculated. These regression weights are shown below in Table 6.

Table 6: Regression weights

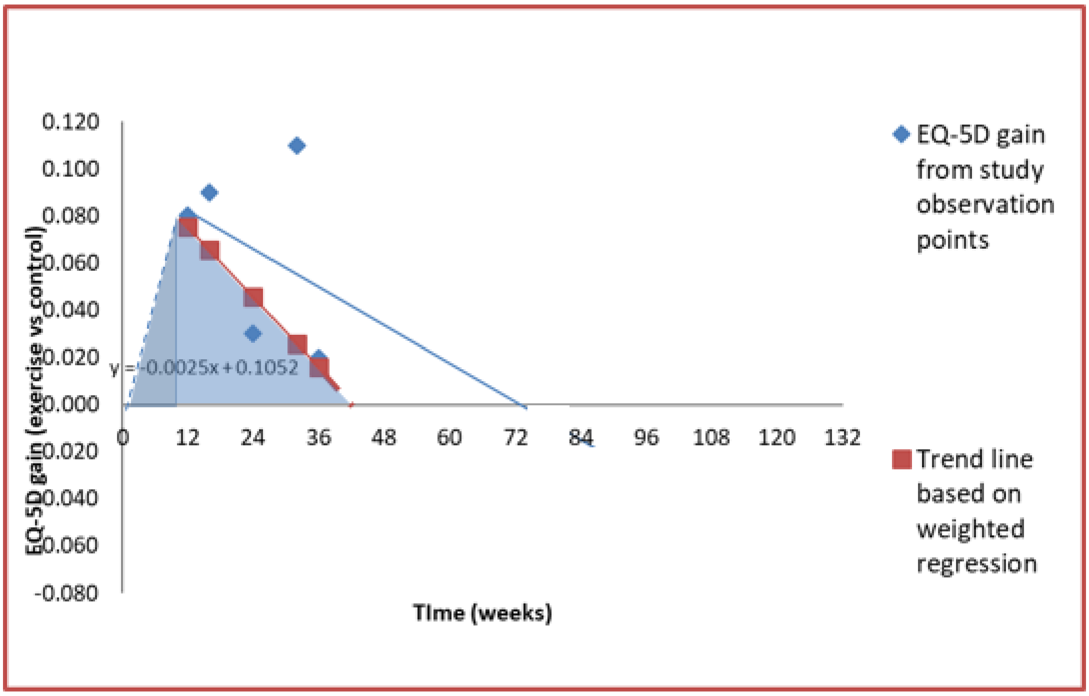
	12 weeks	16 weeks	24 weeks	32 weeks	36 weeks	84 weeks	120 weeks
Base case							
SE	0.02	0.064	0.020	0.069	0.036		
Variance	0.0004	0.004	0.000	0.005	0.001		
Inverse of variance (regression weights)	**2400.9**	**245.9**	**2400.9**	**210.8**	**783.9**		
Including outcomes following a de-training period							
SE			0.018	0.056			
Variance			0.000	0.003			
Inverse of variance (regression weights)			**3135.9**	**317.5**			
Including long term outcomes							
SE						0.071	0.043
Variance						0.0051	0.0019
Inverse of variance (regression weights)						**195.9**	**531.7**

These weights were not varied in the probabilistic analysis.

In the sensitivity analysis using final QoL values in the meta-analysis, the regression weights are different to those in Table 6 because the standard errors around the treatment effects are based on the results of that meta-analysis.

The base case treatment effect over time can be seen in Figure 3. The blue dots represent the treatment effect from each time point from the meta-analysis (with the blue line representing a linear trend of that data. The red dots and corresponding red line show the trend line when the weighted regression is being applied, which is what is being used in the model.

Figure 3: Base case QoL difference over time from exercise



Upward sloping trends in the probabilistic analysis

In the probabilistic analysis, the treatment effect at each timepoint can vary (the probabilistic analysis in this model has 10,000 simulations). It is therefore feasible that the trend line of the treatment effect could be upward sloping in a simulation if treatment effect at later timepoints are higher than treatment effect at shorter timepoints (and also depending on the effect of the regression weightings).

The committee discussed whether an upward sloping trend line would be clinically feasible (i.e. the QoL gain from exercise continuing to improve over time). It was thought possible that some people could continue experiencing improvements in QoL from exercise if they were still receiving the intervention, or even after they were no longer receiving the intervention - if they were still exercising on their own (although this is likely to be in the minority, as generally improvement would plateau at some point).

The uncertainty in the model is large, and upward sloping trend lines over time can occur because of the results of two opposing physiological effects that are being pooled:

- A positive effect on QoL: While the intervention is being undertaken. This is confirmed by looking at the data only from outcomes measured right after the end of the intervention (post intervention outcomes) which showed an upward sloping trend line, implying the better the outcome the longer the intervention period.
- A reduced effect on QoL: After the intervention has ended. This is confirmed by looking at the data that also include outcomes measured later in time after the intervention ended (follow-up data) which showed a downward sloping trend line, implying people discontinue (see Error! Reference source not found.).

Therefore, the slope of the line changing in simulations is an appropriate reflection of the uncertainty in the data, and an upward slope only occurs in a small proportion of simulations, but was monitored to assess the impact on the results by comparing the deterministic and probabilistic results (see results section for discussion on this).

Extrapolating treatment effect

The committee discussed whether they wanted to extrapolate beyond the available data. There is a lack of data on whether people continue to exercise beyond the intervention, but the studies that had follow-up outcomes tend to confirm the committee opinion that QoL from exercise would decrease after an intervention ended, as people would discontinue exercising. This is assumed to already be partly captured in the treatment effect from the available data, as some of the outcome measurements were at follow-up. The committee discussed how to extrapolate beyond this data.

The committee agreed that although they were uncertain about what would happen to QoL beyond the available data, following the slope of the trend line (in the base case) seemed reasonable (see Figure 3), as they thought it likely there would be some continuing benefits, even if they reduced, rather than not assuming anything beyond the trial data, which could be underestimating benefits and the cost effectiveness. An alternative base case was therefore modelled where the time horizon of the model was at the end of the trial data (at 36 weeks).

Note that the treatment effect will be extrapolated only until there was no additional QoL benefit from exercise. This is because the committee assumed that over time people would discontinue and any benefit from exercise would reduce back to the baseline (i.e. no difference in QoL between the exercise and control groups). The committee thought that post-exercise QoL in the exercise group was unlikely to reduce below the control group.

Extrapolating treatment effect in this way does not consider the complexities associated with living with the condition. For example, a continuing downward trajectory may not take into account that people may have interventions in the future, or their condition can fluctuate.

However, the data are intended to reflect a population perspective, rather than an individual perspective. The model also assumes that people only receive one course of the intervention.

The exercise interventions in reality are also intended to teach self-management techniques that people could subsequently be practicing themselves. This is partly captured in the model through the follow-up data that is included in the pooled analysis. Therefore, the pooled data represents the average quality of life in populations in which some people may still be exercising.

Further extrapolation assumptions required in the probabilistic analysis

As there is a large amount of uncertainty around each of the QoL gain data points, this can create large changes in the slope of the trend line in each simulation. Each sample from the distribution around each data point can be very different to the last (and even reflect a QoL loss rather than a gain), and therefore also lead to large changes in the slope of the trend line in each simulation. Various scenarios can therefore occur that needed to be identified in the model to avoid unfeasible results, such as QoL gain (or loss) exceeding the maximum difference between the best and worse states on the EQ-5D scale, or QoL accruing beyond feasible survival. These scenarios and their extrapolation assumptions were discussed with the committee when preparing for the probabilistic analysis, because of the uncertainty in the data.

Different extrapolation assumptions were needed depending on:

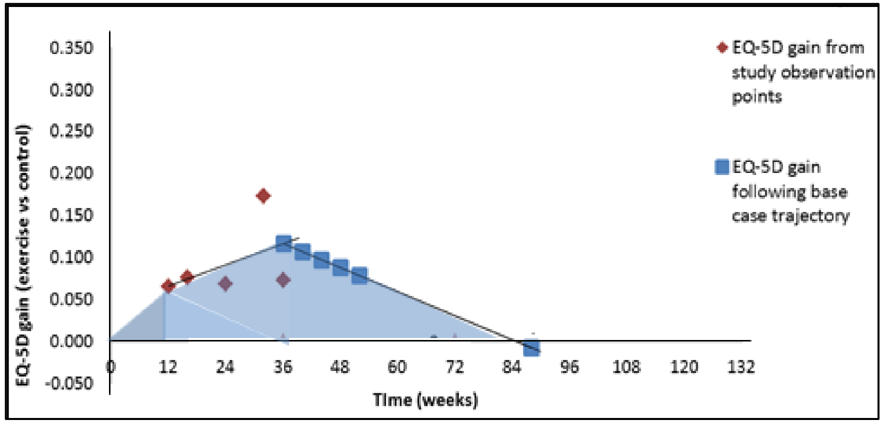
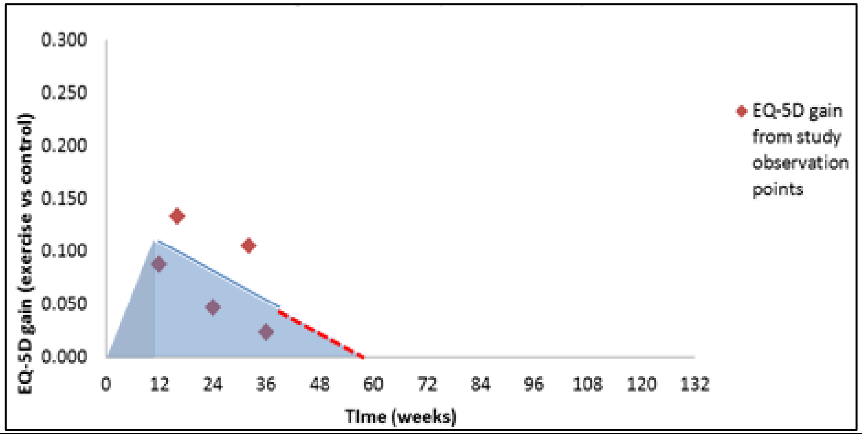
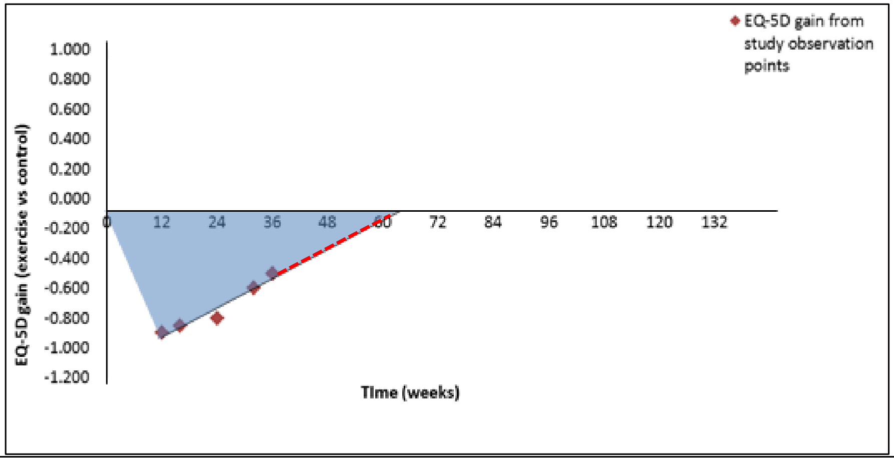
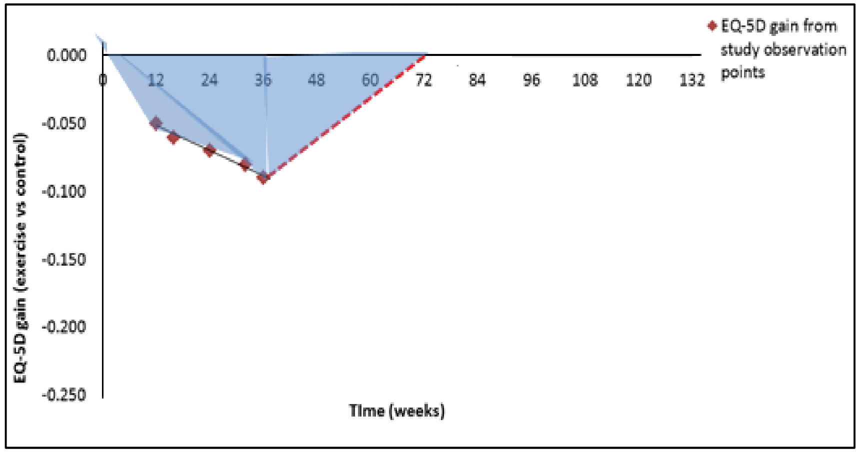
- the slope of the line,
- whether the end of the trend line (based on the final observation point) represented a QoL gain from exercise or a loss.

See Figure 4, and below for more explanation.

