National Institute for Health and Care Excellence

Draft

Chronic pain: assessment and management

Cost-effectiveness analysis: Acupuncture 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 All NICE guidance is subject to regular review and may be updated or withdrawn.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

1 Introduction

A systematic review of the published clinical and economic evidence was undertaken as part of the guideline, comparing acupuncture with usual care, and sham acupuncture.

The clinical evidence showed a benefit of acupuncture compared to both sham acupuncture and usual care, in reducing pain and improving quality of life.

One UK-based within-trial economic analysis was identified for this review, comparing acupuncture in addition to usual care with usual care. This was in people with chronic neck pain and had a 1-year follow-up, although the intervention itself was around 5 months long (up to 12 x 50-minute treatments delivered once per week and then once every 2 weeks). Resource use included all appointments and prescriptions. The study found that acupuncture had an ICER of £18,767 per QALY gained, suggesting acupuncture is cost effective. The 95% confidence interval was very wide (95% CI: £,4,426 to £74,562). However, a sensitivity analysis where missing data was imputed (and 40% of data was missing in the acupuncture arm) showed an ICER of £43,838, again with a very large confidence interval (-£216,427 to £395,047). The committee opinion was that the confidence interval led to uncertainty around cost effectiveness, although this would be the more relevant study as it is from a UK perspective. The costs of providing acupuncture (£35 per session) are likely to be lower than current staff costs that might provide acupuncture in the NHS. This might be because of the date of the costs (2012/13) or also because the costs of the sessions were based on the level of practitioner delivering the intervention in the trial, which was unclear. A second study was identified which was a German within-trial analysis, comparing acupuncture to a waiting list control in people with chronic neck pain, with a three-month follow-up. People in the acupuncture group received between 10 to 15 sessions of acupuncture over the three months. The study considered costs of acupuncture as well as physician visits, medication and hospital stays in both groups. This paper suggested that acupuncture is cost effective compared to waiting list control (ICER: £11,430 per QALY gained). Although acupuncture costs were arbitrarily derived because acupuncture is not reimbursed by health insurance companies in Germany, and the costs per session (€35/£28) seem lower than UK costs. Both studies had limitations regarding intervention costs potentially being underestimated, and uncertainty remained around cost effectiveness.

Acupuncture for chronic pain is not currently used in the NHS, therefore, a recommendation could have a resource impact to the NHS in England 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).[11] An incremental analysis was undertaken.

2.1.1 Comparators

The comparators selected for the model were:

1. Acupuncture
2. No acupuncture

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

The data used from the clinical review were the studies with acupuncture versus usual care comparisons (and not the studies with acupuncture versus sham acupuncture). The committee agreed that this would reflect the real-world impact of acupuncture on people with chronic pain and so was the most appropriate to use in the economic evaluation as this aims to compare real-world alternatives. The committee noted that sham acupuncture would not be used outside of a research study. A more detailed discussion of this decision is provided in section 2.1.1.1 below.

The interventions in this review are all types of acupuncture, and therefore were considered more similar to each other than different types of exercise for example. However, there was still heterogeneity in the data. The committee noted the differences between the studies in terms of: the type of acupuncture (dry needling, traditional Chinese, Japanese style), intensity (i.e. frequency, duration, and total number of sessions), the likely staff delivering the acupuncture (not well reported however), and the varying descriptions of usual care (some studies only allowed medication or certain medication, some stated routine care or usual care without further definition). Noting all the complexities, the committee agreed that pooling the data would give a more reliable overall estimate of the likely cost effectiveness of acupuncture. Clearly, the results would need to be interpreted with caution given the heterogeneity in the data created by pooling different interventions from different time frames 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.

2.1.1.1 Using usual care evidence in the economic analysis

In economic evaluation we compare alternative real-world strategies quantifying the costs and health effects with each in order to inform decisions regarding which is the best option for use in practice given the budgetary constraints of the healthcare system. Arguably the most important input into such an analysis is the effectiveness data used to quantify the differences between alternative strategies. Randomised controlled trials (RCTs) (where the intervention of interest is compared to a control group receiving something else, for example no treatment, the standard treatment or placebo) are usually considered the most appropriate measure of relative treatment effect by NICE.[24]

In the acupuncture review for the guideline, RCTs were included that compared acupuncture to either a placebo (sham acupuncture) or to no acupuncture (that is usual care). The committee agreed that:

- sham evidence is important for assessing whether there are treatment-specific effects from acupuncture;
- however, the data comparing acupuncture as an adjunct to usual care with usual care alone should be used in the economic evaluation (as sham is not a real-world comparator).

This approach was also taken in a recent UK cost effectiveness analysis of acupuncture undertaken as part of an NIHR-funded research programme.[17] This approach was also consistent with the economic analysis undertaken for exercise for the guideline where exercise was compared to usual care. No appropriate 'placebo' was considered feasible for exercise in the clinical review.

A more detailed discussion about the issues and basis for this decision are discussed in detail below.

2.1.1.1.1 More detailed exploration of the issues around the choice of clinical data used in the economic analysis

Placebo-controlled comparisons

Placebos are used in trials to reduce bias as it means that participants, and ideally those administering the treatment, don't know whether they are receiving the active treatment or not. This means that if a treatment effect is observed we can be confident that it is attributable to treatment specific effects rather than say contextual or placebo effects.

For pharmacological agents using a placebo is usually straightforward as it requires simply producing an identical looking treatment without the active agent. However, for non-pharmacological treatments it is often difficult or not possible, for example, surgery or exercise. Where placebos have been developed for non-pharmacological interventions there are often complexities and uncertainties, for example about whether the placebo really is 'inert'. In addition, it is generally not possible for the practitioner to be unaware whether they are giving the real or placebo treatment. These issues can complicate the interpretation of placebo-controlled studies of non-pharmacological interventions. Sham acupuncture is often used as a placebo in acupuncture studies although its use has been much debated.

Usual care comparisons

Comparing an intervention (as an adjunct to usual care) to usual care (alone) is likely to give a better estimate of the real-world impact on outcomes should an intervention be implemented than a placebo comparison. This will include both treatment-specific and non-specific or contextual effects of the intervention. Non-specific effects may for example come from the process of care and information or advice given at the time of treatment. However, it is not possible to tell from a usual care comparison if any of the effect observed is due to treatment-specific effects. In addition, if the usual care in the study is not the same as the usual care in current practice in the health system of interest this may complicate interpretation.

Differences between intervention types

The availability of evidence with each of these types of comparison (placebo or usual care) tends to vary between types of intervention.

The use of placebo-controlled trials is well established and uncontroversial for demonstrating efficacy of a pharmacological intervention (although comparison with an alternative

established pharmacological agent is also commonly used). New pharmacological agents must provide evidence of efficacy from randomised controlled trials as part of the regulatory approval process before they can be used. The aim is to demonstrate with confidence that there is a benefit specifically attributable to the new treatment. After the medicine has been approved the manufacturer will not have much incentive to conduct additional trials comparing it with usual care or other active treatments.

For non-pharmacological interventions placebo-controlled studies are less routinely used because of the difficulties in developing, or in some cases absence of, an appropriate placebo (as described above) and as there is usually no regulatory requirement parallel to that for pharmacological treatments requiring evidence of efficacy. Contextual effects may potentially be more significant with non-pharmacological interventions because they typically involve more interaction between patients and health care practitioners.

Interpretation when there is both usual care and placebo comparisons

Given the issues outlined above, the committee agreed that RCTs comparing acupuncture to either sham or usual care alone should be included and analysed separately in the clinical review. They also agreed that there needed to be evidence of a treatment-specific effect from the placebo (sham)-controlled studies for it to be recommended. However, assuming this was the case the magnitude of effect from the usual care comparison studies would be considered. This was the approach also taken by the committee in the 2016 Low back pain and sciatica guideline.

It is noted that for some non-pharmacological interventions there was not considered to be an adequate placebo, and in these cases decisions had to be made using only usual care comparison studies taking into account the uncertainty this added. This is not uncommon in guidelines (surgery and exercise are common examples).

Appropriate comparisons for economic evaluation

Economic evaluations compare alternative real-world clinical options with the aim of informing decision making about their use. It is often advocated that economic evaluations should be ideally based on 'effectiveness' evidence rather than 'efficacy'.[13, 15, 27] Effectiveness is assessed with 'pragmatic' randomised trials that attempt to replicate real world conditions that would exist if the intervention were to be implemented in routine clinical practice. Hence patients should be typical of normal caseload and comparison should be with a relevant alternative (usual practice or the best alternative treatment strategy) with clinicians and patients un-blinded. [13] This way the incremental costs and health gain should closely reflect what will happen if the intervention is rolled out to the wider health service capturing all treatment-specific and non-specific health effects and only capturing real-world cost differences. A study of acupuncture that was based on a sham comparator would not pick up all the health effects (be they positive or negative) attributable to needling (since both trial arms have needling) but these health effects which would occur as part of routine practice.

In practice the data used in any economic evaluation will be limited by what is available at the time the analysis is undertaken. Economic evaluations of new pharmacological interventions are often undertaken at a time when the key evidence will be from efficacy trials. In addition, there may be a trade-off between different aspects of the study design and quality of the evidence, or different considerations depending on the type of intervention, and a judgement will have to be made about the most appropriate data for an analysis.

While economic evaluations of pharmacological interventions do often incorporate placebo-controlled data this is not the case with non-pharmacological interventions. Work undertaken at the NGC in 2015 (unpublished) found that of 28 economic evaluations of acupuncture for various indications:

- Sixteen of the studies evaluated acupuncture as an adjunct to 'usual care' (or in comparison to waiting list).
- Three studies compared acupuncture to sham (or non-penetrating acupuncture) and 2 studies compared acupuncture with usual care but used a sham control to estimate effectiveness.
- Some studies (either in addition or instead of comparing to usual care or sham) compared acupuncture to specific drug treatments (6 studies) or other active treatments (3 studies), or compared different types of acupuncture (2 studies).

In addition a recent UK cost effectiveness analysis of acupuncture for chronic pain (related to osteoarthritis, chronic or recurrent headaches (e.g. tension or migraine headaches), specific and non-specific shoulder pain, and non-specific back or neck pain) undertaken as part of an NIHR-funded research programme about acupuncture also used data comparing acupuncture with usual care on the basis that sham was not used in practice.[17] Sham and usual care comparisons were included in the systematic review and evidence synthesis.

Given all the above considerations the committee agreed that studies comparing acupuncture as an adjunct to usual care with usual care alone were the most appropriate to use in the economic evaluation of acupuncture for the guideline.

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 acupuncture compared to no acupuncture were calculated based on data from randomised controlled studies identified by the systematic review of the clinical evidence that reported quality of life (QoL) or measures that could be mapped to QoL.

The clinical evidence showed that acupuncture reduced pain and improved quality of life. Mortality is not impacted by treatment. The differences in QALYs between acupuncture and no acupuncture 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), the SF-36 that could be mapped to utilities, or pain scales that could be mapped to utilities (see section 2.3.2.1 for more detail). Note that the approach used here was different to the exercise model, whereby only utilities or SF-36 data that could be mapped to utilities were used (and no mapping from pain), as there was considered to be a sufficient amount of QoL data to use in the exercise analysis. The available data on the difference in utility between acupuncture and no acupuncture 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 acupuncture compared to no acupuncture. 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 acupuncture 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. No studies in the clinical review reported use of healthcare services. The two included economic evaluations also reported other resource use. The UK study showed an increase in healthcare costs in the acupuncture group (particularly due to more practice nurse appointments, outpatient visits, A&E admissions, and day case admissions).[14] Although these differences in resource use were not statistically significant (either individually for each resource or for the healthcare costs overall), this led to an overall cost of healthcare resource use over a year (outside of acupuncture costs) of £558 in the acupuncture group and £484 in the usual care group. The German study founds healthcare costs (other than acupuncture) were numerically slightly lower with acupuncture (2 of 3 categories of cost were slightly lower, and 1 of 3 very slightly higher) but differences were very small and not statistically significant.[32] The committee therefore agreed there remains uncertainty particularly about whether any change in resource use is related to chronic primary pain, and (on the available data) whether acupuncture increases or reduces resource use. Due to this uncertainty, no costs other than the cost of acupuncture itself have been included in the model, as this would have required assumptions in one direction or the other as to whether acupuncture 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 acupuncture in a chronic primary pain population.

Pooling acupuncture studies

It was acknowledged that the intervention was delivered differently in different studies and this 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. 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 acupuncture is likely to be cost effective to the NHS based on the currently available evidence. However, if acupuncture is found to be cost effective, uncertainties will remain due to the heterogeneity in the underlying evidence base, and assumptions about effect beyond the trials. All these considerations should be taken into account when interpreting the results of the analysis.

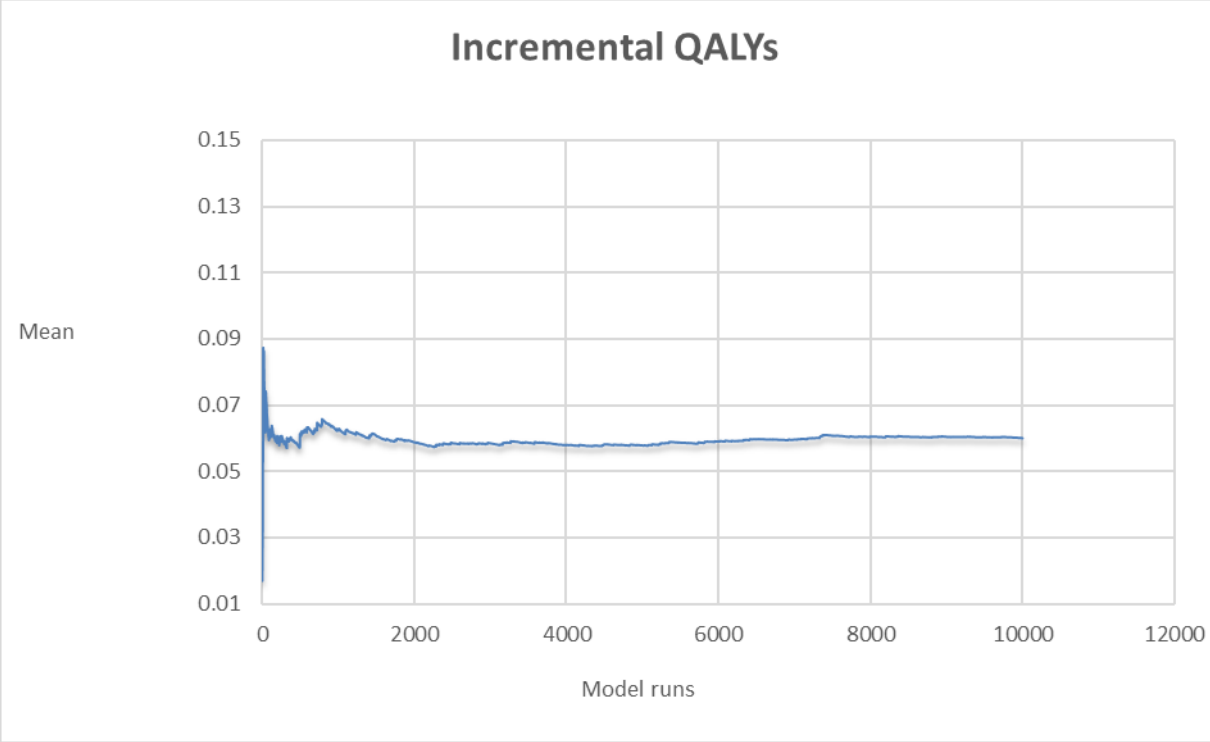
2.2.11 Uncertainty

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 acupuncture 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 acupuncture versus no acupuncture. 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.

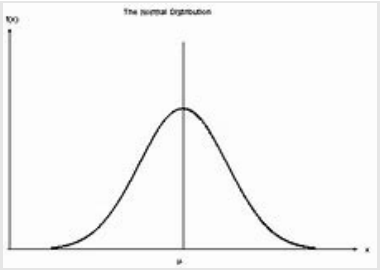
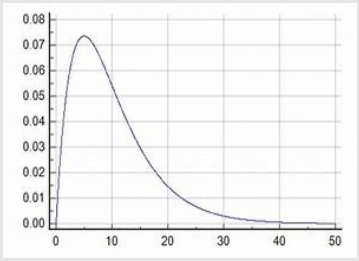
Figure 1: Convergence of incremental QALYs (lifetime analysis)



The way in which distributions are defined reflects the nature of the data. All 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 acupuncture and no acupuncture groups	Normal	The normal distribution is symmetric. Derived from mean and its standard error.

Parameter	Type of distribution	Properties of distribution
		
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 acupuncture 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, or pain scales that could be mapped to EQ-5D. Where a study reported more than one type of outcome, then the following hierarchy was used: EQ-5D, then mapped SF36, then mapped pain. The basis for this being that direct measurement of utilities was preferred over mapped

measures, and where mapping was the only option then mapping SF-36 was preferred over mapping pain, as the SF-36 is more well established and more widely used.

32 clinical studies were included in the acupuncture review in total. Studies comparing acupuncture with usual care were used for the economic analysis (the rationale for this is discussed in Section 2.1.1 Comparators). 9 of the included studies were usual care comparisons. Of these 9: 1 study reported EQ-5D; 2 studies reported SF-36 in enough detail to be mapped to the EQ-5D-3L (no studies reported SF-36 that could not be mapped); and 4 studies reported pain scales that can be mapped to the EQ-5D-3L.

The remaining two studies out of the 9 could not be used in the model because: one study used a composite pain outcome (visual analogue scale (VAS), 15 pain descriptors, and a 1 to 5 present pain intensity scale, together giving a total pain score) whereas it is only the VAS that can be mapped to the EQ-5D, and the other study only reported a discontinuation outcome and no effectiveness outcomes.