1. Where the treatment effect could be upward sloping, with a QoL gain from exercise, it is thought likely that improvements from exercise would not continue increasing indefinitely (and can also only do so to a maximum of 1 for quality of life – an extreme example as we are referring to EQ-5D gain), and although they could initially be increasing, they would at some point plateau. There is little data on how people's behaviour changes following exercise interventions. The committee decided that a conservative estimate would be that when the treatment effect is upward sloping, it should be extrapolated beyond the trial data based on the same slope as the base case treatment effect (hence treatment effect reducing over time) (see Figure 4a).
2. The treatment effect could be upward sloping but the QoL change from exercise could be a loss rather than a gain (i.e. the trend line is in the negative part of the graph). In this case it was assumed the slope of the line continues at the same slope until there is no difference in QoL between exercise and no exercise, as it is unlikely people will continue treatment that was not giving them any benefit. See Figure 4c.
3. The treatment effect could be downward sloping but the QoL change from exercise at each time point was negative (i.e. a loss). In this case, it was assumed that the treatment effect should be sloping up again until there is no treatment benefit (Figure 4d). The point at which there is no treatment benefit was decided as being twice the duration of the last data point (i.e. 72 weeks if the last observation was 36 weeks). This was to create some symmetry not only with the downward sloping part of the trend line, but also with how the converse scenario was dealt with in the positive area of the graph (i.e. when treatment effect was upward sloping then after the last observation point it reverts to a downward sloping line (Figure 4a)).

Note that the scenarios in Figures 4c and 4d only occur in a very small proportion of cases. The committee were concerned that scenarios such as these that result in a QALY loss might be skewing the average result from the probabilistic analysis, so the proportion of times these scenarios were occurring was recorded in the model to check the impact on the overall results.

1 **Figure 4: Extrapolation assumptions**

<table>
<tr><td></td><td>Trend line sloping up</td><td>Trend line sloping down</td></tr>
<tr><td>QoL gain</td><td>Figure 4a: Sloping up with QoL gain

</td><td>Figure 4b: Sloping down with QoL gain (e.g like base case)

</td></tr>
<tr><td>QoL loss</td><td>Figure 4c: Sloping up with QoL loss

</td><td>Figure 4d: Sloping down with QoL loss

</td></tr>
</table>

2 *Note: These are illustrations of the scenarios. Note that 4b is the base case and included in this figure for reference.*

It is important to note that other scenarios could also occur, where the trend line crosses the X axis. In other words: there could be areas of QALY gain along with QALY loss. However, the assumptions remain the same as those in Figure 4, depending on the slope of the trend line and where the trend line ends. For example: if the trend line is downward sloping and starts with a QoL gain from exercise, but ends with a QoL loss from exercise, then the extrapolation assumptions would follow Figure 4d. It was discussed whether probabilistic analysis should allow for QoL losses as well as gains, but again this represents the uncertainty in the data, and also such situations can occur in reality for example exercise making a person's symptoms worse before they make them better.

As mentioned, an alternative base case was undertaken with no extrapolation assumed (i.e. the time horizon was only as long as the last trial observation point (36 weeks in the base case)), as this was the most conservative method of dealing with all the various scenarios that could arise in the simulations.

2.3.14 Life expectancy

In probabilistic analysis where the slope of the trend line was very small, the point at which there is no longer a QoL gain or loss from exercise could be very far into the future, beyond feasible survival. Life expectancy data for each year of age was found from national life tables for England,[27] to cap the duration of treatment benefit so that it cannot go beyond feasible survival. Survival was not assumed to be affected by chronic pain. General population mortality would capture mortality of the average population taking into account that death can be from a number of causes.

The life expectancy by gender was weighted by the distribution of gender from the trial data being used for the economic evaluation.

The age of the average patient was based on taking a weighted average age across the studies informing quality of life data. This was used to determine the total survival time, which was calculated by taking the difference between the age of the average patient at the start, and the weighted average life expectancy. See Table 7 for detail on the population parameters of average age and distribution of gender. A weighted average was used in keeping with how the treatment effect and cost data has been pooled.

Table 7: Population parameters

Parameter description	Point estimate	Source
Population parameters		
Age	53	Weighted average from the RCTs informing treatment effect.
Gender distribution	Men: 12% Women: 88%	The distribution of gender across the RCTs informing treatment effect.

RCT: randomised controlled trial.

2.3.3 Calculating the cost of exercise

As discussed in section 2.2, the committee agreed that the cost of exercise in the model would be based on the pooled resource use from the clinical studies used in the analysis to estimate health benefits. See this section for discussion about pooling.

No other costs were incorporated in the analysis (such as healthcare resource use costs like GP appointments) because there was uncertainty in how other resource use would be impacted from exercise.

2.3.3.1 Resource use

The supervised resource use from each study was identified. This included only the components that could have a cost for the NHS like the time involving staff, or the use of a gym that would require membership. This was either reported as the number of sessions, or the frequency of the intervention per week. The frequency of sessions per week together with the intervention length was used to work out the total number of sessions. This information was combined with the length of sessions to work out the total number of hours of resource use involved in providing the intervention from each study. This is summarised in Table 8. Note that for the studies that had 3 arms (2 exercise interventions and a control group), the 2 intervention arms from those studies are being kept separate for the resource use and cost calculations. This is because there was not necessarily the same resource for the two intervention arms in the same trial, and additionally as it is only the intervention arms that are of interest for resource use (and not the control arm), then there were no issues with double counting the control group participants.

Table 8: Intervention resource use

Study	Intervention classification	Intervention length (weeks)	No. of sessions	frequency (per week)	Session length (a)	Total sessions	Total hours
Mcbeth 2012/ Beasley 2015 (b)	Aerobic	24	6	NR	40	6	4
Lauche 2016	Strength, proprioception and flexibility	12	NR	1	67.5	12	13.5
Lauche 2016	Mind body	12	NR	1	82.5	12	16.5
Von trott 2009	Strength	12	24	2	45	24	18

Von trott 2009	Mind body	12	24	2	45	24	18
Rendant 2011	Mind body	24	18	NR	90	18	27
Rendant 2011	Strength + flexibility	24	18	NR	90	18	27
Baptista 2012	Mind body	16	NR	2	60	32	32
Garcia-martinez 2012	Aerobic, strength and flexibility	12	NR	3	60	36	36
Tomas-carus 2007	Aerobic + strength	12	NR	3	60	36	36
Gusi 2006	Aerobic + strength	12	NR	3	60	36	36
Sanudo 2011	Aerobic + strength	24	NR	2	50	48	40
Tomas-carus 2008/9	Aerobic + strength	32	NR	3	60	96	96
Andrade 2019	Aerobic	16	NR	2	45	32	24
Van Eijk-Hustings 2013	Aerobic	12	NR	2	60	24	24

(a) Where a range of session length was reported in the studies, the midpoint has been used for the session length.
(b) Beasley 2015 was gym based, with participants meeting with a fitness instructor once a month. This is why the number of sessions from this study appear low compared to the other studies in Table 8, as the supervised components were less frequent.

The resource use costed up from the studies is the resource use involved in providing the intervention only for the duration of the trials.

In Beasley 2015, a gym membership was provided for the 6 month trial duration, and this cost has also been included in this analysis to represent accurately the resource use consumed in all the trials (see the next section on costs). In Van Eijk-Hustings, because the intervention also took place in a gym, it has also been assumed that a gym membership would be required. This is not stated in the paper (or in the linked economic evaluation (which was excluded from the review for guideline review due to methodological limitations)[29]), therefore an assumption has been made in keeping with the resource use from Beasley 2015. If this is an overestimate, then it is likely to have little impact on the results as the gym cost is a smaller cost than the staff involved in providing the intervention, but is a conservative assumption.

In order to estimates costs, the level and number of staff involved in providing the interventions in the studies were required. The committee discussed and agreed what would be typically involved in providing the interventions in an NHS setting. All studies except Beasley were assumed to be group based (either because they stated they were, or it could be assumed from the way the intervention was described; Beasley however specifically stated it was individual). An average group was assumed to involve 8 people, and require two staff members, one being more senior (the lower band member therefore acting as an assistant). The committee agreed that in the base case these staff would be a band 6 and a band 4 physiotherapist. Use of other staff bands was also tested in a sensitivity analysis, as well as only having one staff member to teach a class. For Beasley, because the study stated the monthly sessions were with a fitness instructor, the committee thought this would be equivalent to a band 4 staff member (only one member of staff). A summary of the staff costs can be seen in Table 9**.** The assumptions made regarding staffing and total costs per study are shown in Table 10.

The approach of costing up the weighted average of the resource use was used as opposed to determining what exercise looked like in current practice in England and costing that up, because a typical exercise course was difficult to determine due to variability in practice. This would also require the assumption that all exercise is equally effective. There is inadequate information on this, and would also be a strong assumption because even in the studies that were pooled in this analysis; the committee debated whether this was appropriate as they felt that the type of interventions themselves were different and not just different in their duration or intensity (see more discussion on this in approach to modelling section). Therefore, the resource use of the studies used for treatment effect was costed because it was also felt important to keep the relationship between cost and intensity from the clinical studies themselves.

2.3.3.2 Costs

The staff expected to provide exercise interventions would most likely be physiotherapists. The costs of different bands of physiotherapists used in the analysis are presented in Table 9.

Table 9: Physiotherapist costs

Physiotherapist band	Cost per hour	Source
Base case		
6	£64.41	PSSRU 2018[10 a,b]
4	£44.03	PSSRU 2018[10 a,b]
Sensitivity analysis		
5	£51.19	PSSRU 2018[10 a,b]
3	£40.26	Agenda for change pay bands 2018[c], PSSRU 2018 [a,b]

(a) Costs include a ratio of direct to indirect time of 1.37 taken from PSSRU 2018[10], section V.20.
(b) Costs include qualification costs, based on a physiotherapist from PSSRU 2018, section V.18.
(c) Bands below band 4 were not reported in PSSRU 2018. The agenda for change 2018 pay scales were used (https://www.nhsemployers.org/tchandbook/annex-1-to-3/annex-2-pay-bands-and-pay-points-on-the-second-pay-spine-in-england) to work out the salary and on costs of a band 3 staff member, using the midpoint of the ranges within the band. Other assumptions were same as for other bands.

The other cost included was the cost of a monthly gym membership that was needed for two of the studies. The cost per month of local authority commissioned gyms (Better Gyms, a not for profit charitable social enterprise, working in partnership with local authorities) was found online as a proxy for a gym membership. The membership price per month was location specific, so a London area was chosen to be more conservative (£30.95 per month).

The estimated intervention cost by study and the overall weighted average intervention cost used in the analysis can be seen in Table 10. A weighted average cost was calculated by weighting the cost from each study by the number of participants in the intervention arm.

Table 10: Cost of intervention

		Assumptions							
Study	Total hours	No. of staff assumed	Band of staff member 1	Band of staff member 2	No. per group	Supervised cost per pt	Additional resource use	Total cost per patient	N

Mcbeth 2012/Beasley 2015	4	1	4	NA	1	£176	£186 [a]	£362	109
Lauche 2016 - strength	13.5	2	6	4	8	£183	-	£183	37
Lauche 2016 - mind body	16.5	2	6	4	8	£224	-	£224	38
Von trot 2009 - strength	18	2	6	4	8	£244	-	£244	39
Von trott 2009 - mind body	18	2	6	4	8	£244	-	£244	38
Rendant 2011 - mind body	27	2	6	4	8	£366	-	£366	42
Rendant 2011 - strength	27	2	6	4	8	£366	-	£366	39
Baptista 2012	32	2	6	4	8	£434	-	£434	40
Garcia-martinez 2012	36	2	6	4	8	£488	-	£488	14
Tomas-carus 2007	36	2	6	4	8	£488	-	£488	17
Gusi 2006	36	2	6	4	8	£488	-	£488	17
Sanudo 2011	40	2	6	4	8	£542	-	£542	21
Gusi 2008	96	2	6	4	8	£1,301	-	£1,301	17
Andrade 2019	24	2	6	4	8	£325		£325	27
Van Eijk-Hustings 2013	24	2	6	4	8	£325	£93 [a]	£418	47
WEIGHTED AVERAGE COST								**£380**	

(a) Gym membership

Costs were made probabilistic to incorporate uncertainty into the analysis. Although in a sense, there is no uncertainty around the cost within each study because the resource use was fixed, there is variability between studies and so uncertainty in our estimate of average cost to the NHS. The cost of exercise was made probabilistic in the analysis by assuming that each study was a different sample mean. The distribution of the sample mean (i.e. the variability between the studies) is reflected through the standard deviation across all the studies (£264). Standard error reflects the standard deviation of the sample mean distribution, in other words it tells you how close the cost from each study is to the true population mean cost. The standard error (£68) was applied around the cost from each study using the gamma distribution, to generate a probabilistic cost for each study. A weighted average probabilistic cost was then derived weighting by study size in keeping with how the deterministic costs were pooled.

Summary of costs from each study in relation to corresponding treatment effects

As a summary, the costs from each study in relation to the corresponding treatment effects can be seen in Table 11. These are ranked by increasing cost. Note that the treatment effects reported here are the crude mean differences between arms taking into account the baseline mean (difference in difference). Therefore the 2 intervention arms from the 3 arm trials are listed separately here, as the resource use for each arm was considered separately. This includes all data (including the outcomes following de-training, and the longer terms outcomes, that are not included in the base case). Whilst the committee noted the higher cost interventions had higher QoL, they did not feel they could draw conclusions about the correlation between intensity and QoL. There are other variables to take into account such as; the types of exercise are not all the same, and cost also isn't a reflection of intensity in terms of the number of sessions, as the same cost could be reached from a higher number of shorter sessions or fewer longer sessions.

Table 11: Treatment effects and corresponding costs

Study	Time point (weeks)							N (b)	Cost
	12	16	24	32	36	84	120		
	EQ-5D gain								
Lauche (2016) - strength	0.07		0.05					*37*	£183
Lauche (2016) - mind body	0.07		0.06					*38*	£224
Von trott (2009) - strength	0.02		0.01					*39*	£244
Von trott (2009) - mind body	0.03		0.02					*38*	£244
Andrade (2019)		0.02		0.02				*27*	£325
Beasley (2015) (a)			-0.01		0.02		0.04	*109*	£362
Rendant (2011) - strength	0.07		0.08					*39*	£366
Rendant (2011) - mind body	0.09		0.08					*42*	£366
Van Eijk-Hustings (2013)	0.07					0.13		*47*	£418
Baptista (2012)		0.13		0.07				*40*	£434
Garcia-martinez (2012)	0.31							*17*	£488
Tomas-Carus (2007)	0.21		0.15					*17*	£488
Gusi (2006)	0.29		0.16					*17*	£488
Sanudo (2011)			0.10					*21*	£542
Gusi (2008)	0.26			0.21				*17*	£1,301
Meta-analysis estimates (c)	0.08	0.09	0.04	0.08	0.02	**0.13**	**0.04**		**£380**

Colours: Blue = part way through intervention, Green = post intervention, Pink = follow up.
(a) Note that the EQ-5D values taken from Beasley, that are used to work out the EQ-5D gain from the exercise group over time, are the unadjusted EQ-5D values. The study reported adjusted incremental QALYs but not the adjusted EQ-5D values per group, which would be needed here to be pooled with the data from the other studies.
(b) The number of participants are the number in the intervention arm only from each study, as that is the N of interest for the weighted average resource use.
(c) These estimates include all data (including those only included in sensitivity analyses that have longer term outcomes and de-training outcomes. See Table 5 to identify which outcomes these are.

2.4 Computations

The model was constructed in Microsoft Excel 2010, and was evaluated on an individual patient basis. Time dependency was built in by using life expectancy for each year of age and the average age of the populations in the trials informing treatment effect.

A patient starts with zero QoL gain/loss. The maximum time people can derive treatment effect is based on average life expectancy.

The QoL difference from exercise compared to no exercise (taking into account baseline differences) was the treatment effect. This was based on studies in the clinical review that reported EQ-5D utilities or measured QoL through the SF-36 that could be mapped to utilities. QoL differences were based on a meta-analysis of change from baseline scores from the exercise group compared to the no exercise group. The pooled EQ-5D difference at each time point was plotted graphically and a linear trend line fitted to the points based on weighted least squares regression. A linear increase in EQ-5D from zero difference at time zero to the point estimated by the trend line at the first trial observation was also assumed. Treatment effect was extrapolated beyond the trial data using the trajectory of the trend line until there was no additional quality of life benefit from exercise (assumptions about extrapolation could differ in probabilistic analyses depending on the slope of the line and whether the end of the trend line was in the positive or negative part of the graph, see Figure 4).

The area beneath the trend line was considered the area under the curve for calculating QALY gain. Only the incremental QALYs (and costs) are being calculated. QALYs were discounted to reflect time preference (at 3.5%). QALYs during the first year were not discounted. The total discounted QALYs were the sum of the discounted QALYs per year.

Costs were calculated based on average resource use from the trials, and were pooled using a weighted average based on the number of participants in the study. Costs were not discounted because only intervention costs are included and they occur during the first year.

Discounting formula:

$$\text{Discounted total} = \frac{\text{Total}}{(1+r)^n}$$

Where:
r=discount rate per annum
n=time (years)

The incremental cost and QALYs accrued by the patient were used to calculate a cost per QALY for exercise.

2.5 Sensitivity analyses

All the sensitivity analyses were undertaken probabilistically and deterministically, except for the threshold analyses.

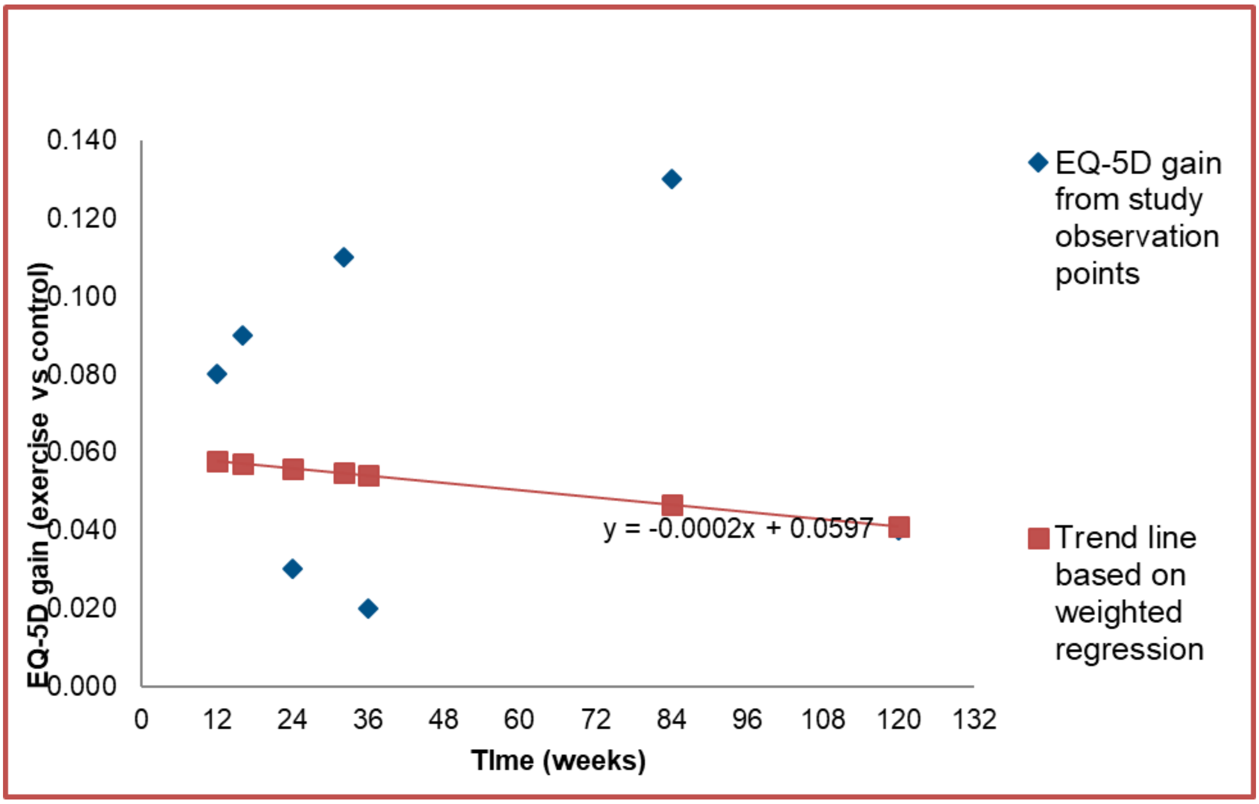
All sensitivity analyses were undertaken for both base cases (extrapolation beyond 36 weeks and truncation at 36 weeks), unless otherwise stated.

2.5.1 SA1: Including long term outcomes (84 week outcome from Van Eijk-Hustings, and 120 week outcome from Beasley)

In the base case analysis, the long term outcomes from Beasley 2015 and Van eijk-hustings (2013) were excluded as the follow up was much longer after the interventions ended compared to other studies (at 120 weeks and 84 week respectively). Also, QoL continued to improve at these follow up points which the committee thought was unlikely to be feasible. In a sensitivity analysis these were included. The additional data points are presented in Figure 5.

The weights used in the regression have the highest weights for the 12, 24 and 32 week timepoints (as seen in Table 6), so the weighted regression trend line is still downward sloping. As the slope is quite flat, then this will lead to an area under the curve that will generate higher QALYs than the base case, because the number of weeks at which the trend line reaches the x axis will be much further into the future.

Figure 5: QoL gain when including longer term outcomes.



For the base case where treatment effect is not extrapolated, including the longer term follow-up data means the maximum time horizon was 120 weeks.

2.5.2 SA2: Including outcomes following a planned de-training period

These outcomes were not included in the base case in order to match the way the clinical review had treated these outcomes, because they were not seen as providing information about the real life scenario. They are included in sensitivity analysis to see if they have any impact on conclusions. The additional data points and revised pooled EQ-5D difference estimates are presented in Table 5

.

2.5.3 SA3: Including outcomes following a planned de-training period AND long term outcomes

This sensitivity analysis is SA1 and SA2 combined. For the lifetime base case, this results in a treatment effect that looks similar to that in Figure 5.

For the base case where treatment effect is not extrapolated, including the longer term follow up data means the maximum time horizon was 120 weeks.

2.5.4 SA4: Using final QoL outcomes in a meta-analysis instead of change from baseline QoL

This sensitivity analysis used final QoL values in the meta-analysis as opposed to change from baseline QoL values. This was to test whether this made any difference to the results of the model, given that the correlation coefficient calculated implied that there was unlikely to be similarity between the baseline and final measurements across participants.

The results of the meta-analysis from final values and the regression weights used based on the uncertainty around each timepoint can be seen in Table 12 and Table 13.

Table 12: EQ-5D mean difference between exercise and no exercise, using final values

	12 weeks	16 weeks	24 weeks	32 weeks	36 weeks (b)	84 weeks (b)	120 weeks (b)
Base case							
Pooled QoL difference	**0.05**	**0.18**	**0.04**	**0.18**	**0.06**	**NA**	**NA**
Uncertainty	0.01 to 0.09	0.05 to 0.31	0 to 0.07	0.05 to 0.31	-0.01 to 0.13		
No. studies informing outcomes	8	2	5	2	1		

(a) Where there was only one study, this was still input into Revman software so that the confidence intervals around the mean difference could be obtained.

Table 13: Regression weights, based on final values

	12 weeks	16 weeks	24 weeks	32 weeks	36 weeks	84 weeks	120 weeks
Base case							
SE	0.020	0.066	0.018	0.066	0.036	NA	NA
Variance	0.000	0.004	0.000	0.004	0.001		
Inverse of variance (regression weights)	**2400.9**	**227.3**	**3135.9**	**227.3**	**783.9**		

2.5.5 SA5: Assuming less staff required

Resource use was varied deterministically by using one staff member instead of two for group interventions, as the committee discussed how this was a possibility in practice. This will lead to a lower cost of the intervention.

2.5.6 SA6: Assuming lower staff bands

The bands of staff involved in providing an intervention were also varied, from a band 6 and 4, to lower bands of bands 5 and 3 for the group interventions, as the most conservative bands were used in the base case. This will lead to a lower cost of the intervention.

2.5.7 SA7: Discounting outcomes at 1.5% (only relevant for extrapolated base case)

QALYs beyond one year were discounted at a rate of 3.5% in the base case, based on the NICE reference case. This is lowered to 1.5% in this sensitivity analysis, as recommended in the NICE guidelines manual.[21]

2.5.8 Threshold analyses

Threshold analyses were undertaken on both what the QALY and cost would need to be, to make the intervention cost effective at a threshold of £20,000 per QALY gained. This was done for both base cases.

2.6 Model validation

The model was developed in consultation with the committee; model structure, inputs and results were presented to and discussed with the committee for clinical validation and interpretation.

The model was systematically checked by the health economist undertaking the analysis; this included inputting null and extreme values and checking that results were plausible given inputs. The model was peer reviewed by a second experienced health economist from the NGC; this included systematic checking of many of the model calculations.

The model was also peer reviewed by a health economist at NICE and an executable version of the model with full technical report was made available to registered stakeholders for review at guideline consultation.

2.7 Estimation of cost effectiveness

The widely used cost-effectiveness metric is the incremental cost-effectiveness ratio (ICER). This is calculated by dividing the difference in costs associated with 2 alternatives by the difference in QALYs. The decision rule then applied is that if the ICER falls below a given cost per QALY threshold the result is considered to be cost effective. If both costs are lower and QALYs are higher the option is said to dominate and an ICER is not calculated.

$$ICER = \frac{Costs(B) - Costs(A)}{QALYs(B) - QALYs \rightleftarrows (A)}$$ Where: Costs(A) = total costs for option A; QALYs(A) = total QALYs for option A	Cost effective if: • ICER < Threshold

2.8 Interpreting results

NICE's report 'Social value judgements: principles for the development of NICE guidance'[24] sets out the principles that committees should consider when judging whether an intervention offers good value for money. In general, an intervention was considered to be cost effective if either of the following criteria applied (given that the estimate was considered plausible):

- The intervention dominated other relevant strategies (that is, it was both less costly in terms of resource use and more clinically effective compared with all the other relevant alternative strategies), or
- The intervention costs less than £20,000 per quality-adjusted life-year (QALY) gained compared with the next best strategy.

Although all the data included in the economic evaluation has been pooled for this analysis, it is important to remember the data is very heterogeneous as they are different interventions. The results need to be interpreted with caution, as the analysis is pooling interventions of different costs, and also different effects from different time points in different study populations. It is likely this analysis could only inform a broad recommendation.

3 Results

3.1 Base case

The probabilistic base case results are presented in the Table 14 and graphically in Figure 6 and Figure 7. Results are presented for both base cases: the extrapolated lifetime analysis and the analysis with a shorter time horizon where treatment effect is not extrapolated.

Exercise is associated with higher costs and higher QALYs. The incremental cost effectiveness ratio is £9,121 per QALY gained from the probabilistic lifetime analysis, and £12,327 when deterministic. When treatment effect was not extrapolated, the ICER was £12,683 in the probabilistic analysis, and £12,739 when deterministic. Both base cases show that the ICER is below the NICE threshold of £20,000, and therefore exercise would be considered cost effective. The probability of exercise being cost effective is also high.

Table 14: Base case results (discounted)

Base case	Analysis	Incremental cost	Incremental QALYs	Cost per QALY gained	Probability cost effective at £20k
Lifetime	Probabilistic	£380	0.04	£9,121	86%
	Deterministic	£380	0.031	£12,327	NA
No extrapolation beyond last trial observation (36 weeks)	Probabilistic	£380	0.03	£12,683	93%
	Deterministic	£380	0.030	£12,739	NA

Abbreviations: QALYs: quality adjusted life years, £20k: £20,000.