The seven studies are summarised in Table 2. One study was a three-arm trial (Cho 2014). In this study the two active acupuncture arms were combined to create a single pairwise comparison, as suggested in the Cochrane Handbook [11](see Appendix C: for how these were combined). Note that two studies are those that already have economic evaluations based on them that were included in this acupuncture review: Essex 2017, [14 14 13 8] and Witt 2006 (used in the Willich 2006 economic evaluation).[32]

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 time point beyond when the intervention had ended (e.g. a 12 week intervention following up patients at 24 weeks).

Table 2: Clinical studies overview

Study	Population	Duration of pain	Level of pain	Measure	Acupuncture type	No of sessions (a)	Intervention length (weeks)	Intervention intensity detail	Follow-up detail	Control arm detail	Number of participants
Witt 2006[33] (b)	Chronic neck pain	6 years	Neck pain and disability scale = 54-55	SF-36	NR	10.2 (mean)	12	NR	Post-intervention outcome at 12 weeks, and follow-up at 24 weeks	Routine care. Allowed any treatment they needed.	3451
Casanueva 2014[5]	Fibromyalgia	NR	Baseline pain VAS =7.8	SF-36	Dry needling	6	6	1 hour sessions	Post-intervention outcome at 6 weeks, and follow-up at 12 weeks	Taking same medical treatment they received before randomization	120
Essex 2017[14] (b) (c)	Chronic neck pain	60-96 months	Northwick park questionnaire = 38%	EQ-5D	Traditional	10 (mean)	20	Offered weekly then fortnightly. 50 min sessions	Follow-up at 24 weeks and 52 weeks	GP care as usual	204
Birch 1995[3]	Chronic myofascial pain	86 months	Baseline pain VAS = 4.8	Looks like NRS rather than VAS (d)	Japanese (shallow needles)	14	10	Twice a week for 4 weeks, once a week for 4 weeks, then every other week for two weeks 30 min sessions	NA – post intervention only	Medication only control. 500mg per day Trilisate.	30
Cho 2014[8]	Chronic neck pain	NR	Baseline pain VAS = 6-6.9	VAS	Traditional	9	3	3 sessions per week Session length NR	Partway through intervention at 1 week, post intervention at 3 weeks, and follow-up at 7 weeks.	NSAID (zaltoprofen, 80 mg daily)	45 (e)

Coan 1981[10]	Neck/hand /arm pain	7-8 years	Baseline pain VAS = 5-6	VAS	Traditional	10.9 (mean)	NR (f)	3 to 4 times per week. Session length NR	Follow-up only at 12 weeks	Usual care. Wait list control	30
Schlaeger 2015[28]	Vulvodynia	5 years	Short form McGill VAS = 5.6	VAS	Traditional	10	5	2 times a week. 30 min sessions	NA – post intervention only	Usual care not further defined. Wait list control.	36

Scales: Where the VAS has been reported this is on a 0-10 scale. The neck pain and disability questionnaire has a scale of 0-100. The Northwick park questionnaire has 2 or 36 questions and scores are expressed as a percentage. The short form McGill questionnaire VAS has a scale of 0-100.

(a) Note that where the mean number of sessions were reported then this has been used in the analysis, rather than the number of sessions that the intervention intended to deliver.

(b) These studies have accompanying economic evaluations for this review (Essex 2017 (same study as in the table above), and Willich 2006.[32] Note that Witt 2006 has SF-36 outcomes and reports these at 12 weeks and 24 weeks. Although only the 12 weeks data was used in the published economic evaluation.[32] Also, the aforementioned economic evaluation also maps the SF-36 data to the SF-6D utility measure to calculate QALYs, whereas here this is being mapped to the EQ-5D.

(c) This is the data from the complete case analysis, not the imputed analysis. This is a limitation. The raw EQ-5D data for the imputed analysis is not reported and has been requested from the authors. Note also this was a 5-month intervention, but the first outcome measurement timepoints is at 6 months, so this has been labelled as post intervention.

(d) This is being treated as a VAS study for the mapping. Both are on a 0-10 scale.

(e) Note this study had 3 arms but 2 arms were acupuncture alone and acupuncture with NSAIDs (the third being NSAIDs alone), and these have been combined using Cochrane methodology.

(f) The intervention must be around 4 weeks long, as the follow-up was at 12 weeks, and the narrative says the follow-up was on average 8 weeks after the treatment was completed.

2.3.2 Calculating the difference in QALYs

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

Most of the studies measured outcomes 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, data was extracted for all time points at which the relevant 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 acupuncture over time after the intervention had ended.

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

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

The raw data extracted from these studies is included in Appendix A:.

2.3.2.2 Mapping to EQ-5D

2.3.2.2.1 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 SDs were reported, the confidence intervals were calculated in Revman software using: the number of participants analysed 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.[1] This is a well-established mapping study. However, to account for some of the uncertainty in the mapping, a variance adjustment method was used. This is explained in more detail in section 2.3.2.2.3.

More discussion on mapping can be found in the discussion section.

Full details on the data extracted (or calculated) from the studies including the resulting mapped EQ-5D values, can be seen in Appendix A: and Appendix B:.

2.3.2.2.2 Mapping pain to EQ-5D

For studies that reported pain, the mean scores 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 pain scores were also extracted. Where only SD's were reported, the confidence intervals were calculated in Revman software using: the number of participants analysed in the study; the mean; and the SD.

The pain scores and their confidence intervals were mapped onto the EQ-5D-3L (UK tariff) using the regression by Maund 2012.[18] Note that the regression used by Maund was based on a dataset using the VAS on a 0-100 scale. The data used in this acupuncture analysis reported VAS on the 0-10 scale, and therefore these were multiplied by 10 to convert them to the 0-100 scale.

Maund 2012 was a systematic review and cost-effectiveness analysis, that derived QoL needed for the cost utility analysis by creating a regression to map from the visual analogue pain scale to the EQ-5D. The dataset used to generate the regression was the SAPPHIRE trial (2008),[31] which was a trial in a population with rotator cuff disease (N = 200).

The analysis with the largest population was used, which was the analysis using patient-level data reported at 1, 3 and 12 months (n= 491, 295 in the estimation data set (60%), and 196 in the validation data set). The OLS model including the squared VAS interaction term was used. Although other models were also available like a TOBIT model, this did not report the R squared statistic which was needed (see section 2.3.2.2.3 for explanation).

The model goodness of fit was fairly poor, with an R squared of 0.1. Although this implies a poor fit, with the authors stating so themselves, this is the only mapping study identified that maps the VAS scale onto the EQ-5D, that also doesn't include other scales in the same regression. This has also been used for mapping in other cost effectiveness studies, notably a large acupuncture piece of work.[17] To account for this uncertainty in the mapping, represented by the low R squared, a variance adjustment method was used. This is explained in more detail in section 2.3.2.2.3.

Full details on the data extracted (or calculated) from the studies including the resulting mapped EQ-5D values, can be seen in Appendix A: and Appendix B:.

2.3.2.2.43 Adjusting mapping for uncertainty in the regression

Several publications have suggested that there is a problem with underestimation of uncertainty of utilities derived from mapping algorithms.[9],[2],[16] 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.[6]

A high level of unexplained variation was found in the mapping algorithms used for this analysis, that is, relatively low R squared (more so in the pain mapping study). 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[6] 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 the same pain mapping algorithm.[17]

This adjustment factor was applied to the variance of the mapped EQ-5D values for both utilities mapped from the VAS (R squared = 0.1), and utilities mapped from the SF-36 (R squared = 0.59). See Appendix B: 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. One of these studies in particular, which happens to be the EQ-5D study (Essex)[14], showed a continued improvement in QoL at follow-up, whereas other studies showed that QoL gain reduced at follow-up.

It is difficult to explain why QoL in Essex study remained stable, especially given that acupuncture is not an intervention that can be continued by the person themselves once the intervention has ended (unlike exercise). This might be to do with other interventions that people are having after acupuncture has ended. The Essex study also had the largest time interval between its follow-up outcomes, as the first outcome was at 6 months, which was 1 month after the intervention ended, and there was second follow-up at 1 year. Given that the outcome at 1 year was very far from the end of the intervention, and also given that this showed a sustained benefit, which the committee were not confident was clinically plausible, this outcome at 1 year was excluded from the base case and only included in a sensitivity analysis. With regards to the Essex study, it is also important to note that the EQ-5D values reported in the paper are the complete case data. The study also undertook an analysis using imputed data, which led to lower QALYs, but the EQ-5D imputed data was not reported. This was requested from the authors, but no response was received.

See section 2.3.2.4 on how the data was meta-analysed and used in the analysis.

It is also important to note that because the QoL values represent acupuncture treatment effect as the QoL *gain (or loss)* from acupuncture 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 acupuncture group remained stable, or: the acupuncture group had improved QoL, and the usual care group remained stable, or both groups improved similarly leading to small QoL gains from acupuncture. The baseline differences and direction of these QoL changes varied between the studies, as can be seen from Figure 2.

Some studies had very small baseline differences. How baselines were accounted for in the meta-analyses where studies were pooled is discussed in the next section.

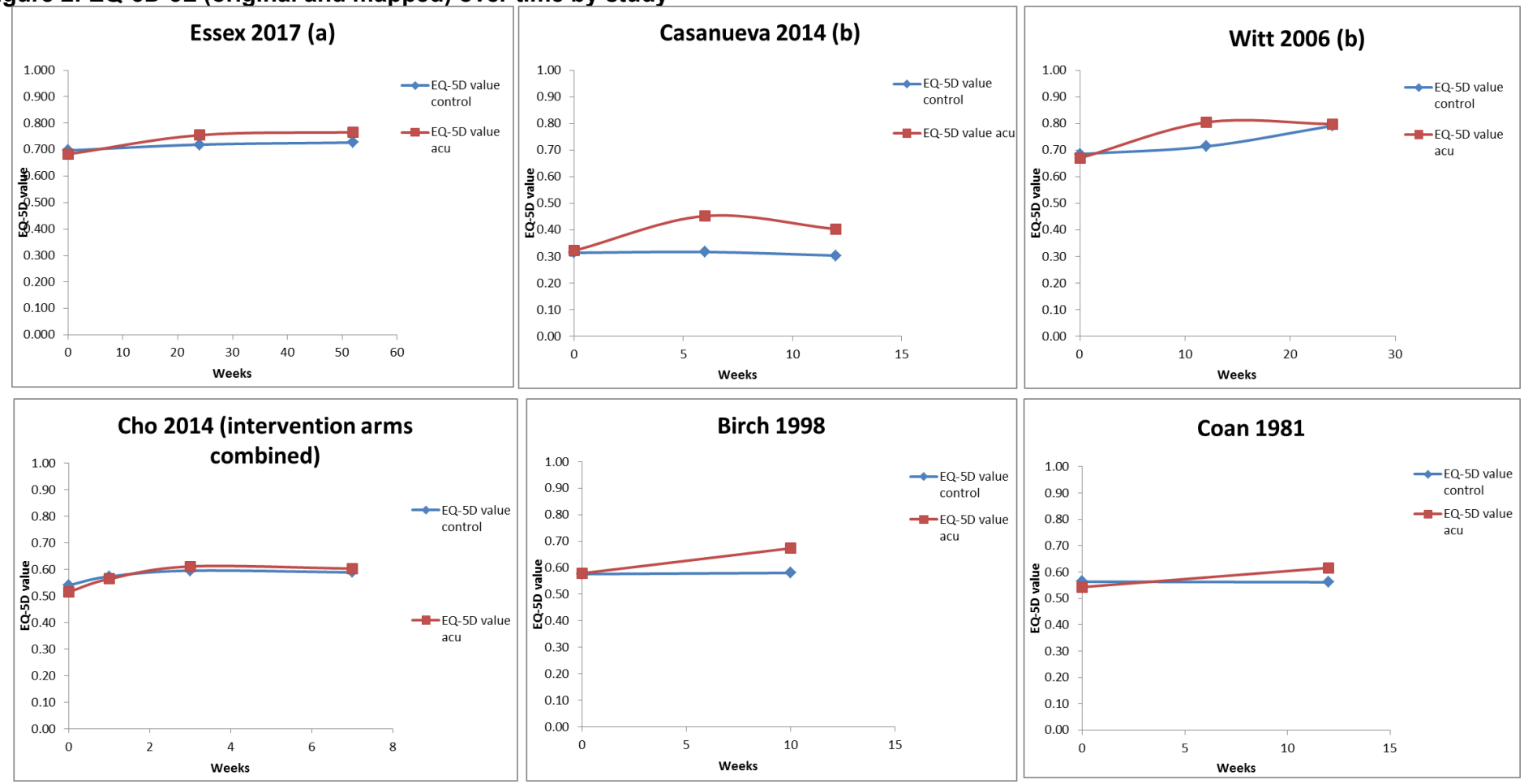
Table 3: EQ-5D-3L (original and mapped) over time by study

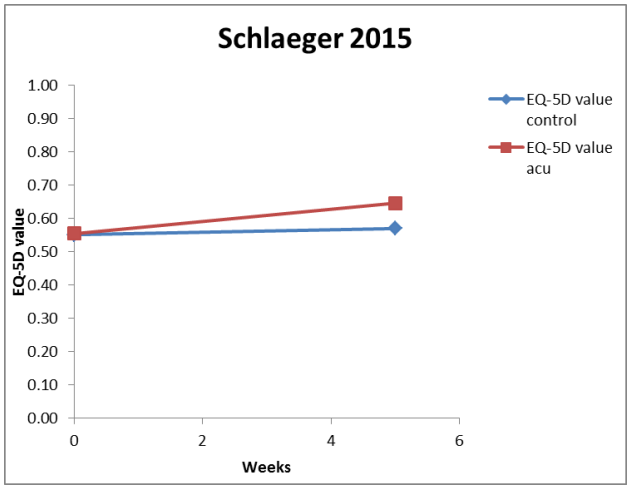
Study	Timeframe (weeks) (a)	EQ-5D value usual care	EQ-5D value acupuncture
Essex 2017 (b)	0	0.697	0.683
	24	0.72	0.76
	52	0.73	0.77
Casanueva 2014 (c)	0	0.31	0.32
	6	0.32	0.45
	12	0.30	0.40
Witt 2006 (c)	0	0.69	0.67
	12	0.71	0.81
	24	0.79	0.80
Cho 2014 (d)	0	0.54	0.51

	1	0.57	0.56
	3	0.60	0.61
	7	0.59	0.60
Birch 1998	0	0.58	0.58
	10	0.58	0.67
Coan 1981	0	0.56	0.54
	12	0.56	0.62
Schlaeger 2015	0	0.55	0.55
	5	0.57	0.65

(a) Timeframe 0 is the baseline.
(b) This study reported EQ-5D-3L data.
(c) These studies reported SF-36 data.
(d) This study had three arms, but the two acupuncture arms have been combined in to a single arm following Cochrane methodology.[11] See Appendix C:

Figure 2: EQ-5D-3L (original and mapped) over time by study





Note: 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. Studies with no footnote are pain studies.

(a) Reported EQ-5D-3L data.

(b) Mapped from SF-36 data.

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 acupuncture together in order to analyse the cost effectiveness of acupuncture versus no acupuncture. As quality of life benefits may change over time it was agreed that pooling should be done by time point. At many time points there was only one data point and so meta-analysis was not required, but where there were multiple data points these were meta analysed.

All studies reported baseline data and final values at one or more other timepoint. Although meta-analysis could be undertaken simply using the final values at each timepoint, it was decided that meta-analysing EQ-5D change scores (i.e. change from baseline in the acupuncture and usual care groups from each study) would be the most precise way of using the data from the trials, capturing any baseline differences between studies. This was also consistent with the approach taken in the exercise modelling undertaken as part of this guideline development. At timepoints where there was only one data point and meta-analysis was not undertaken, change scores were also calculated.

Standard deviations of the means are needed to undertake the meta-analysis. As most of the data was mapped from pain or SF-36 to EQ-5D, then the uncertainty around these mapped values was in the form of confidence intervals (as the pain or 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 analysed in each arm. More detail can be found below on how the standard deviations around change from baseline scores was calculated.

Calculating standard deviations of change scores

As described above, to capture any baseline differences between studies, it was decided that meta-analysing EQ-5D change scores (i.e. change from baseline in the acupuncture and control groups from each study) would be a more precise way of using the data from the trials. However, all the trials 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[11] 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

Correlation coefficients lie between –1 and 1. Cochrane methodology[11] 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.

As no study was available that reported both EQ-5D change from baseline standard deviations as well as baseline and final value standard deviations, then the correlation coefficient was assumed to conservatively be 0.5. This assumption has been used elsewhere in the literature.[21],[26] As Table 3 did not show any large differences in baselines between the groups of any of the studies, then this estimate seemed appropriate. Note that in the exercise analysis, a sensitivity analysis using treatment effects based on a meta-analysis of final QoL values was tested, however there were much larger baseline differences there, and so the impact of the different meta-analyses might be higher, however in this model as baseline differences are not too concerning then this sensitivity analysis was not felt necessary. However, a sensitivity analysis was undertaken varying the correlation coefficient to a higher value of 0.7 to see the impact of this. See the sensitivity analysis section for more explanation on this.

The equation showing how standard deviations were imputed using this correlation coefficient is shown below. 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.

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

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

Once the change from baseline SD's could be calculated, then data was in a form that could be meta-analysed in RevMan. More detail on deciding how to pool the data together in a meta-analysis is discussed in the next section.

2.3.2.5 Using the EQ-5D data in the economic analysis

In the economic analysis, the EQ-5D data from different time points (meta-analysed if there was more than one measurement at a particular time point) were used to estimate QALY gain with acupuncture.

Looking at the pattern of the QoL improvement from acupuncture over time plotted graphically showed that there was an increasing QoL trend up to 12 weeks in the data, and then a somewhat decreasing trend. It was also noted that in studies that measured QoL at the end of the intervention and then again at a later follow-up point, the QoL gain at the follow point was lower. It was agreed that two linear trend lines should be used to represent this pattern over time in the economic analysis.

One trend line was estimated using all observed data points up to and including 12 weeks. A second trend line was estimated based on follow-up datapoints only, where time point zero was the end of the intervention. The linear trend lines were generated using weighted least squares regression to apply a higher weight to the treatment effect from timepoints that had smaller variance.

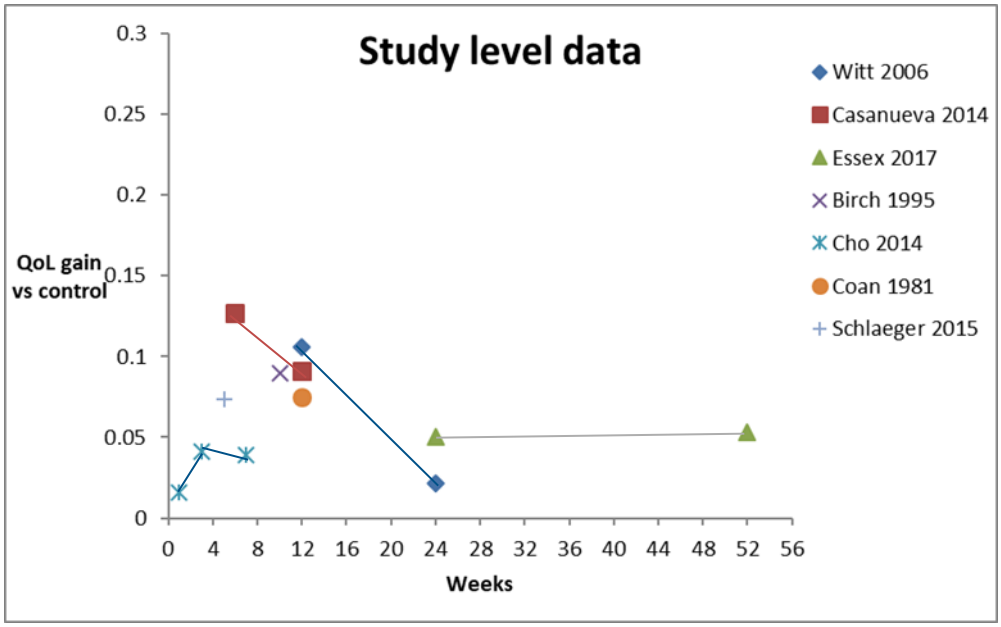
In the economic analysis QOL gain over time was initially modelled using the ≤12 weeks trend line. A linear increase in EQ-5D from zero difference at time zero to the point estimated by the trend line at the first trial observation was also assumed. After 12 weeks the slope from the follow-up analysis trend line was applied up to 24 four weeks (12 weeks follow-up data) in the base case. Analyses were included with and without further extrapolation of treatment effect beyond this point. QALY gain with acupuncture was estimated by calculating the area under the curve.

More detailed information about how the data was analysed and used in the model can be found below.

2.3.2.5.1 Pooling the data and generating trend lines

The committee considered how best to pool the data together to generate a picture of the treatment benefit from acupuncture over time. As mentioned previously, it was felt appropriate to use the outcomes at the time points that they were being measured. Because acupuncture is unlike exercise in the sense that it is not an intervention that can be continued by the person on their own after their course of treatment ends, the committee were very aware of not wanting to overestimate the long-term treatment benefit. A graph of the study level data can be found below in Figure 3.

Figure 3: Study level EQ-5D gain from acupuncture versus usual care



Additional detail about the studies can be seen in Table 4, including a breakdown of the outcome time points of each study, colour coding to show what these represented in terms of whether they were: during; post intervention; or follow-up outcomes. The length of the intervention is also reported to provide information on how long follow was after the intervention ended.

Table 4: Study time point information (all data)

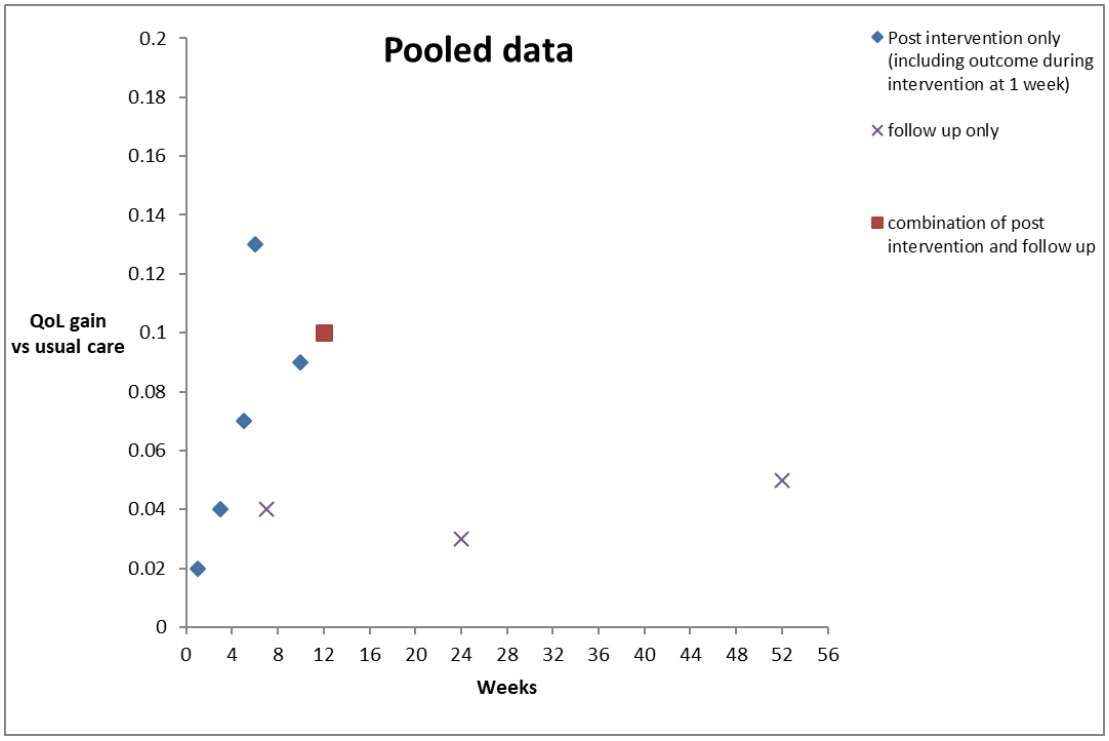
	Time point (weeks from beginning of intervention)										
Study	**1**	**3**	**5**	**6**	**7**	**10**	**12**	**24**	**52**	**N (a)**	**Intervention length**
Cho 2014										*30*	3 weeks
Schlaeger 2015										*18*	5 weeks
Witt 2006										*1753*	12 weeks
Coan 1981										*15*	4 weeks
Casanueva 2014										*60*	6 weeks
Birch 1995										*15*	10 weeks
Essex 2017										*104*	20 weeks

Colours: Blue = part way through intervention, Green = post intervention, Pink = follow-up.

(a) The number of participants is the number in the intervention arm only from each study, as that is the N of interest for the weighted average resource use.

Data that reported outcomes at the same time period could be meta-analysed. A graphical representation of treatment effect over time when including all data can be seen in Figure 4. Time points that had multiple studies that could be meta-analysed have been highlighted in the footnote, and it is also highlighted on the graph whether points were follow-up only, post intervention only, or a combination (where they were meta-analysed).

Figure 4: Treatment effect over time (all data)



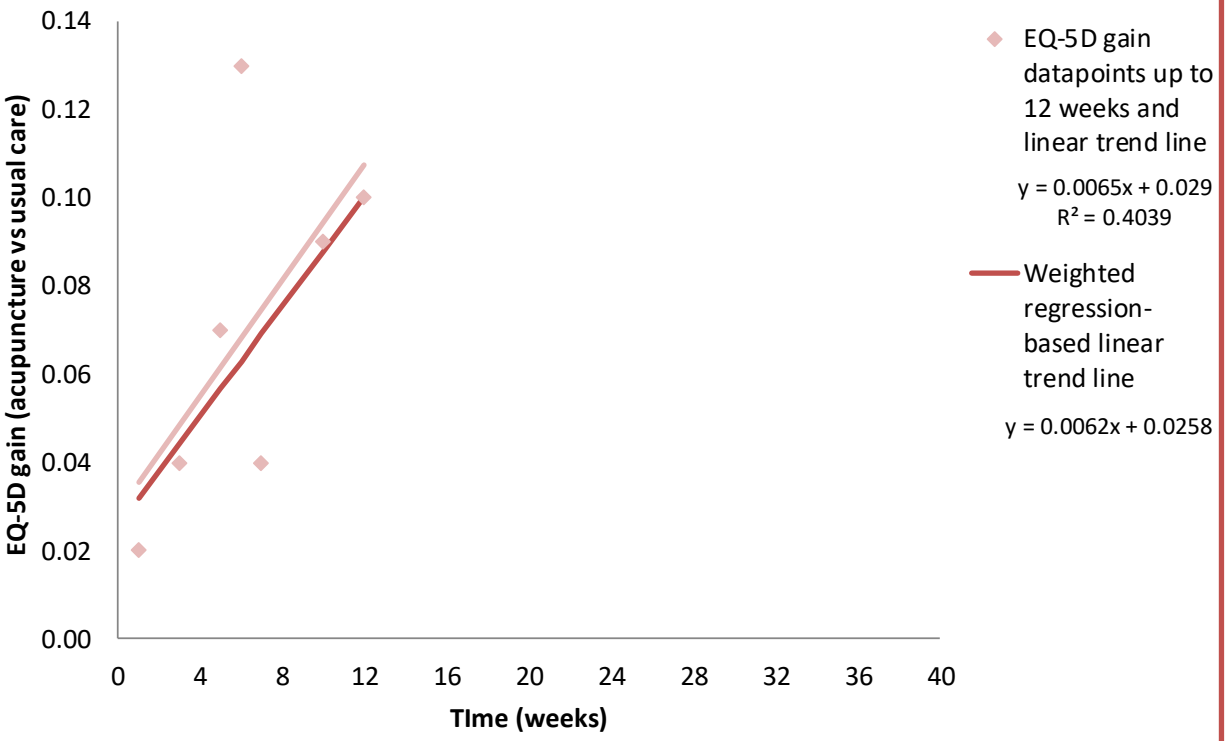
Note: Time points where there was more than one study and therefore a meta-analysis was undertaken was at 12 weeks (there were three studies here and one was a post-intervention outcome and two were follow-up outcomes), and at 24 weeks (where there were two studies and they were both follow-up outcomes).

The committee discussed what all the data over time showed, and how to use it in the economic analysis. A trend could be seen in the early part of the graph that showed QoL gain from acupuncture initially increasing over time, and then reducing later on. Looking at the individual study data in Figure 3 it can also be seen that QOL gains at follow-up were generally lower than when on treatment. Fitting a single linear trend line did not fit the data well for this reason and two linear trend lines were used to model the pattern of the QoL over time representing QOL while on treatment and QOL after treatment (see section further down on other non-linear trend lines that were considered).

As the maximum timeframe that a post-intervention outcome was available was 12 weeks, the committee decided that they would use all the data available up to 12 weeks (pooling at timepoints where there was more than study), regardless of whether these were post intervention or follow-up outcomes. Including any follow-up outcomes that fell in this timeframe was also a more conservative approach towards acupuncture than using only the post-intervention outcomes, as follow-up QoL tended to be lower. Using all the data up to 12 weeks in this way is analogous to the data providing information on the average treatment effect over time during the period of the intervention. Figure 4 shows the data points up to 12 weeks and the associated linear trend line. It also shows a trend line based on a weighted regression that attaches more importance to data points that had greater certainty thus better

taking into account uncertainty in the treatment effect data points. As could be seen in Figure 4, there was an increasing trend up to 12 weeks and**Error! Reference source not found.** Table 6 shows a summary of the <12 week data meta-analysed at each time point.

Figure 5: ≤12 week QoL gain data



As there were no studies that reported outcomes immediately after an intervention longer than 12 weeks, then the treatment benefit beyond 12 weeks is what might be considered the follow-up treatment effect. One method of capturing how treatment effect from acupuncture is likely to diminish after the intervention ends over time, is to plot only the follow-up outcomes on a graph but having time zero as the end of the intervention. Table 4 showed which outcomes were follow-up outcomes from each study, and therefore the difference in time between the end of the intervention and when follow-up outcomes were measured, are the timeframes of interest here. These can also be seen in Table 5. Note only 5 out of the 7 studies included reported follow-up outcomes.

Table 5: Study time point information (all data)

Study	Intervention length	Follow-up measurement time 1	Follow-up measurement time 2	Time from post intervention to follow-up 1 (a) (b)	Time from post intervention to follow-up 2 (c)
Cho 2014	3 weeks	7 weeks		4 weeks	
Witt 2006	12 weeks	24 weeks		12 weeks	
Coan 1981	4 weeks	12 weeks		8 weeks	
Casanueva 2014	6 weeks	12 weeks		6 weeks	
Essex 2017	20 weeks	24 weeks	52 weeks	4 weeks	32 weeks

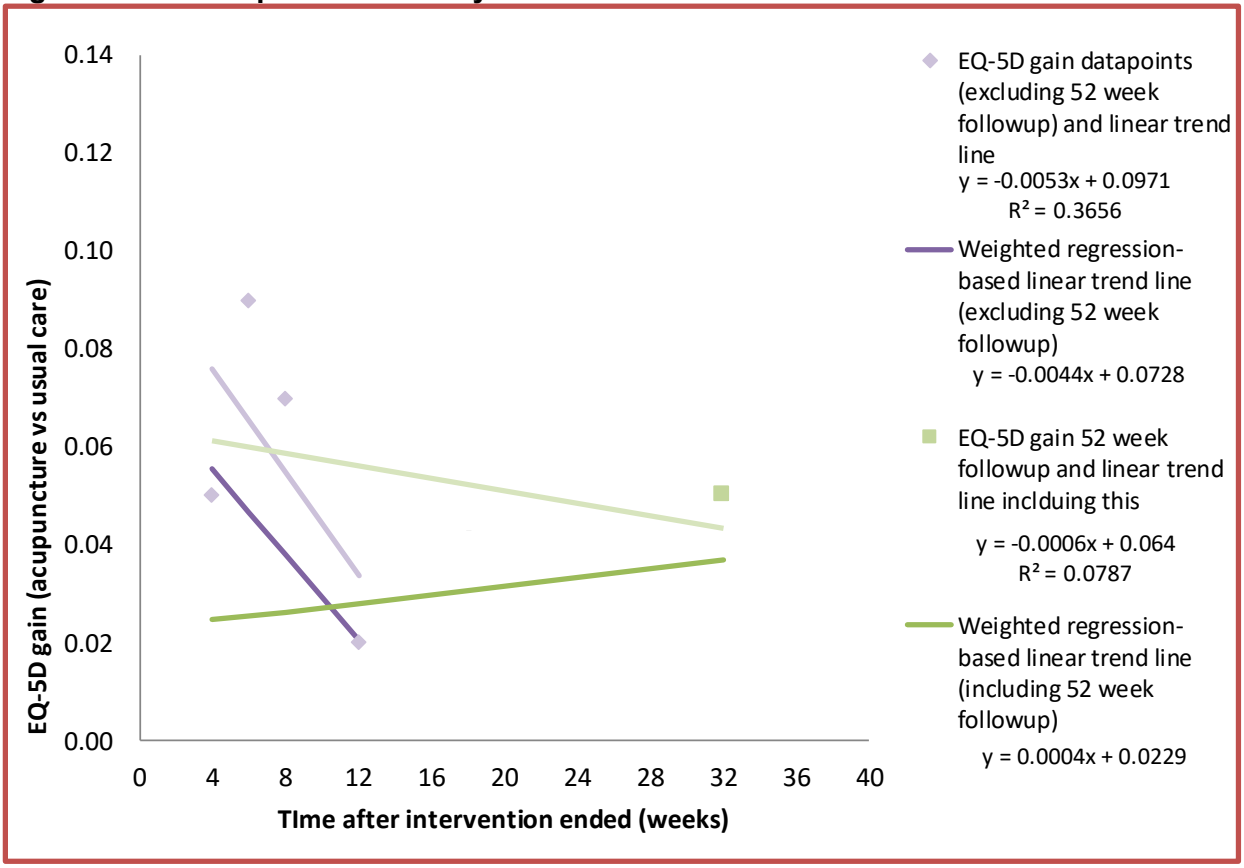
(a) Follow-up measurement time 1 minus intervention length.
(b) The two studies that had a follow-up time post intervention of 4 weeks could be pooled.
(c) Follow-up measurement time 2 minus intervention length.

Figure 6 shows the follow-up outcomes graphically with EQ-5D gain (acupuncture versus usual care) plotted against time after the end of the intervention. The associated linear trend lines with and without the 52 week outcome from Essex are shown, as are trend lines based on weighted regressions that attaches more importance to data points that had greater certainty thus better taking into account uncertainty in the treatment effect data points (note that the intervention length was 5 months (20 weeks) in this study and so the 52 week outcome is 32 weeks post intervention).

This approach is supported by a published systematic review that looked at the association of factors of acupuncture treatment schedule and pain relief and found that the longer the follow-up, the smaller the improvement in pain (which would correspond to higher pain and therefore lower QoL).[7] Vickers 2018[30] also found that there was a reduction in effect size from acupuncture the longer the time since treatment.

The summary of treatment effect from the follow-up data is also summarised in Table 7.

Figure 6: Follow-up outcomes only



Note: time zero here is the end of the intervention. Also note that the 52 week outcome from Essex is follow-up after a 20 week outcome and is therefore the 32 week point on the graph.

The committee noted that the weighted regression trend line when the Essex 52 week datapoint was included resulted in an upward sloping trendline. This was due to relative weighting of the different data points (weights can be seen in Table 8). The committee were not confident that quality of life continuing to improve from a course of acupuncture would be clinically plausible, especially so long after the interventions ended. For this reason, they decided to exclude this long-term outcome from the base case, and to include it in a sensitivity analysis.