There were some differences in the incremental QALY gain estimates with the probabilistic and deterministic analyses, but this did not impact conclusions. The reasons for differences are discussed below.

Figure 6 and Figure 7 show the cost effectiveness plane showing the 10,000 simulations from the base case probabilistic analysis. As can be seen most of the results are in the top right quadrant where the intervention is both more costly but more effective. The mean result is represented by the black X. Note that there is much less variation around the QALYs in Figure 7 because this is short time horizon only until the end of the trial data, whereas in the lifetime analysis where treatment effect is extrapolated (Figure 6), this leads to much more skewness in the QALYs, mostly because of the extrapolation leading to some scenarios with benefit occurring for a long time. The skewed QALYs are leading to different deterministic and probabilistic results in the lifetime analysis, and this is discussed more in the next section.

Figure 6: Base case results (lifetime): cost effectiveness plane

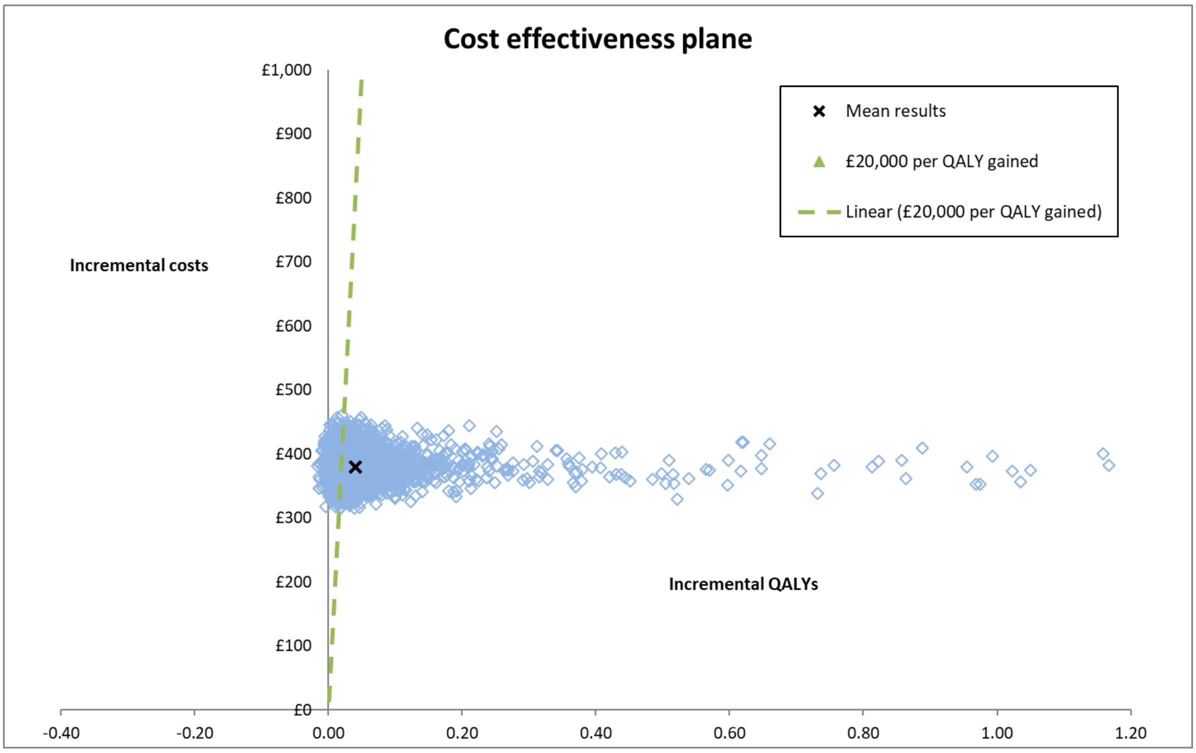
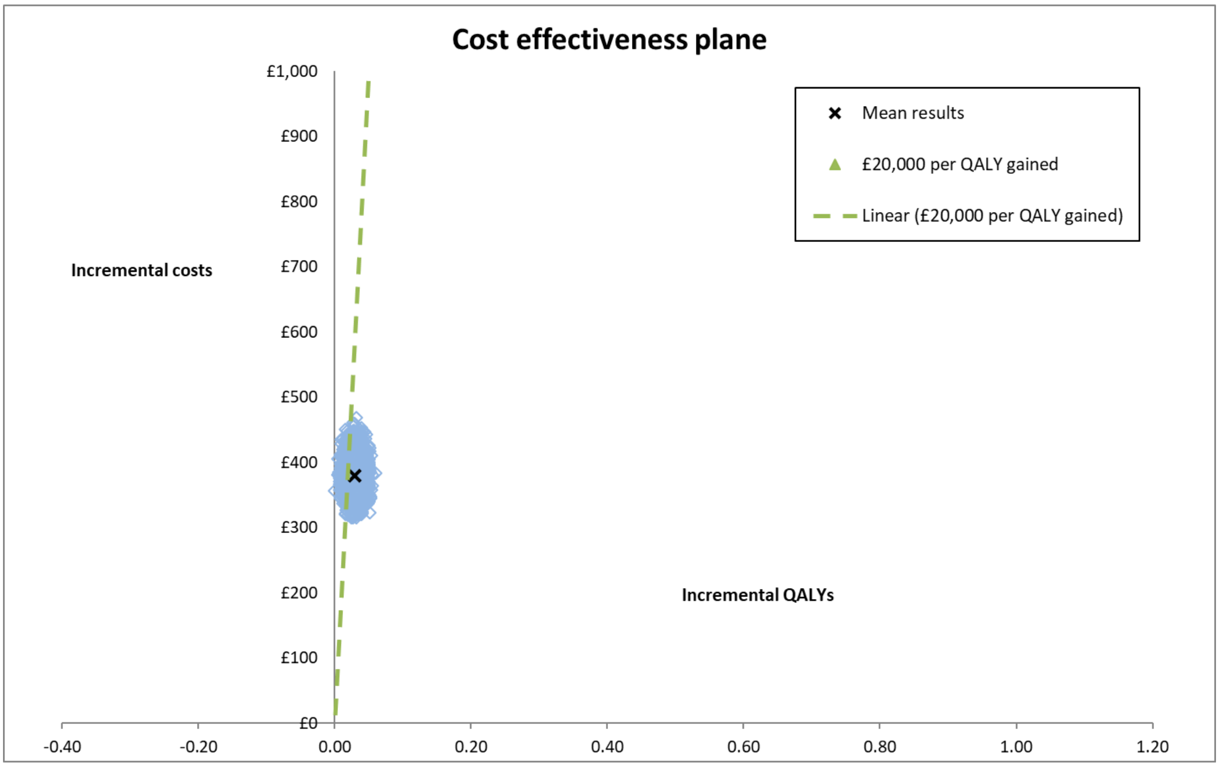


Figure 7: Base case results (no extrapolation): cost effectiveness plane



3.1.1 Differences between deterministic and probabilistic results

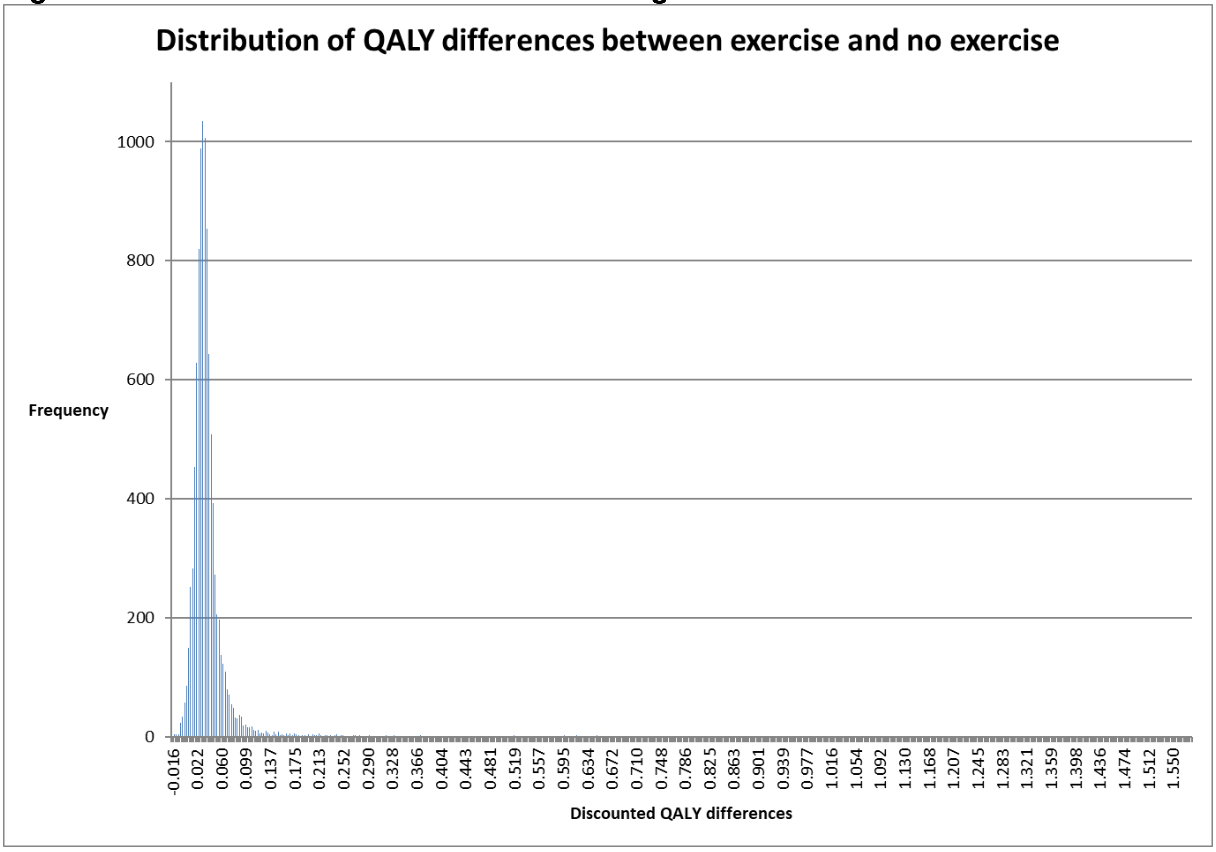
The mean costs and QALYs from the probabilistic analysis are usually considered the best estimate for use in decision making. Deterministic and probabilistic results are often very similar (as the mean of the simulated inputs should always revert to the mean (i.e. the point estimate)). However, this is not always the case, a common example being if models are non-linear. The deterministic analysis (using the input point estimates and not the uncertainty around them) is also calculated and it is routine to consider if these are similar, and if not why not, as it may be the case that differences are due to programming errors in the model. As can be seen above, the incremental QALY estimates in this analysis for the lifetime horizon are somewhat different in the deterministic and probabilistic analysis. This was investigated thoroughly and is considered to be a reflection of the modelling methods used to estimate QALY gain rather than an error. This is discussed further below.

The reason for these differences were because of the extrapolation assumptions, coupled with a skewed distribution of QALY gains in the probabilistic analysis. The most frequent scenario is a downward sloping trend of QALY gain from exercise, but where there are some simulations with quite flat slopes, this leads to a large QALY gain because of the extrapolation assumptions exacerbating the gain, and the point at which there is no longer a difference in treatment effect from exercise being far into the future.

A skewed distribution can be confirmed by viewing the distribution of the QALY changes by plotting the QALY changes from exercise from the base case simulations (10,000 simulations) against their frequency (Figure 8). This confirms there is a skewed distribution with a longer right tail, and therefore even a few simulations with very large QALY gains could be skewing the probabilistic mean.

The deterministic result for the no extrapolation base case is very similar to the probabilistic result (see Table 14), thereby confirming the explanation that the extrapolation of treatment effect can lead to very large QALY gains and a skewed distribution.

Figure 8: Distribution of base case QALY changes from exercise



Some further information that can contribute to what is happening in the probabilistic analysis can be seen in Table 15, where it is recorded how often different scenarios are occurring. Some are occurring very infrequently or not at all, as expected, such as where the trend line is fully in the negative area (i.e. QALY losses), so these are not leading to treatment effect being skewed downward which the committee were concerned about.

Table 15: Occurrence of treatment effect scenarios in lifetime probabilistic analysis

Scenario	Percentage of simulations occurring	Total
Slope direction		
Sloping down	94.51%	100%
Sloping up	5.49%	
Specific scenarios		
Sloping down		
Trend line is fully in positive area	67.56%	100%
Trend line crosses the X axis	26.95%	
Trend line is fully in negative area	0.00%	
Sloping up		
Trend line is fully in positive area	5.49%	
Trend line crosses the X axis	0.00%	
Trend line is fully in negative area	0.00%	

Overall, although it can be explained why the probabilistic and deterministic results are different (due to the uncertainties around the data and how the trend line is behaving in simulations, as well as the extrapolation exacerbating the QALYs), the results are still well below the NICE threshold of £20,000 per QALY gained, and are therefore both in agreement that exercise is likely to be cost effective.

3.2 Sensitivity analyses

The results of the sensitivity analyses are presented in Table 16 and Table 17. These are presented separately for the two base cases. Exercise remained cost effective in all sensitivity analyses. The deterministic results are also reported for each base case in Table 17 because as discussed above, these can differ to the probabilistic results.