Note that the Essex study outcome at 52 weeks, which was only included in a sensitivity analysis, was included in a way that reflects the length of the trial itself (so the follow-up trend line was applied up to 52 weeks). However, as this was actually 32 weeks after the end of the intervention, then an alternative way to apply this would be to apply the follow-up trend

line for only up to 32 weeks after the 12 week trend line. Which would mean the follow-up trend line ends at 44 weeks (as opposed to 52 weeks). This approach was used in an alternative sensitivity analysis. Note that for the base case without the 52 week data point, this alternative method of follow-up time makes no difference because the maximum time from post intervention to follow-up without the 52 week point was 12 weeks, and as the maximum trial length without this outcome was 24 weeks, then both methods would lead to trend lines reflecting trial data at a maximum of 24 weeks anyway.

A summary of the meta-analysed data informing each timepoint for the two trend lines can be seen in Table 6 and Table 7. The full data on the EQ-5D changes from baseline and their SD's from each study can be seen in Appendix A: and Appendix B:. The treatment effect reported here is the mean difference in changes from baseline QoL, between acupuncture and no acupuncture groups.

See Section 2.3.2.5.2 'Resulting base case treatment effect over time in economic analysis and extrapolation beyond the trial data' for details of how the ≤12 week trend line and >12 week trend lines are used together in the economic analysis.

Table 6: EQ-5D mean difference between acupuncture and no acupuncture (up to 12 weeks)

Weeks (time zero being beginning of trial)	1	3	5	6	7	10	12
Base case - all data up to 12 weeks							
Pooled QoL difference	0.02	0.04	0.07	0.13	0.04	0.09	0.1
Uncertainty	-0.09 to 0.12	-0.07 to 0.15	-0.08 to 0.23	-0.01 to 0.27	-0.09 to 0.16	-0.06 to 0.24	0.09 to 0.12
No. studies informing timepoint outcome (a)	1	1	1	1	1	1	3

(a) 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 acupuncture) could be obtained.

Table 7: EQ-5D mean difference between acupuncture and no acupuncture (beyond 12 weeks)

Weeks (time zero being end of intervention)	4	6	8	12	32
Base case – follow-up data (excluding 52 wk outcome)					
Pooled QoL difference	0.05	0.09	0.07	0.02	
Uncertainty	0 to 0.1	-0.05 to 0.24	-0.09 to 0.24	0 to 0.04	
No. studies informing timepoint outcomes (b)	2	1	1	1	
Sensitivity analysis – follow-up data (including 52 wk outcome) (a)					
Pooled QoL difference					0.05
Uncertainty					0 to 0.1
No. studies informing timepoint outcomes (b)					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 acupuncture) could be obtained.

In the probabilistic analysis the QOL difference at each time point was assigned a normal distribution parameterised using the mean estimate and the uncertainty around it. A normal distribution was used as this would not be bounded by zero, and it is possible for there to be

a QoL loss from acupuncture compared to no acupuncture (as well as a QOL gain). The treatment effect (QOL difference) at each time points was varied independently: this means that the slope of the treatment effect lines can change. It was considered whether the QoL changes across time points could be correlated, but as not all the points were from the same study, it was decided to let the uncertainty around QoL estimate for each time point be independent. Therefore, this is a limitation in the model.

Use of linear trend lines in the analysis

As described above the QoL gain from acupuncture over time was modelled using two linear trend lines based on the available data. The first line representing up to 12 weeks and the second after 12 weeks. This was because initially QoL gain increased over time but later on it reduced and it was also noted that in studies that measured QoL at the end of the intervention and then again at a later follow-up point, the QoL gain at the follow point was lower. This was considered to fit with time on treatment and then what happens once treatment has stopped. A trend line gives a smoothed estimate of the treatment effect trend over time. It can also be used to predict the treatment effect for timeframes that go beyond those available.

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 acupuncture 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. A polynomial curve was also considered when taking all the data as a whole, as Figure 4 shows an initially increasing trend and then a decreasing trend. However, a polynomial curve wasn't a good fit because some of the hills and valleys looked like they fitted the data well and some did not. Therefore, it was decided that two linear trend lines were the most appropriate fit and reflection of what was happening to the treatment effect over time.

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, 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 and its reciprocal could be calculated. These are shown below in Table 8.

Table 8: Regression weights

Weeks (time zero being beginning of trial)	1	3	5	6	7	10	12
Base case							
SE	0.05	0.06	0.08	0.07	0.06	0.08	0.01
Variance	0.0029	0.0031	0.0063	0.0051	0.0041	0.0059	0.0001

Weeks (time zero being beginning of trial)	**1**	**3**	**5**	**6**	**7**	**10**	**12**
Inverse of variance (regression weights)	**348.4**	**317.5**	**159.9**	**196.0**	**245.9**	**170.7**	**17073.2**
Weeks (time zero being end of intervention)	**4**	**6**	**8**	**12**	**32**		
Base case – follow-up data (excluding 52 week outcome)							
SE	0.03	0.07	0.08	0.01			
Variance	0.0007	0.0055	0.0071	0.0001			
Inverse of variance (regression weights)	**1536.6**	**182.7**	**141.1**	**9603.6**			
Sensitivity analysis – follow-up data (including 52 week outcome)							
SE					0.03		
Variance					0.0007		
Inverse of variance (regression weights)					**1536.6**		

These weights were not varied in the probabilistic analysis.

2.3.2.5.2 Resulting base case treatment effect over time in economic analysis and extrapolation beyond the trial data

The base case treatment effect over time in the economic analysis can be seen in Figure 7 (analysis with extrapolation) and Figure 8 (analysis without extrapolation). The area under the curve represents the QALY gain.

In both graphs the red line shows the ≤12 week QoL gain with acupuncture based on the weighted regression trend line described above based on the available data up to 12 weeks. 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 assumed – this is shown by the red dashed line. The purple line that starts at 12 weeks, is based on the analysis of follow-up data points described above. It uses the slope of the follow-up data weighted regression trend line applied so that it starts where the ≤12 week line finishes.

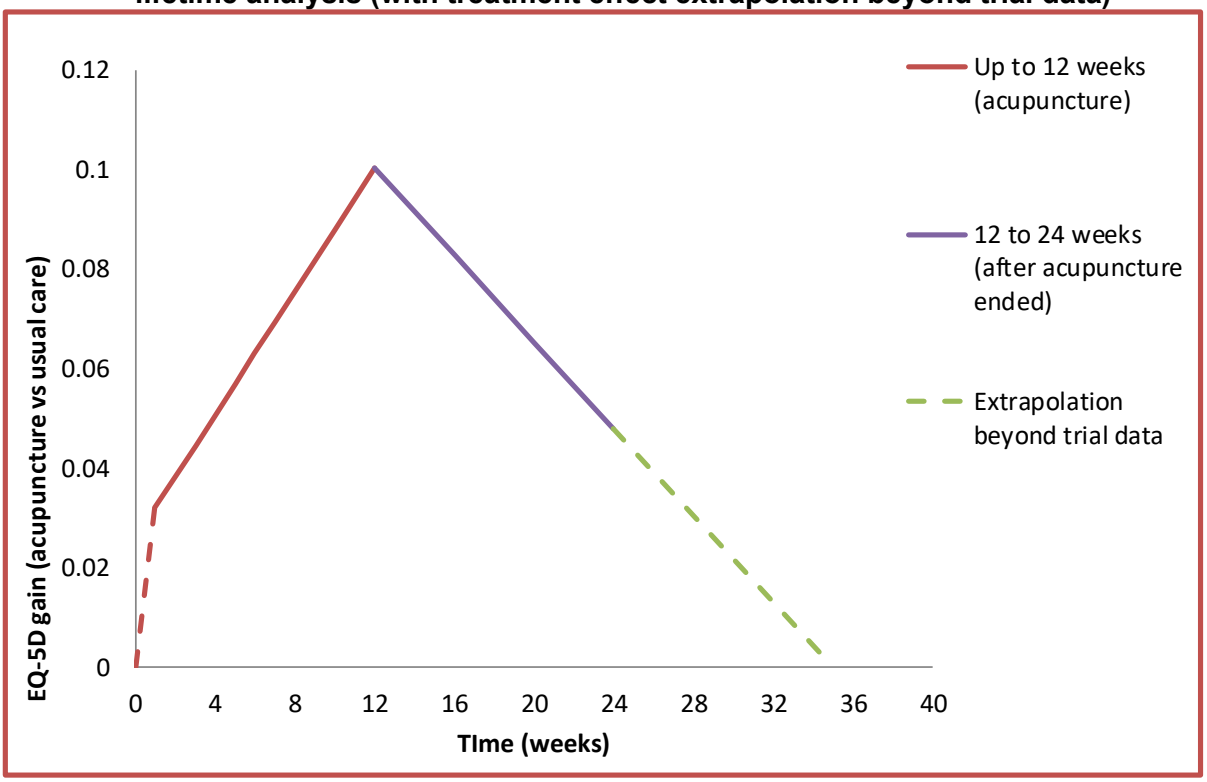
The committee discussed whether they wanted to extrapolate beyond the available data. Any persisting treatment benefit beyond the intervention is assumed to already be partly captured in the treatment effect from the available data, as some of the outcome measurements were at follow-up. The committee discussed how to extrapolate beyond this data.

The committee agreed that benefits beyond the trial data were uncertain but not extrapolating may underestimate benefits and so cost effectiveness. Given this, two base cases were modelled: one where the time horizon of the model was at the end of the trial data (at 24 weeks in the base case; Figure 8), and one where the treatment effect was extrapolated (Figure 7). For the analysis with extrapolation the committee agreed that following the downwards slope of the >12 week trend line that represents the post-intervention treatment effects beyond 12 weeks (in the base case) until there was no difference in QoL with between acupuncture and usual care (that is when the line meets the x axis) seemed reasonable (as shown by green dashed line in Figure 7, as there may be some continuing benefits, even if they reduced.

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.

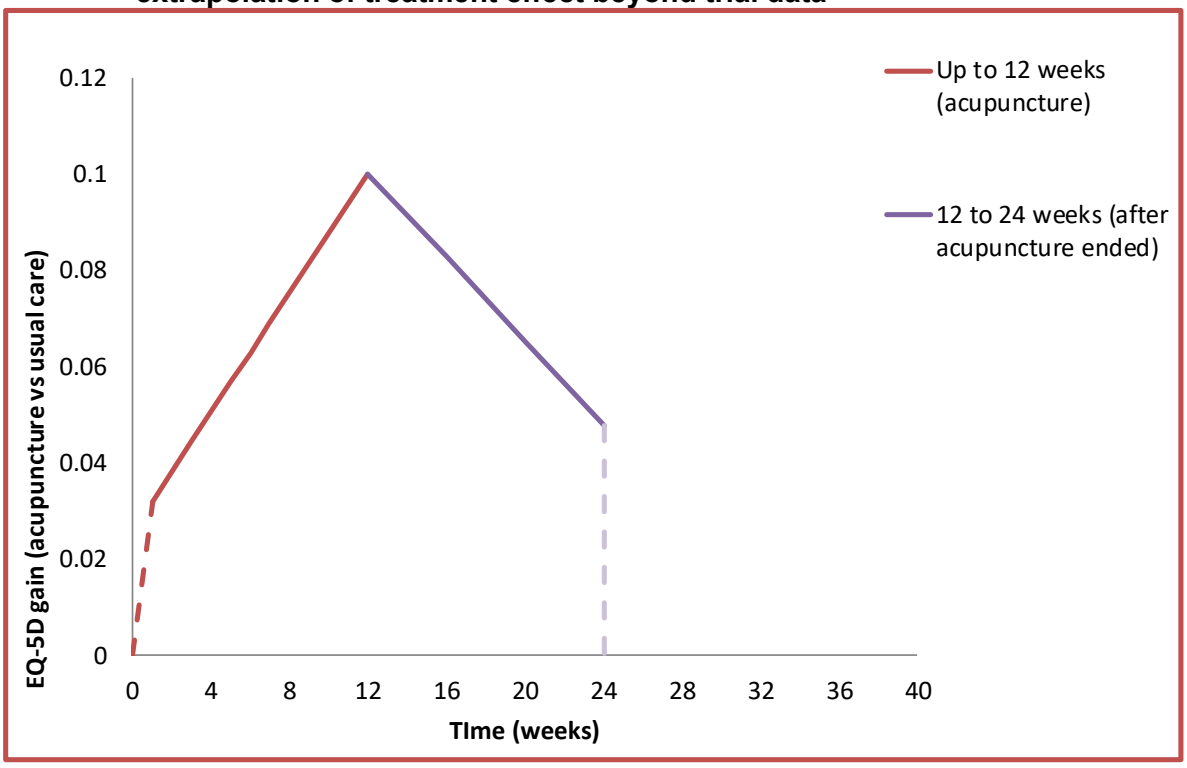
1

Figure 7: QoL difference over time with acupuncture in economic analysis: base case lifetime analysis (with treatment effect extrapolation beyond trial data)



2

Figure 8: QoL difference over time with acupuncture: base case analysis without extrapolation of treatment effect beyond trial data



2.3.2.5.3 Behaviour of the trend lines 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). The uncertainty in the model is large, and each time point is independent. It is therefore feasible that the >12 week trend line 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). Likewise, the <12 week trend line could also be downward sloping.

The committee discussed whether an upward sloping >12 week trend line (representing follow-up treatment effect) would be clinically feasible (i.e. the QoL gain from acupuncture continuing to improve over time after the intervention had ended). It was thought this would be unlikely as people would not be receiving the intervention anymore. And very few people may pay for the intervention themselves. However, the committee acknowledged that the slope of the line changing in simulations is an appropriate reflection of the uncertainty in the data.

To identify the scenarios occurring in probabilistic analysis that needed assumptions, as well as identify their frequency, multiple sets of 10,000 simulations were run. It was identified that some scenarios do not occur at all, and therefore assumptions did not need to be made about them. Scenarios that did occur can be seen in Table 9.

Table 9: Scenarios occurring in probabilistic analyses

	< 12 week line (red)	> 12 week line (purple)
Sloping up		
1. Trend line fully in negative area	X	X
2. Trend line crosses x axis	Yes	X
3. Trend line fully in positive area	Yes	Yes
Sloping down		
4. Trend line fully in negative area	X	X
5. Trend line crosses x axis	X	Yes
6. Trend line fully in positive area	Yes	Yes

The proportion of times that these different scenarios were occurring was monitored to assess the impact on the results by comparing the deterministic and probabilistic results (see results section for discussion on this).

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 means that 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 this can lead to large changes in the slope of the trend line in each simulation of the probabilistic analysis. 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.

Note that it is specifically the behaviour of the >12 week (follow-up) trend line (from Figure 7) that is of interest for extrapolation here, as that is the trend line that will be extrapolated.

Different extrapolation assumptions were needed depending on:

- the slope of the line,
- whether the end of the purple trend line (at 24 weeks in the base case, reflecting the end of the trial data) represented a QoL gain from acupuncture or a loss.

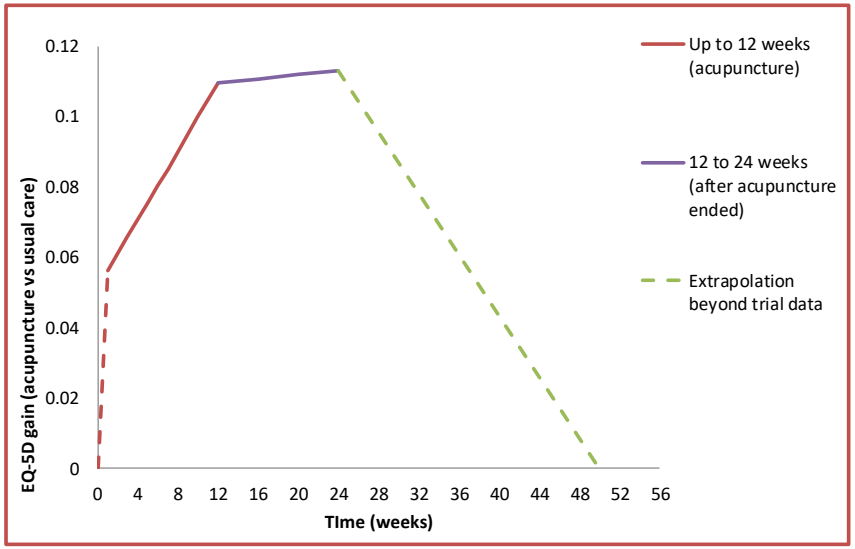
The scenarios that were occurring in the model for the follow-up trend line were shown in Table 9**.** See Figure 9, and below for more explanation on assumptions made for the scenarios occurring.

1. Scenario 3 from Table 9**:** Where the treatment effect could be upward sloping, with a QoL gain from acupuncture, it is thought that improvements from acupuncture 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. The committee decided that a conservative estimate would be that when the treatment effect is upward sloping, it should be extrapolated by assuming that beyond the trial data QOL gain with acupuncture reduces until there is no longer a difference with acupuncture compared to usual care. It was agreed that this reducing treatment effect should be based on the same slope as the base case >12 week trend line (representing follow-up treatment effect) (see Figure 9a).
2. Scenario 5 from Table 9**:** The >12 week treatment effect trend line could be downward sloping and end in the negative part of the graph (which represents a QoL loss from acupuncture). In this case, it was assumed that beyond this point (24 weeks in the base case) the treatment effect line should slope up again until there is no longer a difference in QOL with acupuncture compared to usual care (Figure 9b). The time it would take for the QoL difference from acupuncture to go back to baseline was decided as being the same as the duration of the <12 week trend line (i.e. 12 weeks). Quite a short timeframe was chosen because it was seen as quite unlikely that there would be some adverse impact from acupuncture after the treatment, but there is uncertainty about this. Any adverse impact is more likely to happen at the beginning of the treatment. However, allowing some area of QALY loss would also be more conservative. Note that a sensitivity analysis was done where the trend line stopped at the x axis even if the end of the trend line was below the x axis, as it was discussed whether a QoL loss after treatment was likely (this is discussed more in the sensitivity analyses).

Note that scenario 6 is that of the base case so treatment effect would continue on the same slope until it hits the x axis (no treatment benefit from acupuncture).

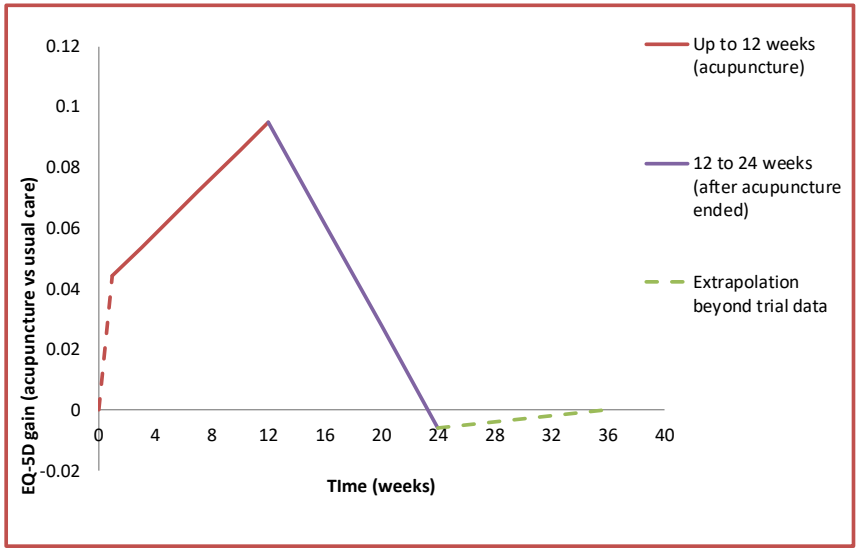
Figure 9: Additional extrapolation assumptions in probabilistic analysis

Figure 9a: >12 week trend line (purple) sloping up with QoL gain



Note: This graph shows an illustrative scenario where the 12-24 weeks line is upward sloping. In these instances in the probabilistic analysis, when treatment effect is extrapolated beyond 24 weeks it is done so by assuming that QoL will decrease over time based on the same slope as the base case >12 week trend line until there is no difference in QoL with acupuncture (that is when the line crosses the x axis).

Figure 9b: >12 week trend line (purple) sloping down and crosses x axis representing some QoL loss



Note: This graph shows an illustrative scenario where the 12-24 weeks line is downward sloping but crosses the x axis representing QoL loss. In these instances in the probabilistic analysis, when treatment effect is extrapolated beyond 24 weeks it is done so by assuming that QoL will increase until there is no difference in QoL with acupuncture (that is when the line crosses the x axis) over a 12 week period.

.

Note that in the probabilistic analysis, the <12 week trend line (representing treatment effect during the intervention period) can also change direction in terms of slope and can also cross the x axis. However as this is the first trend line in the graph, no extrapolation assumptions are needed about this. Where the <12 week trend line starts below the x axis, the area of QALY loss is summed with the overall QALY gain.

As described above, sometimes in the probabilistic analysis the trend lines may be partially in the negative part of the graph which represents QOL loss with acupuncture compared to usual care. It was discussed whether the probabilistic analysis should allow for QoL losses as well as gains but it was agreed that it should because this represents the uncertainty in the data, and because such situations can occur in reality, for example acupuncture making a person's symptoms worse initially before making 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 (24 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.

2.3.2.6 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 acupuncture 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,[29] 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 10 for detail on the population parameters of average age and distribution of gender. These parameters were fixed in the probabilistic analysis. Note that the majority of simulations QOL difference with acupuncture reduced to zero before the age of death in the analysis.

Table 10: Population parameters

Parameter description	Point estimate	Source
Population parameters		
Age	50	Weighted average from the RCTs informing treatment effect.
Gender distribution	Men: 30% Women: 70%	The distribution of gender across the RCTs informing treatment effect.

RCT: randomised controlled trial.

2.3.3 Calculating the cost of acupuncture

As discussed in section 2.2, the committee agreed that the cost of acupuncture 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 acupuncture.

2.3.3.1 Resource use

The resource use from each study was identified. 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 11.

Table 11: Intervention resource use

Study	Intervention classification	Frequency (per week)	Intervention length (weeks)	No. of sessions	Length of sessions	Total minutes	Total hours	N (a)
Witt 2006	NR	NR	12	10.2 (b)	30 (c)	306	5.1	1753
Casanueva 2014	Dry needling	1	6	6	60	360	6.0	60
Essex 2017	Traditional	1, then 0.5	20	10 (b)	50	500	8.3	104
Birch 1995	Japanese (shallow needles)	1, then 0.5 then 0.3	10	14	30	420	7.0	15
Cho 2014	Traditional	3	3	9	30 (c)	270	4.5	30
Coan 1981	Traditional	3 to 4	NR	10.9 (b)	30 (c)	327	5.5	15
Schlaeger 2015	Traditional	2	5	10	30	300	5.0	18
Straight average				10	37	372	6.2	
Weighted average				10.1	31.9	322	5.4	

(a) These are the number of participants analysed in the intervention arm only
(b) This is the mean number of sessions reported. Not the total that the intervention intended to deliver.
(c) The length of the sessions was not reported in these studies and has been assumed to be 30 minutes.

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

Some information on the average intervention information can also be seen in the table. On average across the studies, the resource use is equivalent to 10 sessions of around 30 minutes.

Some studies did not report the length of the sessions, and this has been assumed to be 30 minutes.

In order to estimates costs, the level and number of staff involved in providing the interventions in the studies were required. The committee agreed that in the base case a band 6 staff member would provide the intervention. Use of other staff bands was also tested in a sensitivity analysis. See the section on sensitivity analyses for more detail on these.

The assumptions made regarding staffing and total costs per study are shown in Table 13.

The approach of costing based on the weighted average of the resource use was used, so that this would be more closely related to the treatment effect. Although there is variability in practice of what an acupuncture course might look like, the committee also came up with an

estimate of what a typical course could be, consisting of 6 sessions of 30 minutes each, which was tested in a sensitivity analysis. Another reason this was only used in a sensitivity analysis was because there is uncertainty about whether fewer sessions would lead to the same treatment effect. This is discussed further in the discussion section.

2.3.3.2 Costs

The costs of different bands of staff used in the analysis are presented in Table 12.

Table 12: Staff costs

Band	Cost per hour	Source
Base case		
6	£64.41	PSSRU 2018[12 a,b,c]
Sensitivity analysis		
5	£51.19	PSSRU 2018[12 a,b,c]
7	£77.53	PSSRU 2018[12 a,b,c]

(a) PSSRU staff costs are based on the mean full-time equivalent annual basic salary for each agenda for change band plus salary oncosts (national insurance and pension), overheads and capital overheads.
(b) Costs include a ratio of direct to indirect time of 1.37 taken from PSSRU 2018[12], section V.20.
(c) Costs include qualification costs, based on a physiotherapist from PSSRU 2018, section V.18.

Unit costs for staff from the PSSRU are based on the mean full-time equivalent annual basic salary for each agenda for change band plus salary oncosts (national insurance and pension), overheads and capital overheads. The cost of staff per hour also included a ratio of direct to indirect time, thereby taking into account not just time with patients, but also time spent doing other things related to patient work such as admin. Qualification costs are also included.

The band of staff that would deliver the intervention was discussed extensively with the committee. Theoretically, a band 5 could also deliver the intervention, but would require a lot of managerial support. More generally it was thought a band 6 or above would be more typical. However, this might be the case because of career structure (e.g. more senior staff looking for a new field to train in) rather than a certain grade being a prerequisite for delivering the intervention. The needling itself is a skill that can come with practice. There are also the contextual effects associated with acupuncture, in terms of the way the clinician interacts with the patient for example, and a higher grade individual might provide more of a contextual effect. After discussing all these points, the committee felt that a band 6 staff member should be used in the base case, and a higher and lower band tested in sensitivity analyses.

The cost of needles was also included. These were taken from the NHS supply chain[25] by finding all acupuncture needle products, and taking an average of the cost per needle across all products. The cost per needle was found to be £0.06.

The number of needles needed per session were discussed with the committee. A large acupuncture individual patient meta-analysis reported the number of needles across studies, and the most frequent range was between 10 and 14 needles.[30] The number used depends on the type of acupuncture, with traditional acupuncture using more. The assumption was made to use 14 needles per session. The cost of the needles is small in comparison to the staff costs.

The estimated intervention cost by study and the overall weighted average intervention cost used in the analysis can be seen in Table 13. A weighted average cost was calculated by weighting the cost from each study by the number of participants for whom outcomes were reported in the intervention arm.

Table 13: Intervention cost

		Assumptions					
Study	**Total hours**	**Band of staff member**	**Overlap in treatment (number of people can be seen per session)**	**Supervised cost per patient**	**Additional resource use (needles)**	**Total cost**	**N**
Witt 2006	5.1	6	1	£328	£9	£337	1753
Casanueva 2014	6	6	1	£386	£5	£391	60
Essex 2017	8.3	6	1	£537	£8	£545	104
Birch 1995	7	6	1	£451	£12	£463	15
Cho 2014	4.5	6	1	£290	£8	£297	30
Coan 1981	5.5	6	1	£351	£9	£360	15
Schlaeger 2015	5	6	1	£322	£8	£330	18
WEIGHTED AVERAGE COST						**£350**	

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 acupuncture 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 (£87). 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 (£33) 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 by 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 14. 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). This includes all data (including the outcome at 52 weeks which is not included in the base case). The committee noted that it was not clear that higher cost interventions had higher QoL gain and did not feel they could draw conclusions about the correlation between intensity and QoL gain. There are other variables to take into account such as the type of acupuncture, 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 14: Treatment effects and corresponding costs (all data)

	Time point (weeks from beginning of intervention)										
Study	**1**	**3**	**5**	**6**	**7**	**10**	**12**	**24**	**52**	**N (a)**	**Cost**
		EQ-5D gain									
Cho 2014	0.016	0.041			0.039					*30*	£297
Schlaeger 2015			0.073							*18*	£330
Witt 2006							0.106	0.022		*1753*	£337

Coan 1981							0.075			*15*	£360
Casanueva 2014				0.127			0.091			*60*	£391
Birch 1995						0.090				*15*	£463
Essex 2017								0.050	0.053	*104*	£545

Colours: Blue = part way through intervention, Green = post intervention, Pink = follow-up.
(a) The number of participants is the number in the intervention arm only from each study, as that is the N of interest for the weighted average resource use.

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 acupuncture compared to no acupuncture (taking into account baseline differences) was the treatment effect. This was based on studies in the clinical review that reported EQ-5D utilities or measures that could be mapped to EQ-5D like SF-36 and the pain scales. QoL differences were based on a meta-analysis of change from baseline scores from the acupuncture group compared to the no acupuncture 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 acupuncture (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 9).

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 analysed 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 acupuncture.

2.5 Sensitivity analyses

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

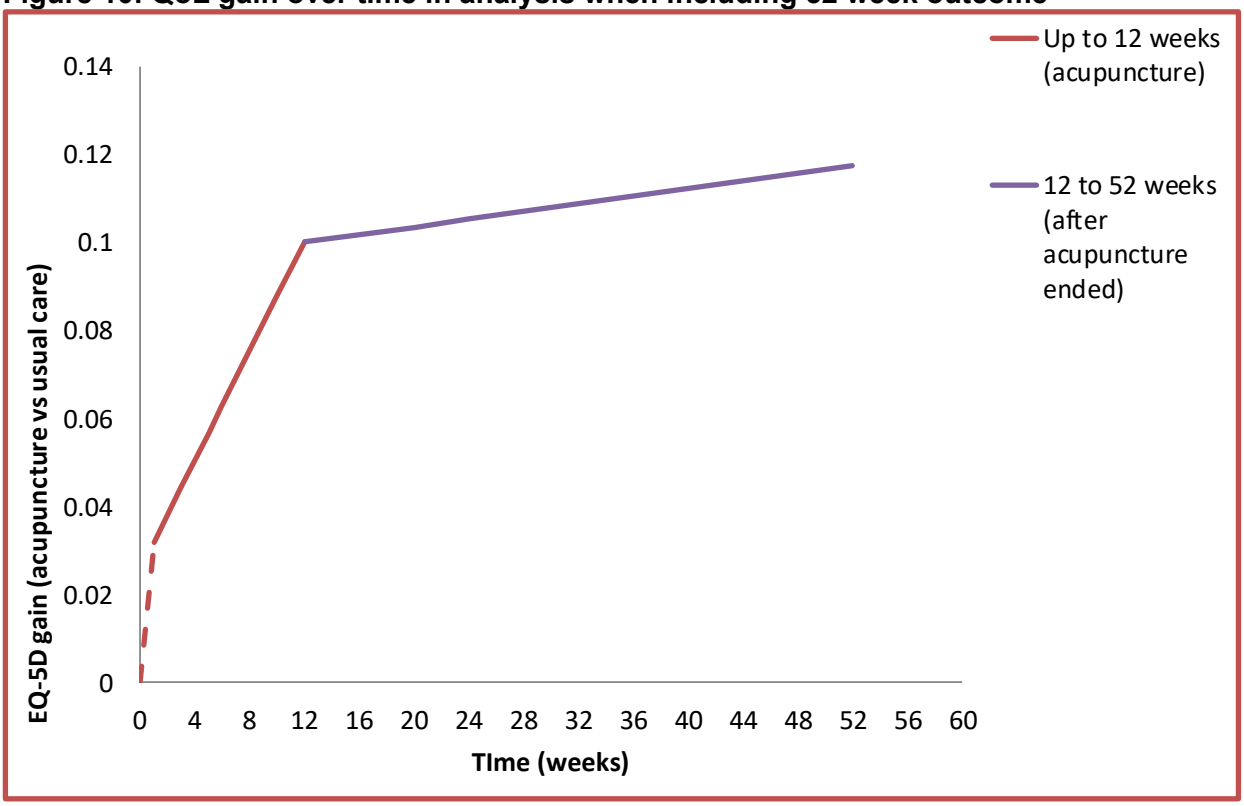
All sensitivity analyses were undertaken for both base cases (extrapolation beyond 24 weeks and truncation at 24 weeks), unless otherwise stated.

2.5.11 SA1: Including 52 week outcome from Essex 2017 (applied at 52 weeks as in the trial)

In the base case analysis, the long term outcome from Essex 2017 were excluded as the follow-up was much longer after the interventions ended compared to other studies, and also QoL continued to improve at this follow-up point which the committee thought was unlikely to be feasible. In a sensitivity analysis this was included and informed the follow-up trend line. In the analysis without extrapolation beyond the trial data, the difference in QoL now end at 52 weeks (which is when follow-up happened in the trial). In the analysis with extrapolation, QoL gain is extrapolated after 52 weeks as was done in the base case probabilistic analysis when there was an upward sloping trend line for the >12 week data (see Figure 9a and accompanying explanation above); it was assumed that QoL will decrease over time based on the same slope as the base case >12 week trend line until there is no difference in QoL with acupuncture (that is when the line crosses the x axis).

The follow-up trend line over time with this data point included can be seen in Figure 10. The weights used in the regression have led to an upward sloping trend line because Table 8 shows that the weight given to the outcome at 52 weeks is quite high, most likely because it has a small SE. This will lead to an area under the curve that will generate higher QALYs than the base case.

Figure 10: QoL gain over time in analysis when including 52 week outcome



Notes: Up to 12 weeks line is based on weighted regression of trial data points ≤12 weeks. 12 to 52 weeks line is based on the slope from the weighted regression trend line from the analysis of follow-up data points applied so that it starts where the ≤12 weeks line finishes. When treatment effect is extrapolated beyond 52 weeks (not shown) it is done so by assuming that QoL will decrease over time based on the same slope as the base case >12 week trend line until there is no difference in QoL with acupuncture (that is when the line crosses the x axis).

2.5.2 SA2: Including 52 week outcome from Essex 2017 (applied at 32 weeks post 12 week trend line)

As discussed in section 2.3.2.5, when including the 52 week outcome from Essex 2017: instead of applying this outcome based on the duration of the trial, which would give the follow-up trend line a maximum duration of 52 weeks, it could be applied on top of the <12 week trend line based on the time between the end of the intervention and when the follow-up outcome was measured (32 weeks). Applying it in this way means that the follow-up trend line would have a maximum duration of 44 weeks (12 + 32). This would mean that in this sensitivity analysis, 44 weeks would be the end of the time horizon in the base case where treatment effect is not extrapolated.

This generates a follow-up trend line with exactly the same slope as in Figure 10, but just ends at 44 weeks (rather than 52 weeks). This will therefore generate fewer QALYs than SA1 for both base cases.

2.5.3 SA3: No QALY loss when >12 week (purple) trend line sloping down

One of the scenarios occurring in some of the simulations in the probabilistic analyses was that the >12 week trend line sloped down and crossed the x axis. This means that the end of the trend line at 24 weeks could be below the x axis, implying that there would be some QoL loss (so QoL being below baseline) the longer the gap between the end of the intervention, and follow-up.

The committee discussed how feasible this might be. Their opinion was that for an intervention like acupuncture, it is unlikely that there would be continuing adverse effects that would worsen over time. Adverse events with acupuncture occur early, and people are likely to recover, whereas people who have a bad experience with exercise for occur, this can occur early or late, and the effects are pervasive for months after.

Therefore, although the committee accepted that the behaviour of the trend line is based on the uncertainty around the data points, and a model is a simplification of reality and therefore may sometimes be behaving in a way that might not make sense clinically: a sensitivity analysis tested the impact of not allowing negative QoL at the end of the >12 week trend line. I.e. QALY gain was calculated only up to where the trend line meets the x axis. This was tested in both the short and long term time horizons.

As this would mean that there would be no QALY loss from the end of the trend line subtracted from the overall QALY gain, then it is anticipated that this would make the QALYs slightly higher, and therefore improve cost effectiveness.

Note that this sensitivity analysis has only been applied to the base case data, and not to the data that includes the 52 week outcome, as simulations showed that the scenario described here of a downward sloping follow-up trend line that crosses the x axis happens less than 1% of the time when the 52 week outcome is included (either in SA1 or SA2).