1 **Table 16: Sensitivity analysis results - probabilistic**

Analysis	Base case 1: Lifetime analysis				Base case 2: No extrapolation of treatment effect analysis			
	Incremental cost	Incremental QALY	ICER (Cost per QALY gained)	Probability cost effective at £20k	Incremental cost	Incremental QALY	ICER (Cost per QALY gained)	Probability cost effective at £20k
Basecase results	£380	0.04	£9,121	86%	£380	0.03	£12,683	93%
Including long term outcomes								
SA1: Including long term outcomes (84 week outcome from Van Eijk-Hustings, and 120 week outcome from Beasley)	£380	0.17	£1,897	95%	£380	0.11	£3,488	99%
Including outcomes following a de-training period								
SA2: Including outcomes following a de-training period	£380	0.04	£8,326	92%	£380	0.03	£12,060	96%
SA3: Including outcomes following a de-training period AND long term outcomes	£380	0.18	£1,874	96%	£379	0.11	£3,404	99%
Using final EQ-5D values meta-analysis								
SA4: Final outcomes EQ-5D meta-analysis	£380	0.08	£4,316	99%	£380	0.03	£11,890	97%
Changing staff bands and numbers								
SA5: Assuming less staff required	£258	0.04	£6,221	94%	£258	0.03	£8,676	99%
SA6: Assuming lower bands of staff	£333	0.04	£7,904	90%	£333	0.03	£11,205	96%
Discount rate								
SA7: Discount rate at 1.5%	£380	0.04	£8,687	86%	NA	NA	NA	NA
Threshold analyses								
Cost at which exercise has an ICER of £20,000 per QALY gained	£789	NA	NA	NA	£599	NA	NA	NA

Analysis	Base case 1: Lifetime analysis				Base case 2: No extrapolation of treatment effect analysis			
	Incremental cost	Incremental QALY	ICER (Cost per QALY gained)	Probability cost effective at £20k	Incremental cost	Incremental QALY	ICER (Cost per QALY gained)	Probability cost effective at £20k
QALY gain which exercise has an ICER of £20,000 per QALY gained	NA	0.019	NA	NA	NA	0.019	NA	NA

Table 17: Sensitivity analysis results - deterministic

Analysis	Base case 1: Lifetime analysis			Base case 2: No extrapolation of treatment effect analysis		
	Incremental cost	Incremental QALY	ICER (Cost per QALY gained)	Incremental cost	Incremental QALY	ICER (Cost per QALY gained)
Basecase results	£380	0.031	£12,327	£380	0.030	£12,739
Including long term outcomes						
SA1: Including long term outcomes (84 week outcome from Van Eijk-Hustings, and 120 week outcome from Beasley)	£380	0.20	£1,911	£380	0.11	£3,558
Including outcomes following a de-training period						
SA2: Including outcomes following a de-training period	£380	0.03	£11,461	£380	0.03	£12,078
SA3: Including outcomes following a de-training period AND long term outcomes	£380	0.19	£1,968	£380	0.11	£3,509
Using final EQ-5D values meta-analysis						
SA4: Final outcomes EQ-5D meta-analysis	£380	0.05	£7,324	£380	0.03	£11,870
Changing staff bands and numbers						

Analysis	Base case 1: Lifetime analysis			Base case 2: No extrapolation of treatment effect analysis		
	Incremental cost	Incremental QALY	ICER (Cost per QALY gained)	Incremental cost	Incremental QALY	ICER (Cost per QALY gained)
SA5: Assuming less staff required	£258	0.03	£8,387	£258	0.03	£8,667
SA6: Assuming lower bands of staff	£333	0.03	£10,806	£333	0.03	£11,168
Discount rate						
SA7: Discount rate at 1.5%	£380	0.03	£12,327	NA	NA	NA
Threshold analyses						
Cost at which exercise has an ICER of £20,000 per QALY gained	£616	NA	NA	£596	NA	NA
QALY gain which exercise has an ICER of £20,000 per QALY gained	NA	0.019	NA	NA	0.019	NA

For all the sensitivity analyses, for both base cases, exercise remains cost effective with an incremental cost effectiveness ratio below £20,000 per QALY gained. When including the longer term quality of life data points, this leads to more QALYs because people are getting treatment effect for longer.

Including outcomes following a de-training period makes little difference to the results.

Alternative assumptions for resource use have generally made the ICER lower because it has lowered the cost of the intervention.

Threshold analyses showed that, other things being equal, the cost of the intervention needs to be below £616 (£596 in no extrapolation base case) to make the intervention cost effective. Note the values from the deterministic sensitivity analyses have been used here as they are more conservative. Keeping the cost the same as the base case, the QALY gain would have to be at least 0.02 (similar in both base cases because the cost is the same) for exercise to be cost effective.

4 Discussion

4.1 Summary of results

Both base cases (the extrapolated lifetime analysis, and the shorter time horizon analysis where treatment effect was not extrapolated) showed that the addition of exercise to usual care is cost effective with a probabilistic ICER of £9,121 and £12,683 respectively, and deterministic ICERs of £12,327 and £12,739 respectively. This conclusion was robust in sensitivity analyses such as including longer follow up data, and using a meta-analysis for QoL values based on final values rather than change from baseline scores.

4.2 Limitations and interpretation

As highlighted in the methods section, this analysis aimed to assess whether exercise is likely to be cost effective for people with chronic pain. However, there are a number of limitations that should be taken into account when interpreting this analysis.

The analysis only used 12 studies from the clinical review to inform treatment effect as only those studies reported quality of life data. There were actually over 87 studies included in the clinical review. It has been investigated whether the studies included in this analysis are representative of the studies included in the wider clinical review, by reviewing the forest plots to check if the studies in this analysis are outliers. This did not appear to be the case, however some studies were not pooled with other studies in forest plots to allow this eye-balling of the data. Reasons for this include: Not many studies reported QoL so there were no other studies using the same outcome scales to pool data with (as the clinical review also compared different intervention types separately), or some studies used in this analysis did not report other outcomes that could be pooled with other studies. Therefore, it is not possible to be 100% certain that the studies included in this analysis are representative of all the studies in the clinical review, but the committee believed the populations and interventions in the studies included in this analysis were broadly generalisable.

The analysis pooled data across clinical studies that had different interventions of different intensities. This is likely to affect costs and also treatment effect, although there is not necessarily an association between the two. Therefore, there is uncertainty around whether the costs that have been pooled appropriately correspond to, or are leading to, the pooled treatment effect. This is because it is unclear what it is about exercise that causes a benefit. The clinical review did not look to identify a relationship between treatment intensity and treatment effect. Therefore, the committee decided it would not be appropriate to explore this relationship de novo, in an economic analysis without supporting evidence from the clinical review. The model results therefore need to be interpreted bearing in mind that the data has been pooled, and can only be treated as a piece of information alongside the committee's interpretation of the clinical evidence as a whole.

Studies were identified that measured outcomes using the EQ-5D, or QoL measures that could be mapped to the EQ-5D like the SF-36. Mapping is considered a second best alternative to using directly measured utilities. However, to account for uncertainty in the mapping regression, an adjustment method was used to adjust the variance of the mapped values.

Pooling the data included studies that were of different time periods. Some had follow-up a long time after the intervention had ended. The committee were not confident that quality of life continuing to improve from a course of exercise would be clinically plausible, especially so long after the interventions ended. For this reason, they decided to exclude these long-term outcomes from the base case, and to include them in a sensitivity analysis.

Data was pooled in a meta-analysis where different studies reported outcomes at the same time point. Although there are benefits to pooling data together to reduce uncertainty, there is a large amount of heterogeneity as the studies are all very different. The model tried to overcome some of this uncertainty by using weighted regression to generate a trend line based on QoL over time that better represented data points that were more certain. The methods of the studies also differed with some specifically trying to assess the programmes' short term impact on long term outcomes by having a 'de-training' period where people were instructed to stop exercising at the end of the intervention. These outcomes following a de-training period were also excluded from the base case but included in a sensitivity analysis.

The linear trend line representing treatment effect over time is a simplification of how people's quality of life (on average on a population level) would fluctuate in reality. This is because the data is not all from the same study and therefore not telling you about the actual pattern on QoL over time. However, data was pooled to reduce uncertainty.

Modelling the effects of the exercise intervention over the remainder of participant's whole life required extrapolation beyond the trial data. The linear extrapolation is a simplification, as for example people may have other interventions in the future that have not been accounted for here, such as attending a second exercise intervention. However, this would have required assumptions and there was no information on this. Additionally, the extrapolation does not take into account the complexities associated with living with the condition such as reacting to an exercise intervention that increases pain, resulting in more sedentary behaviour, which may mean the analysis has in fact overestimated the extrapolated QoL. However, the committee agreed a reasonable assumption was to extrapolate the trend line following the same trajectory of the base case. The alternative base case also tested not extrapolating the trend line to be conservative. It is also important to note that the data reflected here is from a population level, and is also looking at only one course of the intervention.

Adherence might also be different in reality to what takes place in trials. The quality of life gain taken from the studies could also be an overestimate because it is likely that people who respond to follow up questionnaires or that have not dropped out of a trial are those who are more engaged with the intervention. Additionally, it is uncertain what was happening after the intervention and whether people were continuing the intervention, or perhaps their quality of life improvement could be coming from other causes such as social engagement rather than an effect of the exercise specifically.

Given that it was not possible to access the adjusted EQ-5D values from the Beasley study, it is uncertain what impact this would have had on the results. The paper only reports QALYs from the adjusted data of exercise versus usual care using the 30 month outcomes. However, to test the differences in results: the QALY gain calculated in the model only for the Beasley study, using the unadjusted data in the paper, is similar to the published adjusted QALY in the imputed data analysis, but much higher than the published adjusted QALY in the complete case data analysis (by a ratio of 2.5). Using this ratio to reduce the QALY from SA1 (no extrapolation analysis), led to an ICER of £8,902. Therefore the cost-effectiveness conclusion has not been altered by this lower QALY from the study. No other costs have been accounted for in the analysis except for intervention costs. Very little data on whether exercise influences the use of other resources was found from the clinical review, and the data were conflicting. The committee's opinion was that exercise anecdotally reduces other healthcare resource use. Therefore, these were not included in the analysis. We have also assumed no costs associated with the intervention beyond the intervention length in the trials. Ongoing costs (e.g. gym membership) might also imply an association with ongoing benefits, which we would not have been able to capture from the available data, and modelling this would have required more speculative assumptions.

Overall, this analysis has pooled a subset of data from the clinical review that reported quality of life, to estimate the potential cost effectiveness of supervised exercise in general, not being specific to a particular type of exercise. However it is important to consider the differences between the studies, and how few studies were used compared to the review as a whole, when interpreting this analysis.

4.3 Generalisability to other populations or settings

The populations reflected in the trials used for treatment effect in this analysis are mostly people with fibromyalgia, and some people with chronic neck pain. The committee agreed that these populations are likely to be generalisable to the wider chronic primary pain population.

4.4 Comparisons with published studies

One UK published economic evaluation in this area showed that exercise was not cost effective in the complete case analysis, but was in the imputed analysis. That was a gym based exercise program with limited supervision.[6] The intervention resource use (based on what the trial was designed to deliver rather than what people in the trial actually used) and QoL from this trial were used in the guideline economic analysis. The QALYs from the complete case analysis were lower than those found by this model, this is likely to be because treatment effects in this model were from pooling many more studies. The incremental costs of the study were also much larger than this model found, because the published study also included other costs not just intervention costs, and these showed much higher health service costs in the exercise group at 18-24 months after intervention (i.e. they were using more health services). Although this is only one study, so we cannot be certain this is the true direction of effect on resource use. A second Spanish economic evaluation was also identified that showed that pool-based aerobics was cost effective. This study found much higher QALYs than the model in this report because the study has been pooled with other studies in this model that had lower QoL. This study however had limitations in terms of the costs of the staff involved looking very low compared to UK costs, which will impact the cost effectiveness.

Other NICE guidelines have looked at the cost effectiveness of exercise versus no exercise in chronic pain populations. The NICE guidelines on Osteoarthritis,[20] and low back pain[23] also found published economic evidence suggesting exercise was cost effective. Group exercise programs were recommended for the low back pain guideline. Exercise was also recommended in osteoarthritis guideline, but it was not stated specifically whether there was an expectation for the NHS to provide this.

It was also noted that public health guidance on exercise referral schemes found referral for exercise not to be cost effective.[22] However, this guidance is for people who are otherwise healthy but are sedentary or inactive, and referral aims to improve activity to reduce the lifetime risk of coronary heart disease (CHD), stroke and type 2 diabetes. This scenario is different to what is being analysed here because the purpose of the public health interventions are principally to reduce avoidable deaths and wider comorbidities. The benefits captured in the chronic pain model in this write-up focuses on quality of life changes related to symptom benefit in respect of their chronic pain. Because the populations are likely to be different, albeit with some overlap, it is difficult to compare the effectiveness and cost effectiveness of exercise in a population using exercise to reduce future risk of coronary events, to a population using exercise to relieve symptoms from a specific condition.

4.5 Conclusions

Supervised exercise has been found to be cost effective in the chronic primary pain population, using pooled data from various trials to reflect the quality of life improvement over time from exercise, and taking into account the cost of the programmes.

4.6 Implications for future research

This analysis has shown that exercise is likely to be cost effective. However more research should be undertaken on the effectiveness of exercise that also includes utility measures as outcomes, to allow more data to be available for economic evaluations that can avoid mapping methods.