2.5.4 SA4/SA5: Band 5/7 staff member

In the base case, the committee consensus was that a band 6 staff member might be a typical grade of professional that would deliver acupuncture. However, it could be a higher band, or it could be a lower band such as a band 5, providing they had adequate support.

The cost of a band 5 member of staff was used in a sensitivity analysis (SA4), and also the cost of a band 7 staff member (SA5).

2.5.5 SA6: Session length assumed where not reported - 20 min follow-ups

For three studies, the length of the sessions were not reported. In the base case it was assumed that 30 minutes would be a reasonable sessions length where this was not reported. Both because this was a typical sessions length in the UK, and also because the average session length from all the studies in the guideline clinical review for acupuncture was around 30 minutes.

In UK practice it might also be the case that the initial sessions could be 30 minutes, but follow-up sessions could be 20 minutes, based on committee member experience. Therefore, in this sensitivity analyses, for the three studies that the sessions length was not reported, the first session was assumed to be 30 minutes and the follow-up sessions were 20 minutes (as opposed to all being 30 minutes in the base case).

2.5.6 SA7: Overlap in treatment

There do exist clinics which operate by people receiving acupuncture in synchrony, rather than people being seen one at a time in timely sequence. These work by having either several rooms available or a larger space where patients can be separated by curtains, and the clinician moves between patients and can apply treatment to one patient whilst the previous is lying down with needles inserted. What this means is that multiple people can be treated at the same time, so the clinicians time is split across several patients rather than only on one patient at a time.

The studies in the review did not state whether this was the case, so they have assumed to only be treating one patient at a time. However, in a sensitivity analyses, the committee wanted to test the cost of the overlap treatment concept. It was assumed in this sensitivity analyses that two people could be treated during the length of the session from each study. What this essentially means is that the costs will be roughly half that of the base case (won't be exactly half as while staff costs will be halved, needle costs will stay the same), because of these efficiencies in delivering the intervention.

It is important to note that there are uncertainties regarding whether a lower cost (in this case from a different way of providing the intervention) would result in the same treatment effect as that of the studies being used. Therefore, it is important to interpret the results of all the sensitivity analyses around resource use carefully. This is discussed more in the discussion section.

2.5.7 SA8: Typical UK resource use

Resource use more typically associated with the UK was decided on by the committee as being 6 sessions of 30 minutes each. The cost of this was tested in this sensitivity analysis. Note that a band 6 staff member was used like the base case. This equated to a cost of £198. A standard error of 10% was assumed in order to make the cost probabilistic.

The resource use associated with the included studies was on average about 10 sessions of roughly 30 minutes. So this UK resource use would be cheaper, and therefore will lead to a lower ICER. Although again as mentioned above, there is uncertainty around the association between lower cost/fewer sessions and treatment effect.

2.5.8 SA9: Discounting outcomes at 1.5% (only relevant for lifetime horizon)

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.[20]

2.5.9 SA10: Alternative correlation coefficient (0.7) for imputing change from baseline standard deviations

As discussed in section 2.3.2.4, the data was used in the model by calculating change from baseline QoL, to incorporate any baseline differences in the studies. Where change from baseline standard deviations were not available, these were imputed using the baseline and final value standard deviations, and also using a variable known as a correlation coefficient. The correlation coefficient describes how similar the baseline and final measurements were across participants. In other words, it is the within patient correlation between baseline and follow-up measurements. A conservative value is considered to be 0.5. Zero would be no correlation, and 1 would be complete correlation between baseline and follow-up measurements. Baseline and follow-up measurements do tend to be correlated, hence why a value of 0.5 is considered a conservative one in the literature.

As the value of 0.5 used in the model was not based on the data (because no study reported change from baseline SD to calculate this), then this was tested in a sensitivity analysis. The literature varies as to what values are used for correlation coefficients, and justification is rarely provided for the value chosen.[26] A value of 0.7 was arbitrarily chosen as this would be less conservative than 0.5. The value itself is of less importance, but rather the purpose of this analysis is to assess whether a different value to 0.5 would affect the results at all.

Using a different correlation coefficient will not change the point estimates of treatment effect in the analysis, but it will lead to smaller standard deviations, which would have an impact on the confidence intervals of the point estimates, and lead to some impact on the probabilistic sensitivity analysis, and also on the regression weights, as these are based on the standard error of the treatment effect at each time point. The tables below show how the higher correlation coefficient has impacted the uncertainty around the base case treatment effect.

Table 15: EQ-5D mean difference between acupuncture and no acupuncture (up to 12 weeks) – impact of alternative correlation coefficient

Weeks (time zero being beginning of trial)	1	3	5	6	7	10	12
Base case - all data up to 12 weeks							
Pooled QoL difference	0.02	0.04	0.07	0.13	0.04	0.09	0.1
Uncertainty	-0.09 to 0.12	-0.07 to 0.15	-0.08 to 0.23	-0.01 to 0.27	-0.09 to 0.16	-0.06 to 0.24	0.09 to 0.12
Base case - all data up to 12 weeks (with correlation coefficient of 0.7)							
Pooled QoL difference	0.02	0.04	0.07	0.13	0.04	0.09	0.1
Uncertainty	-0.08 to 0.11	-0.06 to 0.14	-0.05 to 0.2	0.02 to 0.24	-0.08 to 0.15	-0.03 to 0.21	0.09 to 0.12

Table 16: EQ-5D mean difference between acupuncture and no acupuncture (beyond 12 weeks) – impact of alternative correlation coefficient

Weeks (time zero being end of intervention)	4	6	8	12	32
Base case – follow-up data (excluding 52 wk outcome)					
Pooled QoL difference	0.05	0.09	0.07	0.02	
Uncertainty	0 to 0.1	-0.05 to 0.24	-0.09 to 0.24	0 to 0.04	
Base case – follow-up data (excluding 52 wk outcome) (with correlation coefficient of 0.7)					
Pooled QoL difference	0.05	0.09	0.07	0.02	
Uncertainty	0.01 to 0.09	-0.02 to 0.2	-0.05 to 0.2	0.01 to 0.04	

Table 17: Regression weights – impact of alternative correlation coefficient

Weeks (time zero being beginning of trial)	1	3	5	6	7	10	12
Base case							
SE	0.05	0.06	0.08	0.07	0.06	0.08	0.01
Variance	0.0029	0.0031	0.0063	0.0051	0.0041	0.0059	0.0001
Inverse of variance (regression weights)	**348.4**	**317.5**	**159.9**	**196.0**	**245.9**	**170.7**	**17073.2**
Base case (with correlation coefficient of 0.7)							
SE	0.05	0.05	0.06	0.06	0.06	0.06	0.01
Variance	0.0023	0.0026	0.0041	0.0031	0.0034	0.0037	0.0001
Inverse of variance (regression weights)	**425.6**	**384.1**	**245.9**	**317.5**	**290.5**	**266.8**	**17073.2**

Weeks (time zero being end of intervention)	4	6	8	12
Base case – follow-up data (excluding 52 wk outcome)				
SE	0.03	0.07	0.08	0.01
Variance	0.0007	0.0055	0.0071	0.0001
Inverse of variance (regression weights)	**1536.6**	**182.7**	**141.1**	**9603.6**
Base case – follow-up data (excluding 52 wk outcome) (with correlation coefficient of 0.7)				
SE	0.02	0.06	0.06	0.01
Variance	0.0004	0.0031	0.0041	0.00006
Inverse of variance (regression weights)	**2400.9**	**317.5**	**245.9**	**17073.2**

2.5.10 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.

A threshold analyses was also undertaken on how many 30 minute sessions could be afforded that would make acupuncture borderline cost effective at the £20,000 per QALY threshold, given the QALY gains estimated using the trial data.

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'[23] 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. 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 deterministic and probabilistic base case results are presented in the Table 18. Probabilistic results are also presented graphically in Figure 11 and Figure 12. 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.

Acupuncture was associated with higher costs and higher QALYs. Higher costs are due to the cost of acupuncture as other costs were not incorporated due to uncertainty over whether they are affected. The incremental cost effectiveness ratio (ICER) for the lifetime analysis was £5,655 per QALY gained in the probabilistic analysis and £9,615 in the deterministic analysis. When not extrapolating beyond the trial data, the ICER was £11,333 in the probabilistic analysis and £11,160 in the deterministic analysis.

Both base cases show that the ICER is below the NICE threshold of £20,000, and therefore acupuncture would be considered cost effective. The probability of acupuncture being cost effective is also high.

Table 18: Base case results (discounted)

Base case	Analysis	Incremental cost	Incremental QALYs	Cost per QALY gained	Probability cost effective at £20k
Lifetime	Probabilistic	£350	0.058	£5,655	96%
	Deterministic	£350	0.036	£9,615	NA
No extrapolation beyond last trial observation (24 weeks)	Probabilistic	£350	0.031	£11,333	98%
	Deterministic	£350	0.031	£11,160	NA

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

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 11 and Figure 12 show the cost effectiveness plane with 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 12 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 11), 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 11: Base case results (lifetime): cost effectiveness plane

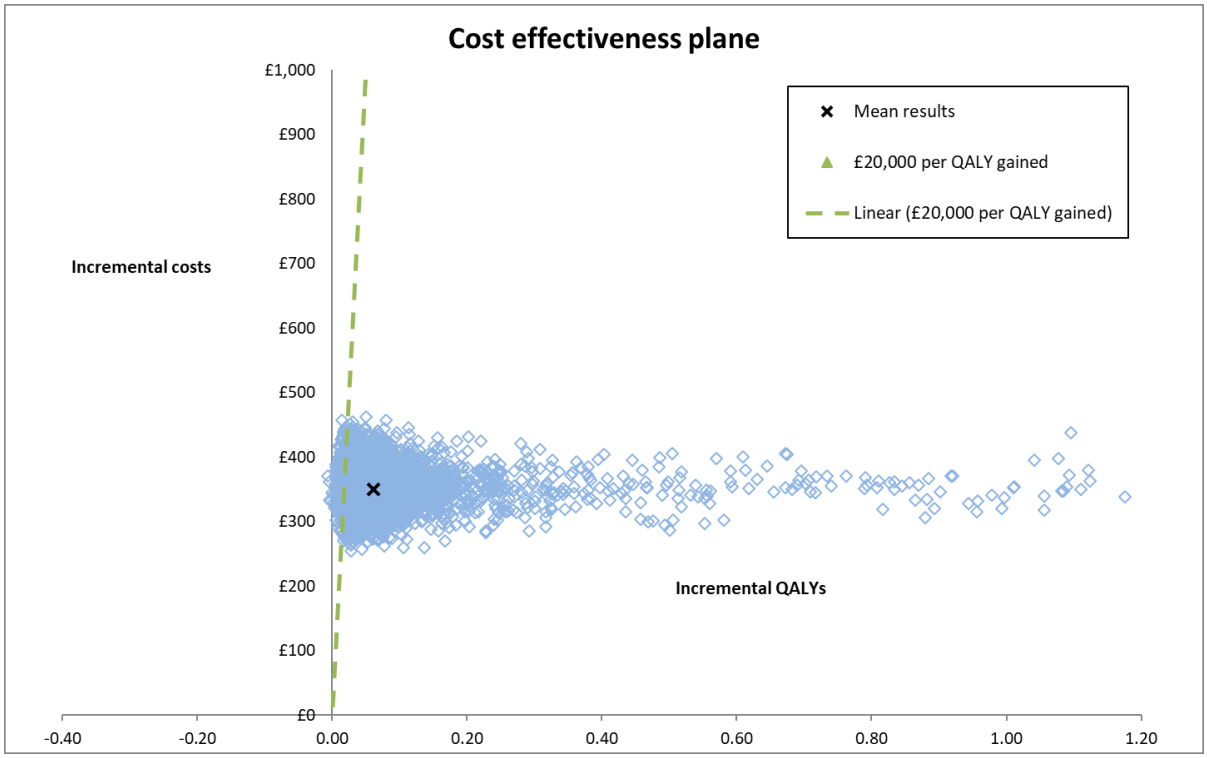
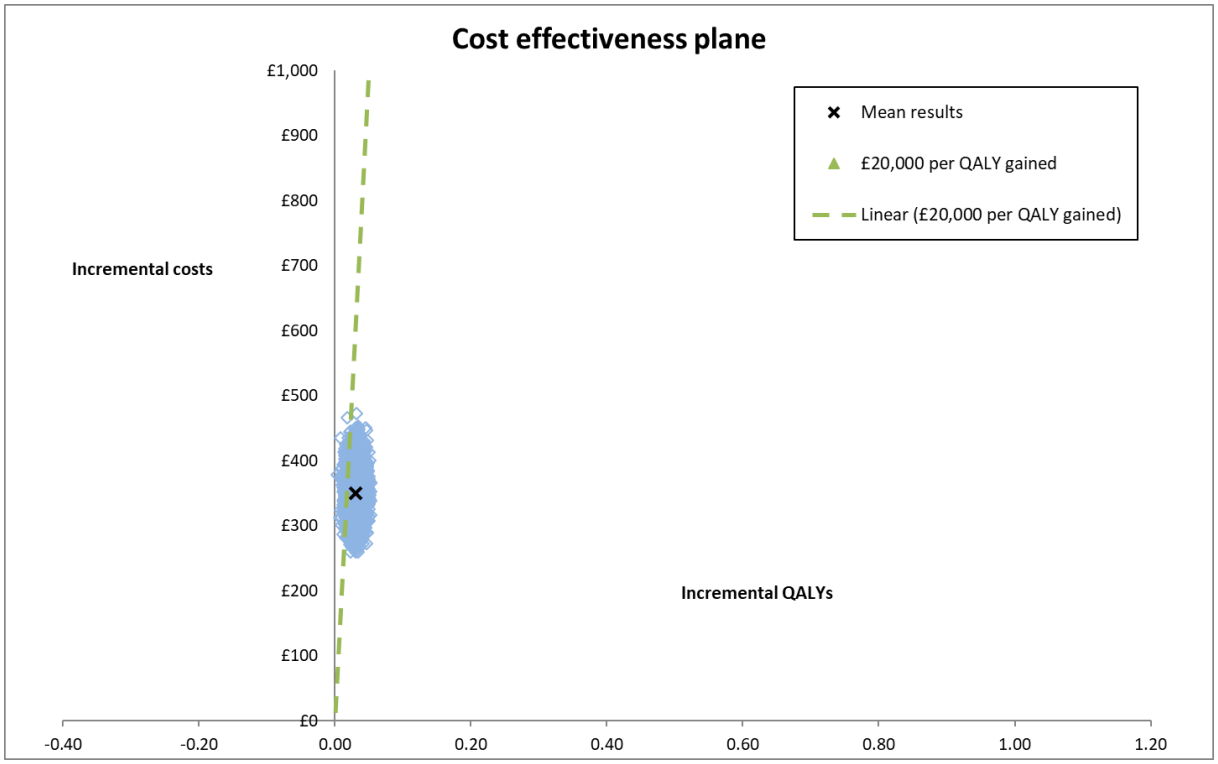


Figure 12: Base case results (no extrapolation): cost effectiveness plane



3.1.5 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

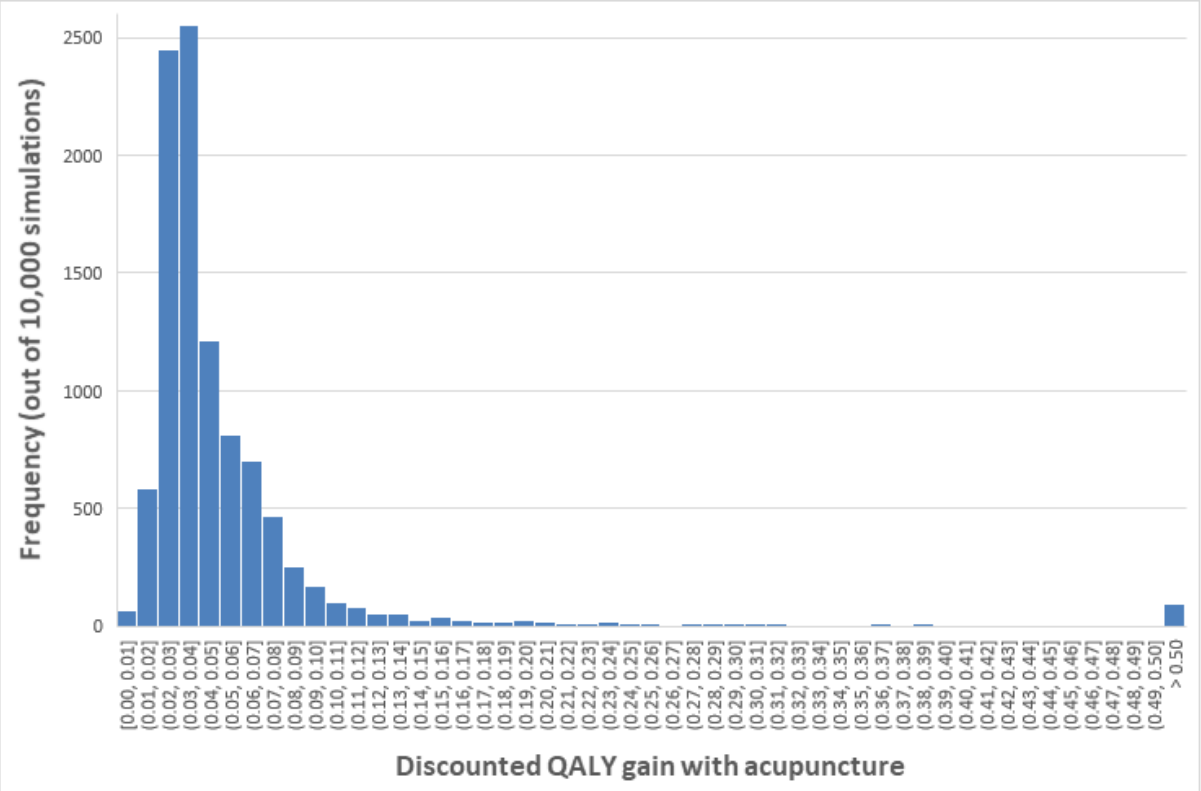
non-linear. The deterministic analysis (using the input point estimates and not the uncertainty around them) is also calculated and it is routine to consider if these are similar, and if not why not, as it may be the case that differences are due to programming errors in the model. As can be seen above, the incremental QALY estimates in this analysis 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 of the >12 week trend line in the base case is a downward sloping trend of QALY gain from acupuncture, 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 acupuncture 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 acupuncture from the base case simulations (10,000 simulations) against their frequency (Figure 13). 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 18), thereby confirming the explanation that the extrapolation of treatment effect can lead to very large QALY gains and a skewed distribution.

Figure 13: Distribution of QALY gain with acupuncture in base case (lifetime) probabilistic analysis



Abbreviations: QALY = quality-adjusted life years

Some further information that can contribute to what is happening in the probabilistic analysis can be seen in Table 19, where it is recorded how often different scenarios are occurring.

Table 19: Occurrence of treatment effect scenarios in base case (lifetime) probabilistic analysis

Scenario	Percentage of simulations occurring	
	< 12 week line (red)	> 12 week line (purple)
Slope direction		
Sloping down	3.83%	90.24%
Sloping up	96.17%	9.76%
Specific scenarios		
Sloping up		
1. Trend line fully in negative area	0%	0%
2. Trend line crosses x axis	19.61%	0%
3. Trend line fully in positive area	76.56%	9.76%
Sloping down		
4. Trend line fully in negative area	0%	0%
5. Trend line crosses x axis	0%	11.84%
6. Trend line fully in positive area	3.83%	78.40%

Overall, although 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 in both analyses are still well below the NICE threshold of £20,000 per QALY gained, and are therefore both in agreement that acupuncture is likely to be cost effective.

3.2 Sensitivity analyses

The results of the sensitivity analyses are presented in Table 20 and Table 21. These are presented separately for the two base cases. Acupuncture remained cost effective in all sensitivity analyses. The deterministic results are also reported for each base case in Table 21 because as discussed above, these can differ to the probabilistic results.

Table 20: Sensitivity analysis results (probabilistic)

Analysis	Lifetime analysis				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
Base case results	£350	0.058	£5,655	96%	£350	0.031	£11,333	98%
Including long term outcomes								
SA1: Including 52 week outcome from Essex 2017 (applied at 52 weeks as in trial)	£350	0.208	£1,495	100%	£350	0.097	£3,599	100%
SA2: Including 52 week outcome from Essex 2017 (applied at 32 weeks post 12 week trend line)	£350	0.190	£1,626	100%	£350	0.080	£4,377	100%
Avoiding QALY loss at end of >12 week trend line								
SA3: No QALY loss when >12 week (purple) trend line sloping down and last point in negative area	£350	0.060	£5,407	98%	£350	0.031	£11,237	98%
Resource use								
SA4: Band 5 staff member	£280	0.059	£4,442	98%	£279	0.031	£9,028	99%
SA5: Band 7 staff member	£420	0.061	£6,364	92%	£420	0.031	£13,491	93%
SA6: Session length assumed where NR - 20 min follow-ups	£260	0.061	£3,945	99%	£262	0.031	£8,425	99%
SA7: Overlap in treatment	£179	0.060	£2,796	99%	£179	0.031	£5,761	100%
SA8: Typical resource use: 6 sessions of 30 mins	£198	0.057	£3,278	99%	£198	0.031	£6,411	100%
Discount rate								
SA9: Discount rate at 1.5%	£350	0.061	£5,519	96%	NA	NA	NA	NA
Using alternative correlation coefficient for imputing change from baseline standard deviations								

Analysis	Lifetime analysis				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
SA10: Using alternative correlation coefficient for imputing change from baseline standard deviations	£350	0.055	£6,047	98%	£350	0.031	£11,159	99%
Threshold analyses								
Cost at which acupuncture has an ICER of £20,000 per QALY gained	£1,166	NA	NA	NA	£617	NA	NA	NA
QALY gain which acupuncture has an ICER of £20,000 per QALY gained	NA	0.018	NA	NA	NA	0.017	NA	NA
No. of sessions that would be cost effective (assuming 30 mins each and band 6)	35.3				18.7			

Note: Note that the sensitivity analysis on omitting QALY loss (SA2) only applies to the probabilistic analyses and not to the deterministic because that scenario only occurs in some probabilistic simulations.

Table 21: Sensitivity analysis results (deterministic)

Analysis	Lifetime analysis			No extrapolation of treatment effect analysis		
	Incremental cost	Incremental QALY	ICER (Cost per QALY gained)	Incremental cost	Incremental QALY	ICER (Cost per QALY gained)
Base case results	£350	0.036	£9,615	£350	0.031	£11,160
Including long term outcomes						
SA1: Including 52 week outcome from Essex 2017 (applied at 52 weeks as in trial)	£350	0.127	£2,747	£350	0.098	£3,571

Analysis	Lifetime analysis			No extrapolation of treatment effect analysis		
	Incremental cost	Incremental QALY	ICER (Cost per QALY gained)	Incremental cost	Incremental QALY	ICER (Cost per QALY gained)
SA2: Including 52 week outcome from Essex 2017 (applied at 32 weeks post 12 week trend line)	£350	0.108	£3,229	£350	0.080	£4,364
Resource use						
SA4: Band 5 staff member	£280	0.036	£7,690	£280	0.031	£8,925
SA5: Band 7 staff member	£420	0.036	£11,527	£420	0.031	£13,379
SA6: Session length assumed where NR - 20 min follow-ups	£261	0.036	£7,174	£261	0.031	£8,327
SA7: Overlap in treatment	£179	0.036	£4,924	£179	0.031	£5,715
SA8: Typical resource use: 6 sessions of 30 mins	£198	0.036	£5,447	£198	0.031	£6,322
Discount rate						
SA9: Discount rate at 1.5%	£350	0.036	£9,615	NA	NA	NA
Using alternative correlation coefficient for imputing change from baseline standard deviations						
SA10: Using alternative correlation coefficient for imputing change from baseline standard deviations	£350	0.037	£9,578	£350	0.032	£11,032
Threshold analyses						
Cost at which acupuncture has an ICER of £20,000 per QALY gained	£728	NA	NA	£627	NA	NA
QALY gain which acupuncture has an ICER of £20,000 per QALY gained	NA	0.018	NA	NA	0.018	NA
No. of sessions that would be cost effective (assuming 30 mins each and band 6)	22			19		

1

For all the sensitivity analyses, for both base cases, and whether deterministic or probabilistic, acupuncture remains cost effective with an incremental cost effectiveness ratio below £20,000 per QALY gained.

When including the 52 week follow-up data point, this leads to more QALYs than the base case because this led to an upward sloping trend which would create a bigger area under the curve than the base case.

When avoiding an area of QALY loss at the end of the follow-up trend line, this made little difference to the results, as doesn't happen in a high proportion of simulations (as can be seen from Table 9 (scenario 5).

When different resource use assumptions were tested, as expected, the analysis that had the largest impact was that of using a band 7 staff member, as this led to a higher cost. Although this still showed that acupuncture would be cost effective.

Using an alternative correlation coefficient had little impact on the results.

Threshold analyses show that, other things being equal, the cost of the intervention needs to be below £728 (£627 in no extrapolation base case) to make the intervention cost effective given the QALY gains estimated using the trial data. Note that the results of these threshold analyses are from the deterministic results, as the deterministic analyses had lower QALYs and therefore these are more conservative estimates of the cost threshold. This threshold analysis shows that the cost difference between acupuncture and usual care would have to be over twice the cost difference modelled for acupuncture not to be cost effective. This also provides some reassurance that should other healthcare costs be higher in the acupuncture group, as was suggested in the included economic evaluations, then this would still need to be a large difference to change the result.

A threshold analyses also looked at how many sessions of 30 minutes could be afforded at the cost thresholds identified above. This showed that if acupuncture was borderline cost effective at the £20,000 threshold, then this could afford 22 sessions of 30 minutes (or 19 sessions from the no extrapolation analysis). This would be much higher than might be typically delivered in England.

Keeping the cost the same as the base case, the QALY gain would have to be at least 0.018 (similar in both base cases because the cost is the same) for acupuncture 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 is not extrapolated) showed that the addition of acupuncture to usual care is cost effective with probabilistic ICERs of £5,655 and £11,333 respectively, and deterministic ICERs of £9,615 and £11,160 respectively. This conclusion was robust in sensitivity analyses such as varying staff members providing the intervention.

4.2 Limitations and interpretation

As highlighted in the methods section, this analysis aimed to assess whether acupuncture 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 7 studies in total. Although this is the majority of the studies that had usual care comparisons from the guideline review, this is still not a large number, and only one study was informing most timepoints because of the different lengths of interventions and timeframes that outcomes were reported. The populations in the studies however were felt to be representative of the chronic pain population.

Studies were used that either reported the utility measure EQ-5D or reported other measures that could be mapped to the EQ-5D. Measures reported that were mapped included the non-utility QoL measure SF-36 and pain measured on the VAS scale. Mapping of pain is less well established than mapping SF-36 but this increased the number of studies that could be used in the analysis from only 3 to 7.

Mapping is not without its limitations and is considered a second-best method of deriving utilities compared to direct elicitation using a utility instrument such as EQ-5D. Mapping from the SF-36 to the EQ-5D is well established and has been used in many models. Mapping from pain scales is however less common. The characteristics of particularly the pain mapping study[18] were investigated in more detail to assess its appropriateness and any limitations. The NICE Decision Support Unit (DSU), which produces training and materials to support the NICE technology appraisal programme, has produced a series of materials on utilities, and some on mapping specifically. Decision Support Unit document number 10[16] is on the use of mapping methods to estimate health state utility values, and documents methods that are considered good practice when undertaking a mapping exercise. Criteria laid out in the DSU document include; the characteristics of the estimation sample should be similar to the target sample for the mapping analysis. The population of the dataset that was used to derive the pain mapping algorithm (the SAPPHIRE trial) was that of rotator cuff disease, which is not too dissimilar to a chronic pain population. The average age in the SAPPHIRE trial was stated as a range of 55-59, with a mean VAS of 68.4 (on a 0-100 scale), and a mean EQ-5D of 0.45 to 0.51. The average age (non-weighted) of the chronic pain population was found to be 53 from the studies used in the exercise modelling, and 45 from the studies used in this acupuncture modelling. But it is important to note that this isn't the whole literature base for the guideline, but only the studies used for modelling. The range of VAS scores was found to be similar to the SAPPHIRE trial: with a range of 50-77 from the exercise modelling trials and 5-7 (all on 0-10 scale) from the acupuncture modelling trials that reported this. SF-36 mapping is much more established in cost-effectiveness analyses as a way to map to utilities. The study used here had a sample of over 6,000 people and used a different dataset for validation. The population is however very mixed and from lots of different disease areas because it is based on various RCT's and observational studies. It does however include some pain populations such as back pain and osteoarthritis. The DSU document also outlines the type of statistical tests that should be done to determine what

regression model to use, and that the range of the observed EQ-5D values should be reported to show whether the predicted utilities might involve extrapolation (where values predicted were not based on any observations). The pain mapping study had a much smaller sample than the SF-36 mapping study, and the range of the pain data is not reported for the larger sample of outcomes from 1,3 and 12 months that informs their regression, but only for 3 month and 12 month outcomes. The SF-36 mapping study stated that its dataset covered the whole range of the EQ-5D values. In terms of goodness of fit, both studies reported various statistics. The R squared was much higher in the SF-36 mapping study than the pain mapping study. However, it is less useful to compare this statistic from different regressions, than it is to compare it for different models based on the same dataset. Also, explanatory power is not a useful basis for assessing model performance, since the purpose of mapping functions is to predict values in other data sets. Other measures include looking at the difference between predicted and observed values at either the aggregate level by calculating Mean Error (ME) or at the individual level by calculating the Mean Absolute Error (MAE) or the Root Mean Squared Error (RMSE).[4] The smaller the value the better, and comparing the RMSE of the two mapping studies showed that the SF-36 study (0.178) did have smaller errors than the pain study (0.265).

Overall, although there were some concerns with the quality of the pain mapping study, this was the only paper identified that mapped from the VAS to the EQ-5D (without the inclusion of other QoL measures also), and has been used in other economic evaluations, and steps were also taken in this analysis to try and account for the uncertainty in the mapping using methods suggested in published literature.

This acupuncture analysis pooled data across clinical studies that had different intensities (in terms of frequency of sessions and overall number of sessions) of acupuncture, and also differences in the type of acupuncture. This may have an impact on treatment effect. 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 acupuncture that causes a benefit (i.e. the frequency, or the number of sessions, or the training and experience of the individual and therefore the extent of the contextual effect). 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.

Pooling the data included studies that were of different time periods. One had follow-up a long time after the intervention had ended and the quality of life benefits persisted over this time. The committee were not confident that quality of life continuing to improve from a course of acupuncture would be clinically plausible, especially so long after the interventions ended. For this reason, they decided to exclude this long-term outcome from the base case, and to include it 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 in the studies. The model tried to overcome some of this uncertainty by using weighted regression to generate trend lines based on QoL over time that better represented data points that were more certain.

The linear trend lines 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 acupuncture 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 acupuncture 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 the fluctuation of the underlying condition. 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.

The imputed EQ-5D from Essex 2017 could not be obtained from the authors. In terms of what impact this might have had on the results, this is likely to have led to a lower QoL improvement than the data currently used in the model and may have led to a slightly smaller treatment effect overall. However, given that the ICER was well below £20,000, this is unlikely to have made a large difference to the analysis. In addition, as this study had two follow-up timepoints, and only one was used in the base case, this mitigates the impact of this paper in the analysis somewhat.

Various sensitivity analyses tested assumptions about resource use. There is however uncertainty regarding whether the same treatment effect might be gained from fewer sessions for example, or whether a higher grade of staff could actually lead to more treatment effect. There are many aspects to an acupuncture intervention that could not be unpicked, such as the needling effects themselves, the contextual effects, the practitioner effects, as well as the uncertainty around any interaction between these effects. The opinion of experts on the committee who undertake (or have undertaken) acupuncture was that non-specific effect/contextual effect may be greater with a more experienced staff member. A large meta-analysis on acupuncture,[30] undertook analyses investigating the impact of characteristics of acupuncture treatment on treatment effect size, and found that there was a positive relationship between treatment effect and the number of sessions when acupuncture was compared to no acupuncture. Therefore, the results of sensitivity analyses around resource use need to be interpreted with some caution as the changes in resource use tested could also impact treatment effect but this is not captured.

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 other interventions.

No other costs have been accounted for in the analysis except for intervention costs. No data on whether acupuncture influences the use of other resources was found from the clinical review, however the two economic evaluations included in the guideline on acupuncture did report higher other healthcare resource use for people in the acupuncture group. The committee's opinion was that acupuncture anecdotally reduces other healthcare resource use, and so taking both the limited data found and the committee opinion, other resource use was omitted from the analysis as it was uncertain what assumptions should be made. We have also assumed no costs associated with the intervention beyond the intervention length in the trials. Results of the threshold analysis on costs found that costs in the acupuncture group would have to be much higher for acupuncture not to be cost effective, and so this provides some reassurance that even with additional costs, acupuncture could still be cost effective.

Overall, this analysis has pooled the available data from the clinical review that compared acupuncture to usual care, and reported EQ-5D or measures that could be mapped to EQ-5D, to estimate the potential cost effectiveness of acupuncture for a population with chronic

pain in general. The heterogeneity of the studies, and the number of studies used, should be taken into account when interpreting this analysis.

One important thing to take into consideration when considering the results of this analysis is that in addition to the studies that were used in this analysis that compared acupuncture and usual care, the clinical review also found evidence of treatment benefit in studies comparing acupuncture with sham acupuncture. This committee agreed that these provide evidence of treatment-specific effects of acupuncture in the chronic primary pain population. Other NICE guidelines have looked at the cost effectiveness of acupuncture versus no acupuncture in other chronic pain populations. The NICE guidelines on osteoarthritis,[19] and low back pain[22] also found published economic evidence suggesting acupuncture was cost effective. Neither guideline recommended acupuncture however. In low back pain this was because the committee concluded there was insufficient evidence of an overall treatment-specific effect to support a recommendation for acupuncture and so consideration of cost-effectiveness was not considered relevant. In the osteoarthritis guideline, the same reasoning applied whereby there wasn't considered to be a clinically important benefit above sham treatment.

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 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 there was uncertainty around the cost effectiveness of acupuncture, as the ICER was below £20,000 in the authors complete case base case analysis (with very wide confidence intervals), but was above £20,000 when missing data for EQ-5D and costs were imputed (again with very large confidence intervals).[14] The amount of missing data was quite high at around 40%. QoL from this trial was used in the guideline economic analysis (the complete case data, not the imputed). The overall QALYs in the complete case analysis (at 1 year) and in this acupuncture model when treatment effect was not extrapolated, were similar. The duration of effect when data was not extrapolated in the model was 24 weeks, therefore much less than 1 year, and yet the QALYs are similar, which can be explained by the fact that treatment effects in this model were from pooling many studies, some of which had higher QoL than this published study. In addition, the difference in ICER can be explained by the difference in incremental costs, as the study also included other costs not just intervention costs, and these showed higher health service costs in the acupuncture group (i.e. they were using more health services). QALYs were higher in the lifetime analysis of this model than in the published study because this also included assumptions about extrapolating treatment effect.

A second German economic evaluation was also identified that showed that acupuncture was cost effective.[32] The QoL from this study was also used in this analysis. The QALYs from this study were lower than those in the non-extrapolated analysis of this model. This was because this study was only 12 weeks long. In addition, the incremental costs are lower than those in this model. This is because the intervention costs used were much lower than UK costs.

Both studies also had limitations in terms of the costs of the staff involved looking low compared to UK costs, which will impact the cost effectiveness.