References

1. Andrade CP, Zamuner AR, Forti M, Tamburus NY, Silva E. Effects of aquatic training and detraining on women with fibromyalgia: controlled randomized clinical trial. European journal of physical & rehabilitation medicine. 2019; 55(1):79-88

2. Ara R, Brazier J. Deriving an algorithm to convert the eight mean SF-36 dimension scores into a mean EQ-5D preference-based score from published studies (where patient level data are not available). Value in Health. 2008; 11(7):1131-43

3. Assumpcao A, Matsutani LA, Yuan SL, Santo AS, Sauer J, Mango P et al. Muscle stretching exercises and resistance training in fibromyalgia: which is better? A three-arm randomized controlled trial. European Journal of Physical and Rehabilitation Medicine. 2018; 54(5):663-670

4. Baptista AS, Villela AL, Jones A, Natour J. Effectiveness of dance in patients with fibromyalgia: a randomized, single-blind, controlled study. Clinical and Experimental Rheumatology. 2012; 30(6 Suppl 74):18-23

5. Barton GR, Sach TH, Jenkinson C, Avery AJ, Doherty M, Muir KR. Do estimates of cost-utility based on the EQ-5D differ from those based on the mapping of utility scores? Health Qual Life Outcomes. 2008; 6:51

6. Beasley M, Prescott GJ, Scotland G, McBeth J, Lovell K, Keeley P et al. Patient-reported improvements in health are maintained 2 years after completing a short course of cognitive behaviour therapy, exercise or both treatments for chronic widespread pain: long-term results from the MUSICIAN randomised controlled trial. RMD Open. 2015; 1:e000026

7. Chan KK, Willan AR, Gupta M, Pullenayegum E. Underestimation of uncertainties in health utilities derived from mapping algorithms involving health-related quality-of-life measures: statistical explanations and potential remedies. Medical Decision Making. 2014; 34(7):863-72

8. Chuang LH, Whitehead SJ. Mapping for economic evaluation. British Medical Bulletin. 2012; 101:1-15

9. Cochrane Handbook for Systematic Reviews of Interventions 5.1.0 [updated March 2011]. Higgins J, Green S. The Cochrane Collaboration. 2011. Available from: www.cochrane-handbook.org

10. Curtis L, Burns A. Unit costs of health and social care 2018. Canterbury. Personal Social Services Research Unit University of Kent, 2018. Available from: https://www.pssru.ac.uk/project-pages/unit-costs/unit-costs-2018/

11. Falla D, Lindstrom R, Rechter L, Boudreau S, Petzke F. Effectiveness of an 8-week exercise programme on pain and specificity of neck muscle activity in patients with chronic neck pain: a randomized controlled study. European Journal of Pain. 2013; 17(10):1517-28

12. Garcia-Martinez AM, De Paz JA, Marquez S. Effects of an exercise programme on self-esteem, self-concept and quality of life in women with fibromyalgia: a randomized controlled trial. Rheumatology International. 2012; 32(7):1869-76

13. Gusi N, Tomas-Carus P. Cost-utility of an 8-month aquatic training for women with fibromyalgia: a randomized controlled trial. Arthritis Research & Therapy. 2008; 10(1):R24

14. Gusi N, Tomas-Carus P, Hakkinen A, Hakkinen K, Ortega-Alonso A. Exercise in waist-high warm water decreases pain and improves health-related quality of life and strength in the lower extremities in women with fibromyalgia. Arthritis and Rheumatism. 2006; 55(1):66-73

15. Kayo AH, Peccin MS, Sanches CM, Trevisani VF. Effectiveness of physical activity in reducing pain in patients with fibromyalgia: a blinded randomized clinical trial. Rheumatology International. 2012; 32(8):2285-92

16. Lauche R, Stumpe C, Fehr J, Cramer H, Cheng YW, Wayne PM et al. The effects of tai chi and neck exercises in the treatment of chronic nonspecific neck pain: A randomized controlled trial. Journal of Pain. 2016; 17(9):1013-27

17. Lynch M, Sawynok J, Hiew C, Marcon D. A randomized controlled trial of qigong for fibromyalgia. Arthritis Research & Therapy. 2012; 14(4):R178

18. MacPherson H, Vickers A, Bland M, Torgerson D, Corbett M, Spackman E et al. Acupuncture for chronic pain and depression in primary care: a programme of research. NIHR Journals Library Programme Grants for Applied Research. 2017; 5(3):342

19. Michalsen A, Traitteur H, Ludtke R, Brunnhuber S, Meier L, Jeitler M et al. Yoga for chronic neck pain: a pilot randomized controlled clinical trial. Journal of Pain. 2012; 13(11):1122-30

20. National Clinical Guideline Centre. Osteoarthritis: care and management in adults. NICE clinical guideline 177. London. National Clinical Guideline Centre, 2014. Available from: http://guidance.nice.org.uk/CG177

21. National Institute for Health and Care Excellence. Developing NICE guidelines: the manual. London. National Institute for Health and Care Excellence, 2014. Available from: http://www.nice.org.uk/article/PMG20/chapter/1%20Introduction%20and%20overview

22. National Institute for Health and Care Excellence. Physical activity: exercise referral schemes. NICE public health guidance 54. London. National Institute for Health and Care Excellence, 2014. Available from: http://guidance.nice.org.uk/PH54

23. National Institute for Health and Care Excellence. Low back pain and sciatica in over 16s: assessment and management. NICE guideline 59. London. National Institute for Health and Care Excellence, 2016. Available from: https://www.nice.org.uk/guidance/ng59

24. National Institute for Health and Clinical Excellence. Social value judgements: principles for the development of NICE guidance. London. National Institute for Health and Clinical Excellence, 2008. Available from: https://www.nice.org.uk/media/default/about/what-we-do/research-and-development/social-value-judgements-principles-for-the-development-of-nice-guidance.pdf

25. Rendant D, Pach D, Ludtke R, Reisshauer A, Mietzner A, Willich SN et al. Qigong versus exercise versus no therapy for patients with chronic neck pain: a randomized controlled trial. Spine. 2011; 36(6):419-27

26. Sanudo B, Galiano D, Carrasco L, de Hoyo M, McVeigh JG. Effects of a prolonged exercise program on key health outcomes in women with fibromyalgia: a randomized controlled trial. Journal of Rehabilitation Medicine. 2011; 43(6):521-6

27. Statistics OfN. Life tables for England. 2018. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lif

eexpectancies/datasets/nationallifetablesenglandreferencetables Last accessed: 24/09/2019

28. Tomas-Carus P, Raimundo A, Timon R, Gusi N. Exercise in warm water decreases pain but no the number of tender points in women with fibromyalgia: a randomized controlled trial. Seleccion. 2007; 16(2):98-102

29. van Eijk-Hustings Y, Kroese M, Creemers A, Landewe R, Boonen A. Resource utilisation and direct costs in patients with recently diagnosed fibromyalgia who are offered one of three different interventions in a randomised pragmatic trial. Clinical Rheumatology. 2016; 35(5):1307-1315

30. van Eijk-Hustings Y, Kroese M, Tan F, Boonen A, Bessems-Beks M, Landewe R. Challenges in demonstrating the effectiveness of multidisciplinary treatment on quality of life, participation and health care utilisation in patients with fibromyalgia: a randomised controlled trial. Clinical Rheumatology. 2013; 32(2):199-209

31. von Trott P, Wiedemann AM, Ludtke R, Reishauer A, Willich SN, Witt CM. Qigong and exercise therapy for elderly patients with chronic neck pain (QIBANE): a randomized controlled study. Journal of Pain. 2009; 10(5):501-8

Appendix A: Data extracted from studies

A.1 SF-36 raw data

Intervention	Measurement timeframe		SF-36 domain								EQ-5D Mapped from SF-36	EQ-5D change from baseline	EQ-5D improvement from exercise (a)
			Physical functioning	Social role	Physical role	Emotional role	Mental health	Vitality	Bodily pain	General health			
Sanudo (2011) (b)(c)													
Exercise	Baseline	Mean	50	55.2	13.5	53.3	51.3	29.4	23.2	39.8	0.53		
		Lower CI	39.7	44.8	5.6	32.7	42.7	22.4	15.3	32.5	0.40		
		Upper CI	60.3	65.6	21.4	73.9	59.9	36.4	31.1	47.1	0.64		
	Post intervention (at 24 weeks)	Mean	56.8	63.9	21.3	71.1	60	41.3	29.9	43.1	0.62	0.09	0.098
		Lower CI	48.9	53.1	9.2	52.2	53.2	35.0	22.3	38.1	0.52		
		Upper CI	64.7	74.7	33.4	90.0	66.8	47.6	37.5	48.1	0.70		
Control	Baseline	Mean	44.6	48.6	19.8	45.6	44	27.7	23.6	33.4	0.47		
		Lower CI	37.4	41.1	7.2	27.2	34.6	19.7	15.5	27.9	0.35		
		Upper CI	51.8	56.1	32.4	64.0	53.4	35.7	31.7	38.9	0.58		
	Post intervention (at 24 weeks)	Mean	45.2	52.2	19.4	52.1	44.2	28.6	19.5	33.5	0.46	-0.01	
		Lower CI	38.8	42.6	6.2	31.9	33.3	20.0	11.3	28.3	0.34		
		Upper CI	51.6	61.8	32.6	72.3	55.1	37.2	27.7	38.7	0.57		
Tomas-carus (2007) (c)													
Exercise	Baseline	Mean	36.0	54.0	35.0	37.0	48.0	30.0	21.0	32.0	0.44		
		Lower CI	24.2	36.5	16.5	13.9	37.7	22.3	11.2	19.7	0.26		
		Upper CI	47.8	71.5	53.5	60.1	58.3	37.7	30.8	44.3	0.59		
	Post intervention (at 12 weeks)	Mean	55.0	79.0	34.0	65.0	66.0	47.0	44.0	40.0	0.69	0.25	0.209
		Lower CI	39.6	66.1	15.0	41.3	54.7	36.2	32.2	27.7	0.53		
		Upper CI	70.4	91.9	53.0	88.7	77.3	57.8	55.8	52.3	0.82		
	Follow up (at 24 weeks)	Mean	48.0	60.0	29.0	75.0	62.0	35.6	43.0	33.0	0.64	0.20	0.154
		Lower CI	37.2	43.0	7.9	56.5	48.1	22.7	33.2	19.1	0.48		
		Upper CI	58.8	77.0	50.1	93.5	75.9	48.5	52.8	46.9	0.76		
Control	Baseline	Mean	33.0	52.0	25.0	33.0	51.0	20.0	23.0	29.0	0.44		
		Lower CI	23.2	38.6	12.1	11.9	38.7	12.8	13.2	21.3	0.27		

Intervention	Measurement timeframe		SF-36 domain								EQ-5D Mapped from SF-36	EQ-5D change from baseline	EQ-5D improvement from exercise (a)
			Physical functioning	Social role	Physical role	Emotional role	Mental health	Vitality	Bodily pain	General health			
		Upper CI	42.8	65.4	37.9	54.1	63.3	27.2	32.8	36.7	0.58		
	Post intervention (at 12 weeks)	Mean	37.0	57.0	25.0	31.0	50.0	25.0	28.0	27.0	0.48	0.04	
		Lower CI	28.3	44.7	10.6	13.5	39.7	17.3	17.7	19.3	0.33		
		Upper CI	45.7	69.3	39.4	48.5	60.3	32.7	38.3	34.7	0.61		
	Follow up (at 24 weeks)	Mean	37.0	57.0	22.0	31.0	50.0	25.0	28.0	27.0	0.48	0.04	
		Lower CI	28.3	44.7	8.6	13.5	39.7	17.3	17.7	19.3	0.33		
		Upper CI	45.7	69.3	35.4	48.5	60.3	32.7	38.3	34.7	0.61		
Baptista (2012) (c)													
Exercise	Baseline	Mean	44.9	52.6	24.7	34.2	46.0	41.3	29.6	46.0	0.52		
		Lower CI	44.3	43.7	14.4	22.4	39.6	35.3	24.0	39.1	0.45		
		Upper CI	45.5	61.5	35.0	46.0	52.4	47.3	35.2	52.9	0.58		
	Post intervention (at 16 weeks)	Mean	52.9	64.1	40.5	55.0	54.2	50.0	44.7	45.0	0.65	0.13	0.128
		Lower CI	46.2	55.1	30.7	44.3	47.6	42.7	38.1	38.2	0.56		
		Upper CI	59.6	73.1	50.3	65.7	60.8	57.3	51.3	51.8	0.73		
	Follow up (at 32 weeks)	Mean	56.3	57.2	36.5	51.9	52.3	47.6	46.0	44.9	0.65	0.14	0.071
		Lower CI	49.9	48.6	26.1	39.2	45.6	40.0	39.9	39.9	0.57		
		Upper CI	62.7	65.8	46.9	64.6	59.0	55.2	52.1	49.9	0.73		
Control	Baseline	Mean	32.6	47.6	8.8	21.2	43.4	29.0	25.7	38.0	0.42		
		Lower CI	26.6	40.2	3.1	10.6	35.7	23.2	21.4	32.7	0.33		
		Upper CI	38.6	55.0	14.5	31.8	51.1	34.8	30.0	43.3	0.51		
	Post intervention (at 16 weeks)	Mean	33.1	47.6	10.4	17.5	44.5	30.7	25.1	38.1	0.42	0.00	
		Lower CI	27.2	39.8	3.5	9.2	36.0	24.9	20.6	32.2	0.33		
		Upper CI	39.0	55.4	17.3	25.8	53.0	36.5	29.6	44.0	0.51		
	Follow up (at 32 weeks)	Mean	39.1	51.3	13.8	31.5	46.2	37.1	29.1	41.5	0.49	0.06	
		Lower CI	32.1	43.1	5.3	19.1	39.0	30.1	22.4	34.7	0.38		
		Upper CI	46.1	59.5	22.3	43.9	53.4	44.1	35.8	48.3	0.58		
Von trott (2009) (c)													
Exercise (Qigong)	Baseline	Mean	32.9	44.1	35.7	38.8	43.1	43.6	26.9	35.8	0.41		
		Lower CI	29.7	40.6	31.6	34.3	40.2	41.0	25.6	33.5	0.37		
		Upper CI	36.1	47.6	39.8	43.3	46.0	46.2	28.2	38.1	0.45		
	Post intervention (at 3 months)	Mean	33.5	45.6	37.1	43.0	43.9	42.1	27.8	36.3	0.42	0.01	0.031
		Lower CI	30.2	42.6	34.0	39.3	40.4	39.6	26.2	33.3	0.38		

Intervention	Measurement timeframe		SF-36 domain								EQ-5D Mapped from SF-36	EQ-5D change from baseline	EQ-5D improvement from exercise (a)
			Physical functioning	Social role	Physical role	Emotional role	Mental health	Vitality	Bodily pain	General health			
		Upper CI	36.8	48.6	40.2	46.7	47.4	44.6	29.4	39.3	0.46		
	Follow up (at 6 months)	Mean	33.5	40.4	35.6	38.6	40.3	40.5	27.2	36.1	0.40	-0.01	0.016
		Lower CI	29.9	37.0	31.9	33.9	37.1	37.8	25.8	33.3	0.35		
		Upper CI	37.1	43.8	39.3	43.3	43.5	43.2	28.6	38.9	0.44		
Exercise (neck exercises)	Baseline	Mean	30.8	43.9	35.1	43.0	44.9	42.1	27.9	37.0	0.41		
		Lower CI	27.9	40.0	31.8	38.4	41.9	39.1	26.4	34.1	0.37		
		Upper CI	33.7	47.8	38.4	47.6	47.9	45.1	29.4	39.9	0.45		
	Post intervention (at 3 months)	Mean	30.3	44.6	37.0	42.1	43.9	42.3	28.4	37.2	0.41	-0.00	0.015
		Lower CI	27.4	41.4	33.2	37.6	40.3	38.8	26.7	34.9	0.36		
		Upper CI	33.2	47.8	41.2	46.6	47.5	45.8	30.1	39.5	0.45		
	Follow up (at 6 months)	Mean	30.5	42.9	34.8	39.2	41.9	41.8	27.3	34.8	0.39	-0.02	0.008
		Lower CI	27.0	39.6	31.3	34.9	37.8	38.9	26.0	31.4	0.34		
		Upper CI	34.0	46.2	38.3	43.5	46.0	44.7	28.6	38.2	0.44		
Control	Baseline	Mean	32.8	47.4	36.5	43.8	43.3	43.6	27.2	39.4	0.42		
		Lower CI	29.0	44.6	33.0	39.8	39.9	40.6	25.9	36.8	0.37		
		Upper CI	36.6	50.2	40.1	47.8	46.7	46.6	28.5	42.0	0.46		
	Post intervention (at 3 months)	Mean	30.8	44.5	36.4	42.8	43.4	41.6	26.6	36.4	0.40	-0.02	
		Lower CI	27.2	40.9	32.5	38.5	39.8	38.5	25.3	33.3	0.35		
		Upper CI	34.4	48.1	40.3	47.1	47.0	44.7	27.9	39.5	0.44		
	Follow up (at 6 months)	Mean	30.9	42.5	35.5	36.5	40.7	42.4	27.7	36.9	0.39	-0.03	
		Lower CI	26.9	39.0	32.2	32.6	37.4	39.5	26.4	34.0	0.34		
		Upper CI	34.9	46.0	38.8	40.4	44.0	45.3	29.0	39.8	0.44		
Garcia-martinez (2012) (c)													
Exercise	Baseline	Mean	33.9	54.2	11.1	37.0	46.6	24.4	26.7	30.0	0.44		
		Lower CI	23.2	44.7	-3.5 (d)	9.1	36.1	14.6	14.3	24.4	0.27		
		Upper CI	44.6	63.7	25.7	64.9	57.1	34.2	39.1	35.6	0.59		
	Post intervention (at 3 months)	Mean	50.6	80.8	47.2	59.2	61.7	38.9	47.5	30.6	0.67	0.22	0.31
		Lower CI	34.8	68.3	25.0	30.7	51.0	26.6	34.9	18.2	0.50		
		Upper CI	66.4	93.3	69.4	87.7	72.4	51.2	60.1	43.0	0.80		
Control	Baseline	Mean	39.6	47.1	5.8	28.2	48.0	36.1	34.6	29.6	0.50		
		Lower CI	29.5	32.9	-2.8 (d)	2.3	35.7	26.2	25.9	17.1	0.34		
		Upper CI	49.7	61.3	14.4	54.1	60.3	46.0	43.3	42.1	0.64		

Intervention	Measurement timeframe		SF-36 domain								EQ-5D Mapped from SF-36	EQ-5D change from baseline	EQ-5D improvement from exercise (a)
			Physical functioning	Social role	Physical role	Emotional role	Mental health	Vitality	Bodily pain	General health			
	Post intervention (at 3 months)	Mean	35.4	42.3	7.7	23.0	41.5	21.5	25.2	28.9	0.41	-0.09	
		Lower CI	25.5	30.3	-8.3	0.3	29.8	11.5	17.8	17.6	0.25		
		Upper CI	45.3	54.3	23.7	45.7	53.2	31.5	32.6	40.2	0.55		
Rendant (2011) (e)													
Exercise (Qigong)	Baseline	Mean	77.8	73.8	62.5	77.0	64.1	44.4	48.8	60.4	0.78		
		Lower CI	71.7	66.9	51.6	66.1	59.0	38.5	42.8	54.3	0.72		
		Upper CI	83.9	80.7	73.4	87.9	69.2	50.3	54.8	66.5	0.83		
	Partway through intervention (at 3 months)	Mean	82.9	82.2	78.0	88.3	70.4	57.5	63.3	61.9	0.86	0.08	0.092
		Lower CI	78.8	75.6	67.2	81.2	65.0	52.7	56.9	56.5	0.81		
		Upper CI	86.9	88.9	88.7	95.4	75.8	62.4	69.8	67.2	0.90		
	Post intervention (at 6 months)	Mean	80.2	81.1	77.8	76.6	68.7	51.5	63.6	62	0.85	0.07	0.083
		Lower CI	76.3	73.5	66.8	65.5	63.9	46.9	57.2	57.7	0.80		
		Upper CI	84.1	88.8	88.7	87.8	73.4	56.1	70.1	66.3	0.89		
Exercise (neck exercises)	Baseline	Mean	77.4	73.4	66.7	67.5	65.5	48.5	48.9	58.4	0.78		
		Lower CI	70.9	65.7	53.2	53.4	59.3	43.2	43.2	52.3	0.72		
		Upper CI	83.9	81.1	80.2	81.6	71.7	53.8	54.6	64.5	0.83		
	Partway through intervention (at 3 months)	Mean	78.5	75.3	62.0	74.5	66.4	49.4	61.8	63.6	0.83	0.06	0.068
		Lower CI	72.1	67.0	48.9	62.7	61.9	44.5	54.6	58.8	0.78		
		Upper CI	84.9	83.6	75.1	86.3	70.8	54.2	69.1	68.4	0.89		
	Post intervention (at 6 months)	Mean	79.2	75.6	63.1	81.2	68.1	49.2	62.5	61.9	0.84	0.06	0.076
		Lower CI	73.7	69.9	51.6	72.1	62.6	43.7	56.3	56.5	0.79		
		Upper CI	84.7	82.1	74.5	90.3	73.5	54.7	68.8	67.2	0.89		
Control	Baseline	Mean	77.8	79.6	67.7	80.5	68.6	49.0	50.7	60.9	0.80		
		Lower CI	72.1	72.4	56.9	70.0	63.8	43.6	45.7	54.9	0.75		
		Upper CI	83.5	86.8	78.5	91.0	73.4	54.4	55.7	66.9	0.84		
	Partway through intervention (at 3 months)	Mean	75.1	74.2	63.4	70.9	63.9	43.2	53.6	55.7	0.79	-0.01	
		Lower CI	71.4	68	52.4	61.3	59.4	38.7	48.5	51.8	0.74		
		Upper CI	78.8	80.4	74.3	80.4	68.5	47.8	58.8	59.5	0.83		
	Post intervention (at 6 months)	Mean	74.8	74.1	60.6	75.8	62.1	43.1	54.2	57.6	0.79	-0.01	
		Lower CI	70	68.2	49.6	66.5	57.9	38.5	48.8	53.5	0.74		
		Upper CI	79.5	80.1	71.6	85.1	66.3	47.7	59.6	61.8	0.83		
Lauche (2016) (c)													

Intervention	Measurement timeframe		SF-36 domain								EQ-5D Mapped from SF-36	EQ-5D change from baseline	EQ-5D improvement from exercise (a)
			Physical functioning	Social role	Physical role	Emotional role	Mental health	Vitality	Bodily pain	General health			
Exercise (Tai chi)	Baseline	Mean	78.5	73.0	62.5	64.0	68.9	51.4	46.3	68.3	0.78		
		Lower CI	74.2	65.1	51.7	51.9	63.6	46.3	37.9	63.5	0.71		
		Upper CI	82.8	80.9	73.3	76.1	74.2	56.5	54.7	73.1	0.83		
	Post intervention (at 12 weeks)	Mean	81.1	79.2	70.0	68.3	67.8	56.5	58.5	70.7	0.83	0.05	0.070
		Lower CI	75.5	71.4	57.6	54.6	61.7	50.8	52.5	65.5	0.78		
		Upper CI	86.7	87.0	82.4	82.0	73.9	62.2	64.5	75.9	0.88		
	Follow up (at 24 weeks)	Mean	79.6	77.9	67.7	68.4	68.4	55.6	58.6	68.3	0.83	0.05	0.060
		Lower CI	74.0	69.8	55.5	56.5	61.8	48.9	51.2	63.0	0.77		
		Upper CI	85.2	86.0	79.9	80.3	75.0	62.3	66.0	73.6	0.88		
Exercise (neck exercises)	Baseline	Mean	77.4	68.9	51.4	72.1	68.2	48.2	45.1	64.4	0.77		
		Lower CI	72.3	62.3	39.8	61.1	64.0	43.2	40.6	58.5	0.72		
		Upper CI	82.5	75.5	63.0	83.1	72.4	53.2	49.6	70.3	0.81		
	Post intervention (at 12 weeks)	Mean	80.3	72.6	66.1	72.1	69.9	52.5	55.2	64.6	0.82	0.05	0.065
		Lower CI	43.1	67.0	56.7	62.7	65.2	47.6	51.1	59.5	0.68		
		Upper CI	117.5	78.2	75.5	81.5	74.6	57.4	59.3	69.7	0.84		
	Follow up (at 24 weeks)	Mean	77.4	71.2	60.2	65.4	69.4	50.7	56.9	61.9	0.81	0.05	0.055
		Lower CI	71.6	64.4	50.0	54.7	64.4	44.8	51.6	55.9	0.76		
		Upper CI	83.2	78.0	70.4	76.1	74.4	56.6	62.2	67.9	0.86		
Control	Baseline	Mean	79.1	75.6	53.2	70.9	66.8	49.9	50.6	67.4	0.79		
		Lower CI	74.7	69.1	42.5	58.0	61.5	44.3	44.7	61.2	0.74		
		Upper CI	83.5	82.1	63.9	83.8	72.1	55.5	56.5	73.6	0.84		
	Post intervention (at 12 weeks)	Mean	74.6	70.3	53.4	62.9	65.9	49.7	50.3	64.5	0.78	-0.02	
		Lower CI	68.3	63.9	43.1	50.4	60.2	44.2	46.5	58.7	0.73		
		Upper CI	80.9	76.7	63.7	75.4	71.6	55.2	54.1	70.3	0.82		
	Follow up (at 24 weeks)	Mean	74.0	68.9	49.9	65.2	65.9	47.6	53.6	59.7	0.79	-0.01	
		Lower CI	67.8	61.5	42.2	53.1	60.5	41.1	48.5	53.7	0.73		
		Upper CI	80.2	76.3	57.6	77.3	71.3	54.1	58.7	65.7	0.84		
Andrade (2019) (c)													
Exercise	Baseline	Mean	44.6	48.1	10.2	24.7	48.6	33.5	31.8	43.1	0.53		
		Lower CI	37.6	41.0	-0.9	10.7	39.9	26.1	25.4	35.6	0.42		
		Upper CI	51.6	55.2	21.3	38.7	57.3	40.9	38.2	50.6	0.62		
		Mean	50.5	54.3	29.8	32.1	46.8	37.9	36.7	48.9	0.58	0.05	0.02

Intervention	Measurement timeframe		SF-36 domain									EQ-5D Mapped from SF-36	EQ-5D change from baseline	EQ-5D improvement from exercise (a)
			Physical functioning	Social role	Physical role	Emotional role	Mental health	Vitality	Bodily pain	General health				
	Post intervention (at 16 weeks)	Lower CI	43.5	45.5	13.6	16.0	37.7	29.0	20.5	40.4		0.43		
		Upper CI	57.5	63.1	46.0	48.2	55.9	46.8	52.9	57.4		0.71		
Control	Baseline	Mean	38.2	44.5	11.0	18.7	37.8	25.4	25.5	44.1		0.43		
		Lower CI	32.7	36.5	1.1	7.1	31.5	19.6	21.1	36.2		0.34		
		Upper CI	43.7	52.5	20.9	30.3	44.1	31.2	29.9	52.0		0.51		
	Post intervention (at 16 weeks)	Mean	38.0	45.4	13.8	22.4	43.4	30.2	29.2	41.0		0.47	0.03	
		Lower CI	32.2	36.3	2.8	8.4	36.6	24.2	24.4	32.9		0.37		
		Upper CI	43.8	54.5	24.8	36.4	50.2	36.2	34.0	49.1		0.55		

Note: Blue in the table means outcome is measured partway through the intervention. Green in the table means outcomes are measured right after the intervention ended (post-intervention outcomes). Beige in the table means outcomes measured later after the intervention ended (follow-up outcomes).

(a) EQ-5D change from baseline in the exercise group minus the EQ-5D change from baseline in the control group. This is calculated for each measurement point, of which some trials have more than one (e.g. outcomes in some trials are measures at the end of the intervention but also have a later follow-up). For example: For Tomas-Carus (2007), outcomes are measured at 12 weeks and at 24 weeks. So the EQ-5D improvement at 12 weeks is the change in baseline in the exercise group at 12 weeks minus the change in baseline in the control group at 12 weeks (0.25 - 0.04 = 0.209). The same is then calculated for the 24 week outcomes.

(b) Labelled as CI's but some are bigger than the mean so have been treated as SD's

(c) Calculated CI's from SDs reported in paper using revman software.

(d) Some confidence intervals that were calculated for the SF-36 returned negative values. This was not an issue in this study because the regression that the mapping function is based on does not involve all the domains, and the physical role domain is one of these domains and therefore did not influence the mapping.

(e) Paper reported confidence intervals.

A.2 EQ-5D raw data

Intervention	Measurement timeframe		EQ-5D value	EQ-5D change from baseline	EQ-5D improvement from exercise (a)
Gusi (2006)					
Exercise	Baseline (a)	Mean	0.29		
		Lower CI	0.15		
		Upper CI	0.43		
	Post intervention (at 12 weeks) (b) (c)	Mean	0.56	0.27	0.29
		Lower CI	0.41		
		Upper CI	0.71		
	Follow up (at 24 weeks)		0.43	0.14	0.16
			0.26		
			0.61		
Control	Baseline (a)	Mean	0.32		
		Lower CI	0.16		
		Upper CI	0.48		
	Post intervention (at 12 weeks) (b) (c)	Mean	0.30	-0.02	
		Lower CI	0.14		
		Upper CI	0.45		
	Follow up (at 24 weeks)	Mean	0.30	-0.02	
		Lower CI	0.15		
		Upper CI	0.45		
Beasley (2015)					
Exercise	Baseline	Mean	0.69		
		Lower CI	0.65		
		Upper CI	0.73		
	Post intervention (at 6 months)	Mean	0.72	0.03	-0.01
		Lower CI	0.67		
		Upper CI	0.76		
	Follow up (at 9 months)	Mean	0.71	0.02	0.023
		Lower CI	0.66		
		Upper CI	0.75		
	Follow up (at 30 months)	Mean	0.71	0.03	0.044
		Lower CI	0.65		
		Upper CI	0.77		
Control	Baseline	Mean	0.65		
		Lower CI	0.61		
		Upper CI	0.69		
	Post intervention (at 6 months)	Mean	0.69	0.04	
		Lower CI	0.63		
		Upper CI	0.74		
	Follow up (at 9 months)	Mean	0.65	-0.00	
		Lower CI	0.63		
		Upper CI	0.75		
	Follow up (at 30 months)	Mean	0.63	-0.02	
		Lower CI	0.56		
		Upper CI	0.70		
Van Eijk-Hustings (2013)					
Exercise	Baseline (a)	Mean	0.41		
		Lower CI	0.40		
		Upper CI	0.43		
		Mean	0.47	0.06	0.07

	Post intervention (at 12 weeks) (a)	Lower CI	0.46		
		Upper CI	0.59		
	Follow up (at 21 months) (a)	Mean	0.54	0.13	0.13
		Lower CI	0.53		
		Upper CI	0.56		
Control	Baseline (a)	Mean	0.51		
		Lower CI	0.50		
		Upper CI	0.52		
	Post intervention (at 12 weeks) (a)	Mean	0.50	-0.01	
		Lower CI	0.49		
		Upper CI	0.51		
	Follow up (at 21 months) (a)	Mean	0.51	0.00	
		Lower CI	0.46		
		Upper CI	0.56		
Gusi (2008)					
Exercise	Baseline	Mean	0.32		
		Lower CI	0.16		
		Upper CI	0.47		
	Partway through intervention (at 3 months)	Mean	0.58	0.27	0.263
		Lower CI	0.43		
		Upper CI	0.73		
	Post intervention (at 8 months)	Mean	0.53	0.21	0.21
		Lower CI	0.38		
		Upper CI	0.68		
Control	Baseline	Mean	0.33		
		Lower CI	0.15		
		Upper CI	0.51		
	Partway through intervention (at 3 months)	Mean	0.33	0.003	
		Lower CI	0.18		
		Upper CI	0.50		
	Post intervention (at 8 months)	Mean	0.33	0.00	
		Lower CI	0.18		
		Upper CI	0.49		

Note: Blue in the table means outcome is measured partway through the intervention. Green in the table means outcomes are measured right after the intervention ended (post-intervention outcomes). Beige in the table means outcomes measured later after the intervention ended (follow-up outcomes).
(a) Calculated CI's from SDs reported in paper using revman software.
(b) Reported as change scores so back calculated to derive EQ-5D.
(c) Confidence interval was reported for the change scores.

Appendix B: Data for meta-analysis

B.1 Data for meta-analyis

										Feeding into meta-analysis						
Study	**Intervention**	**EQ-5D baseline mean**	**EQ-5D mean - follow up 1**	**EQ-5D mean - follow up 2**	**EQ-5D mean - follow up 3**	**Baseline SD**	**Follow up 1 SD**	**Follow up 2 SD**	**Follow up 3 SD**	**EQ-5D change from baseline (timepoint 1) (b)**	**EQ-5D change from baseline (timepoint 2) (b)**	**EQ-5D change from baseline (timepoint 3) (b)**	**change from baseline SD (timepoint 1) (a)**	**change from baseline SD (timepoint 2) (a)**	**change from baseline SD (timepoint 3) (a)**	**N**
Sanudo (2011)	Exercise	0.528	0.617			0.276	0.211			0.089			0.257			21
	control	0.472	0.463			0.264	0.274			-0.009			0.278			21
Tomas-carus (2007)	Exercise	0.441	0.691	0.636		0.344	0.297	0.296		0.250	0.196		0.333	0.333		17
	control	0.438	0.480	0.480		0.321	0.288	0.288		0.042	0.042		0.315	0.315		17
Baptista (2012)	Exercise	0.518	0.649	0.653		0.203	0.264	0.251		0.131	0.135		0.247	0.238		40
	control	0.422	0.425	0.486		0.290	0.300	0.323		0.003	0.064		0.305	0.318		40
Von trott (2009) (intvn arms combined)	Exercise	0.409	0.413	0.395		0.124	0.131	0.141		0.005	-0.014		0.132	0.138		77
	control	0.417	0.399	0.391		0.138	0.144	0.147		-0.018	-0.026		0.146	0.148		40
Garcia-martinez (2012)	Exercise	0.443	0.667			0.304	0.290			0.224			0.307			14
	control	0.498	0.410			0.288	0.285			-0.088			0.296			14
Rendant (2011) (intvn arms combined)	Exercise	0.778	0.847	0.846		0.179	0.161	0.153		0.068	0.068		0.176	0.173		81
	control	0.800	0.788	0.787		0.152	0.143	0.152		-0.012	-0.012		0.152	0.157		41
Lauche (2016) (intvn arms combined)	Exercise	0.773	0.823	0.820		0.163	0.215	0.165		0.050	0.047		0.200	0.169		75
	control	0.795	0.777	0.785		0.158	0.155	0.170		-0.018	-0.010		0.161	0.170		39
	Exercise	0.527	0.577	0.580		0.262	0.366	0.250		0.050	0.053		0.336	0.264		27

										Feeding into meta-analysis						
Andrade (2019)	control	0.431	0.466	0.466		0.227	0.239	0.239		0.034	0.034		0.241	0.241		27
Gusi (2006)	Exercise	0.290	0.560	0.430		0.280	0.316	0.368		0.270	0.140		0.316	0.368		17
	control	0.320	0.300	0.300		0.320	0.326	0.316		-0.020	-0.020		0.326	0.316		17
Beasley (2015)	Exercise	0.686	0.716	0.705	0.712	0.213	0.233	0.245	0.305	0.030	0.019	0.026	0.231	0.238	0.279	109
	control	0.649	0.688	0.645	0.631	0.219	0.289	0.305	0.378	0.039	-0.004	-0.018	0.269	0.280	0.337	109
Van eijk-hustings (2013)	Exercise	0.410	0.470	0.540		0.051	0.051	0.051		0.060	0.130		0.053	0.053		47
	control	0.510	0.500	0.510		0.041	0.041	0.051		-0.010	0.000		0.042	0.048		48
Gusi (2008)	Exercise	0.316	0.582	0.528		0.324	0.310	0.310		0.266	0.212		0.328	0.328		17
	control	0.331	0.334	0.334		0.368	0.326	0.324		0.003	0.003		0.360	0.360		16

Note: Blue means studies that had SF-36 data and therefore EQ-5D mean and follow up was mapped from SF-36, as well as their confidence intervals. Green means reported in the paper. Orange means calculated from change scores. Yellow means transformed using confidence intervals and the number of participants in the study. Follow up 1 = the first follow up point, and so on. SD = standard deviation.

(a) Calculated using the imputing SD formula from the Cochrane (Equation 2)
(b) Calculated by taking the difference from the follow up and baseline values.

B.2 Adjusted standard deviations for mapping uncertainty

						Unadjusted SD's				Adjusted SD's			
Study	**Intervention**	**EQ-5D baseline mean**	**EQ-5D mean - follow up 1**	**EQ-5D mean - follow up 2**	**EQ-5D mean - follow up 3**	**Baseline SD**	**Follow up 1 SD**	**Follow up 2 SD**	**Follow up 3 SD**	**Baseline SD**	**Follow up 1 SD**	**Follow up 2 SD**	**Follow up 3 SD**
Sanudo (2011)	Exercise	0.528	0.617			0.276	0.211			0.361	0.276		
	control	0.472	0.463			0.264	0.274			0.345	0.358		
Tomas-carus (2007)	Exercise	0.441	0.691	0.636		0.344	0.297	0.296		0.450	0.388	0.387	
	control	0.438	0.480	0.480		0.321	0.288	0.288		0.419	0.376	0.376	
Baptista (2012)	Exercise	0.518	0.649	0.653		0.203	0.264	0.251		0.265	0.346	0.328	
	control	0.422	0.425	0.486		0.290	0.300	0.323		0.379	0.392	0.423	
	Exercise	0.409	0.413	0.395		0.124	0.131	0.141		0.162	0.171	0.184	

						Unadjusted SD's				Adjusted SD's			
Von trott (2009) (intvn arms combined)	control	0.417	0.399	0.391		0.138	0.144	0.147		0.181	0.189	0.193	
Garcia-martinez (2012)	Exercise	0.443	0.667			0.304	0.290			0.398	0.379		
	control	0.498	0.410			0.288	0.285			0.376	0.373		
Rendant (2011) (intvn arms combined)	Exercise	0.778	0.847	0.846		0.179	0.161	0.153		0.233	0.210	0.200	
	control	0.800	0.788	0.787		0.152	0.143	0.152		0.199	0.187	0.198	
Lauche (2016) (intvn arms combined)	Exercise	0.773	0.823	0.820		0.163	0.215	0.165		0.214	0.281	0.216	
	control	0.795	0.777	0.785		0.158	0.155	0.170		0.207	0.202	0.223	
Andrade (2019)	Exercise	0.527	0.577	0.580		0.262	0.366	0.250		0.343	0.478	0.326	
	control	0.431	0.466	0.466		0.227	0.239	0.239		0.296	0.312	0.312	

Appendix C: Combining intervention arms of 3 arm trials

Study		N	mean baseline EQ-5D	mean EQ-5D follow up 1	mean EQ-5D follow up 2	Baseline SD	Follow up 1 SD	Follow up 2 SD
Von trott (2009)	strength arm	39	0.409	0.406	0.391	0.126	0.131	0.147
	mind body arm	38	0.408	0.421	0.398	0.124	0.132	0.136
	control	40	0.42	0.40	0.39	0.138	0.144	0.147
	COMBINED ARMS	77	0.41	0.41	0.39	0.1240	0.1309	0.1410
Rendant (2011)	strength + flexibility arm	39	0.778	0.834	0.842	0.181	0.175	0.157
	mind body arm	42	0.779	0.859	0.850	0.179	0.148	0.151
	control	41	0.800	0.788	0.787	0.152	0.143	0.152
	COMBINED ARMS	81	0.78	0.85	0.85	0.1786	0.1609	0.1530
Lauche (2016)	strength, proprioception and flexibility arm	37	0.769	0.817	0.814	0.138	0.261	0.150
	mind body arm	38	0.777	0.830	0.827	0.187	0.161	0.180
	control	39	0.79	0.78	0.78	0.158	0.155	0.170
	COMBINED ARMS	75	0.77	0.82	0.82	0.1634	0.2147	0.1649

Note: Follow up 1 = first follow up time point, follow up 2 = second follow up time point, SD = standard deviation