4.5 Conclusions

Acupuncture 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

acupuncture, and taking into account the cost of the intervention. The heterogeneity of the studies, and the number of studies used, should be taken into account when interpreting this analysis.

4.64 Implications for future research

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

References

1. 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

2. 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

3. Birch S, Jamison RN. Controlled trial of Japanese acupuncture for chronic myofascial neck pain: assessment of specific and nonspecific effects of treatment. Clinical Journal of Pain. 1998; 14(3):248-55

4. Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. European Journal of Health Economics. 2010; 11(2):215-25

5. Casanueva B, Rivas P, Rodero B, Quintial C, Llorca J, Gonzalez-Gay MA. Short-term improvement following dry needle stimulation of tender points in fibromyalgia. Rheumatology International. 2014; 34(6):861-6

6. 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

7. Chen YJ, Chen CT, Liu JY, Shimizu Bassi G, Yang YQ. What is the appropriate acupuncture treatment schedule for chronic pain? Review and analysis of randomized controlled trials. Evidence-Based Complementary Alternative Medicine. 2019; 2019:5281039

8. Cho JH, Nam DH, Kim KT, Lee JH. Acupuncture with non-steroidal anti-inflammatory drugs (NSAIDs) versus acupuncture or NSAIDs alone for the treatment of chronic neck pain: an assessor-blinded randomised controlled pilot study. Acupuncture in Medicine. 2014; 32(1):17-23

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

10. Coan RM, Wong G, Coan PL. The acupuncture treatment of neck pain: a randomized controlled study. American Journal of Chinese Medicine. 1981; 9(4):326-32

11. 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

12. 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/ 2/12/2015

13. Drummond. M, O'Briend. B, Stoddart. G, Torrance. G. Methods for the economic evaluation of health care programmes. 4th Edition ed. Oxford. 2015. Available from: https://global.oup.com/academic/product/methods-for-the-economic-evaluation-of-health-care-programmes-9780199665884?cc=gb&lang=en&

14. Essex H, Parrott S, Atkin K, Ballard K, Bland M, Eldred J et al. An economic evaluation of Alexander Technique lessons or acupuncture sessions for patients with chronic neck pain: A randomized trial (ATLAS). PloS One. 2017; 12(12):e0178918

15. Glick HA, Doshi JA. Economic Evaluation in Clinical Trials (Handbooks in Health Economic Evaluation). Oxford. 2014. Available from: https://global.oup.com/academic/product/economic-evaluation-in-clinical-trials-9780199685028?cc=gb&lang=en&

16. Longworth L, Rowen D. NICE DSU Technical Support Document 10: The use of mapping methods to estimate health state utility values. 2011. Available from: http://www.nicedsu.org.uk

17. 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

18. Maund E, Craig D, Suekarran S, Neilson A, Wright K, Brealey S et al. Management of frozen shoulder: a systematic review and cost-effectiveness analysis. Health Technology Assessment. 2012; 16(11)

19. 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 3/7/2014

20. 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

21. National Institute for Health and Care Excellence. Addendum to Clinical Guideline 72, Attention deficit hyperactivity disorder [dietary interventions]: Clinical Guideline Addendum 72.1. London. National Institute for Health and Care Excellence, 2016. Available from: https://www.nice.org.uk/guidance/ng87/evidence/dietary-interventions-pdf-4844210798

22. 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

23. 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

24. National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal 2013. London. National Institute for Health and Clinical Excellence, 2013. Available from: http://publications.nice.org.uk/pmg9

25. NHS Supply Chain Catalogue. NHS Supply Chain, 2019. Available from: http://www.supplychain.nhs.uk/

26. Pearson MJ, Smart NA. Reported methods for handling missing change standard deviations in meta-analyses of exercise therapy interventions in patients with heart failure: A systematic review. PloS One. 2018; 13(10):e0205952

27. Ramsey S, Willke R, Briggs A, Brown R, Buxton M, Chawla A et al. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA Task Force report. Value in Health. 2005; 8(5):521-33

28. Schlaeger JM, Xu N, Mejta CL, Park CG, Wilkie DJ. Acupuncture for the treatment of vulvodynia: a randomized wait-list controlled pilot study. Journal of Sexual Medicine. 2015; 12(4):1019-27

29. Statistics OfN. Life tables for England. 2018. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesenglandreferencetables Last accessed: 24/09/2019

30. Vickers AJ, Vertosick EA, Lewith G, MacPherson H, Foster NE, Sherman KJ et al. Acupuncture for Chronic Pain: Update of an Individual Patient Data Meta-Analysis. Journal of Pain. 2018; 19(5):455-474

31. Watson J, Helliwell P, Morton V, Adebajo A, Dickson J, Russell I et al. Shoulder acute pain in primary healthcare: is retraining effective for GP principals? SAPPHIRE--a randomized controlled trial. Rheumatology. 2008; 47(12):1795-802

32. Willich SN, Reinhold T, Selim D, Jena S, Brinkhaus B, Witt CM. Cost-effectiveness of acupuncture treatment in patients with chronic neck pain. Pain. 2006; 125(1-2):107-13

33. Witt CM, Jena S, Brinkhaus B, Liecker B, Wegscheider K, Willich SN. Acupuncture for patients with chronic neck pain. Pain. 2006; 125(1-2):98-106

Appendix A: Data extracted from studies and associated mapped EQ-5D values

A.1 SF-36 raw data and mapped EQ-5D values

Intervention	Measurement timeframe		SF-36 domain								EQ-5D Mapped from SF-36	EQ-5D change from baseline	EQ-5D improvement from acupuncture (a)
			Physical functioning	Social role	Physical role	Emotional role	Mental health	Vitality	Bodily pain	General health			
Casanueva (2014) (b)													
Acupuncture	Baseline	Mean	26.8	39.3	6.5	23.3	37.3	17.1	20.5	22.0	0.32		
		Lower CI	22.2	32.4	1.0	13.1	31.8	13.5	16.2	18.4	0.24		
		Upper CI	31.4	46.1	12.0	33.5	42.8	20.7	24.8	25.6	0.40		
	Post intervention (at 6 weeks)	Mean	33.2	45.6	23.0	35.1	44.9	26.1	33.6	27.6	0.45	0.13	0.127
		Lower CI	27.5	39.1	14.0	23.4	40.3	21.4	28.4	24.0	0.37		
		Upper CI	38.9	52.1	32.0	46.8	49.5	30.8	38.8	31.2	0.53		
	Follow-up (at 12 weeks)	Mean	31.1	45.4	18.6	38.1	41.0	21.1	28.0	23.6	0.40	0.08	0.091
		Lower CI	26.2	37.7	9.4	26.1	35.4	16.6	22.0	19.3	0.31		
		Upper CI	36.0	53.1	27.8	50.1	46.7	25.6	34.1	27.9	0.49		
Control	Baseline	Mean	26.6	36.5	5.7	25.0	39.2	12.3	17.3	22.4	0.31		
		Lower CI	22.5	29.3	1.1	15.0	33.7	8.8	13.8	19.3	0.24		
		Upper CI	30.6	43.6	10.4	35.0	44.7	15.8	20.8	25.5	0.38		
	Post intervention (at 6 weeks)	Mean	28.8	34.8	4.3	28.7	37.9	9.7	17.1	19.1	0.32	0.003	
		Lower CI	24.9	29.1	-0.6	18.2	33.0	6.2	12.8	15.2	0.24		
		Upper CI	32.7	40.5	9.2	39.2	42.8	13.2	21.4	23.0	0.39		
	Follow-up (at 12 weeks)	Mean	28.6	34.7	4.8	17.2	36.3	13.7	16.0	20.7	0.30	-0.01	
		Lower CI	24.2	28.7	0.0	7.5	30.9	10.5	12.4	17.6	0.23		
		Upper CI	33.0	40.7	9.6	26.9	41.7	16.9	19.6	23.8	0.38		
Witt (2006) (b)													
Acupuncture	Baseline	Mean	63.6	63.3	38.9	59.4	57.7	40.0	37.9	52.6	0.67		
		Lower CI	62.6	62.2	37.1	57.4	56.8	39.2	37.1	51.7	0.66		

Intervention	Measurement timeframe		SF-36 domain								EQ-5D Mapped from SF-36	EQ-5D change from baseline	EQ-5D improvement from acupuncture (a)
			Physical functioning	Social role	Physical role	Emotional role	Mental health	Vitality	Bodily pain	General health			
		Upper CI	64.6	64.4	40.7	61.4	58.6	40.8	38.7	53.5	0.68		
	Post intervention (at 12 weeks)	Mean	72.0	75.6	63.4	73.3	66.3	51.0	58.9	58.2	0.81	0.134	0.106
		Lower CI	71.2	74.6	61.5	71.3	65.6	50.2	57.9	57.6	0.80		
		Upper CI	72.8	76.6	65.4	75.4	67.1	51.8	59.9	58.8	0.81		
	Follow-up (at 24 weeks)	Mean	71.6	74.6	62.1	73.8	65.3	50.6	58.2	57.4	0.80	0.128	0.022
		Lower CI	70.8	73.4	60.0	71.6	64.5	49.7	57.0	56.7	0.79		
		Upper CI	72.5	75.8	64.1	75.9	66.2	51.5	59.4	58.1	0.81		
Control	Baseline	Mean	63.9	64.4	40.5	61.0	58.9	42.1	40.6	52.5	0.69		
		Lower CI	62.8	63.2	38.7	59.0	58.0	41.2	39.7	51.6	0.68		
		Upper CI	65.0	65.6	42.3	63.0	59.8	43.0	41.5	53.4	0.70		
	Post intervention (at 12 weeks)	Mean	64.8	66.5	45.6	62.7	60.3	47.2	45.9	52.9	0.71	0.029	
		Lower CI	64.1	65.4	43.8	60.8	59.5	46.5	44.9	52.3	0.71		
		Upper CI	65.6	67.5	47.5	64.7	61.0	48.0	46.9	53.5	0.72		
	Follow-up (at 24 weeks)	Mean	70.6	73.9	61.1	70.8	65.2	51.6	57.6	56.8	0.79	0.106	
		Lower CI	69.8	72.7	58.9	68.6	64.3	50.7	56.4	56.0	0.78		
		Upper CI	71.5	75.1	63.3	73.0	66.1	52.5	58.8	57.5	0.80		

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). Light orange in the table means outcomes measured later after the intervention ended (follow-up outcomes).

(a) EQ-5D change from baseline in the acupuncture 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 Casanueva (2014), outcomes are measured at 6 weeks and at 12 weeks. So the EQ-5D improvement at 6 weeks is the change in baseline in the acupuncture group at 6 weeks minus the change in baseline in the control group at 6 weeks (0.13 - 0.003 = 0.127). The same is then calculated for the 12 week outcomes. These are crude estimates for illustration as in the model the changes from baseline in each arm were input into Revman do dervie the QoL difference between the groups.

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

A.2 EQ-5D raw data

Intervention	Measurement timeframe		EQ-5D value	EQ-5D change from baseline	EQ-5D improvement from acupuncture
Essex (2017) (a)					
Acupuncture	Baseline	Mean	0.683		
		SD	0.179		
	Follow-up (at 24 weeks)	Mean	0.755	0.072	0.05
		SD	0.190		
	Follow-up (at 52 weeks)	Mean	0.766	0.083	0.053
		SD	0.188		
Control	Baseline	Mean	0.697		
		SD	0.179		
	Follow-up (at 24 weeks)	Mean	0.719	0.022	
		SD	0.214		
	Follow-up (at 52 weeks)	Mean	0.727	0.03	
		SD	0.197		

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). Light orange in the table means outcomes measured later after the intervention ended (follow-up outcomes).

(a) Note that the paper reported SD's and they are reported here as this was an EQ-5D paper and SD's are needed for the meta-analysis therefore it was not necessary to calculate confidence intervals.

A.3 Pain VAS raw data and mapped EQ-5D values

Intervention	Measurement timeframe		Pain (on scale 0-10)		EQ-5D mapped from pain scale	EQ-5D change from baseline	EQ-5D improvement from acupuncture
Birch (1998) (a) (b)							
Acupuncture	Baseline	Mean	4.8		0.579		
		Lower CI	3.75		0.547		
		Upper CI	5.85		0.613		
	Post intervention (at 10 weeks)	Mean	1.87		0.673	0.094	0.090
		Lower CI	0.82		0.639		
		Upper CI	2.92		0.708		
Control	Baseline	Mean	4.9		0.576		
		Lower CI	3.76		0.541		
		Upper CI	6.04		0.612		
	Post intervention (at 10 weeks)	Mean	4.76		0.581	0.004	
		Lower CI	3.62		0.545		
		Upper CI	5.90		0.616		
Cho (2014) (a)							
Acupuncture	Baseline	Mean	6.9		0.515		
		Lower CI	6.53		0.504		
		Upper CI	7.27		0.526		
	Partway through intervention (at 1 week)	Mean	5.3		0.564	0.049	0.016
		Lower CI	4.78		0.548		
		Upper CI	5.82		0.580		
	Post intervention (at 3 weeks)	Mean	3.8		0.611	0.096	0.041
		Lower CI	4.52		0.540		

		Upper CI	6.08		0.588		
	Follow-up (at 7 weeks)	Mean	4.05		0.603	0.088	0.039
		Lower CI	3.38		0.582		
		Upper CI	4.72		0.624		
Control	Baseline	Mean	6.07		0.540		
		Lower CI	5.79		0.532		
		Upper CI	6.35		0.549		
	Partway through intervention (at 1 week)	Mean	5		0.573	0.033	
		Lower CI	3.95		0.541		
		Upper CI	6.05		0.606		
	Post intervention (at 3 weeks)	Mean	4.3		0.595	0.055	
		Lower CI	3.36		0.566		
		Upper CI	5.24		0.625		
	Follow-up (at 7 weeks)	Mean	4.5		0.589	0.049	
		Lower CI	3.28		0.551		
		Upper CI	5.72		0.627		
Schlaeger (2015) (a)							
Acupuncture	Baseline	Mean	5.6		0.554		
		Lower CI	4.66		0.526		
		Upper CI	6.54		0.584		
	Post intervention (at 5 weeks)	Mean	2.7		0.646	0.092	0.073
		Lower CI	1.85		0.619		
		Upper CI	3.55		0.674		
Control	Baseline	Mean	5.7		0.551		
		Lower CI	4.55		0.516		
		Upper CI	6.85		0.587		
	Post intervention (at 5 weeks)	Mean	5.1		0.570	0.019	
		Lower CI	3.66		0.526		
		Upper CI	6.54		0.615		
Coan (1981) (a)							
Acupuncture	Baseline	Mean	5.97		0.543		
		Lower CI	4.98		0.513		
		Upper CI	6.96		0.574		
	Follow-up (at 12 weeks)	Mean	3.63		0.616	0.073	0.075
		Lower CI	2.40		0.577		
		Upper CI	4.86		0.656		
Control	Baseline	Mean	5.30		0.564		
		Lower CI	4.02		0.525		
		Upper CI	6.58		0.604		
	Post intervention (at 12 weeks)	Mean	5.37		0.562	-0.002	
		Lower CI	4.14		0.524		
		Upper CI	6.60		0.600		

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). Light orange 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) This study looked like it was using the NRS scale rather than the VAS but has been used as a VAS for the mapping to EQ-5D, as both the NRS and VAS are on the same scale. (0-10).

Appendix B: Data for meta-analysis

B.1 Data for meta-analysis

										Feeding into meta-analysis						
Study	**Intervention**	**EQ-5D baseline mean**	**EQ-5D mean - outcome point 1**	**EQ-5D mean - outcome point 2**	**EQ-5D mean - outcome point 3**	**Baseline SD**	**Outcome point 1 SD**	**Outcome point 2 SD**	**Outcome point 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**
Essex 2017	Acupuncture	0.683	0.755	0.766		0.179	0.190	0.188		0.072	0.083		0.185	0.184		104
	control	0.697	0.719	0.727		0.179	0.214	0.197		0.022	0.030		0.199	0.189		100
Casanueva 2014	Acupuncture	0.322	0.453	0.404		0.407	0.405	0.444		0.131	0.081		0.406	0.427		60
	control	0.315	0.318	0.304		0.374	0.368	0.381		0.003	-0.010		0.371	0.378		60
Witt 2006	Acupuncture	0.671	0.805	0.799		0.288	0.227	0.265		0.134	0.128		0.263	0.277		1753
	control	0.686	0.715	0.792		0.298	0.253	0.267		0.029	0.106		0.279	0.284		1698
Cho 2014	Acupuncture	0.515	0.564	0.611	0.603	0.099	0.143	0.214	0.187	0.049	0.096	0.088	0.127	0.185	0.162	30
	control	0.540	0.573	0.595	0.589	0.053	0.204	0.185	0.238	0.033	0.055	0.049	0.183	0.165	0.217	15
Birch 1998	Acupuncture	0.579	0.673			0.205	0.215			0.094			0.210			15
	control	0.576	0.581			0.221	0.221			0.004			0.221			15
Coan 1981	Acupuncture	0.543	0.616			0.188	0.244			0.073			0.221			15
	control	0.564	0.562			0.247	0.238			-0.002			0.242			15
Schlaeger 2015	Acupuncture	0.554	0.646			0.199	0.187			0.092			0.193			18
	control	0.551	0.570			0.241	0.306			0.019			0.279			18

Note: Blue means studies where EQ-5D was mapped from SF-36 data, pink means studies where EQ-5D was mapped from pain data, and therefore EQ-5D mean and follow-up was mapped, as well as their confidence intervals. Green means reported in the paper. 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.
(c) Yellow cells have been adjusted using variance adjustment method to account for uncertainty in the mapping.

B.2 Adjusted standard deviations for mapping uncertainty

						Unadjusted SD's				Adjusted SD's			
Study	**Intervention**	**EQ-5D baseline mean**	**EQ-5D mean - outcome point 1**	**EQ-5D mean - outcome point 2**	**EQ-5D mean - outcome point 3**	**Baseline SD**	**Outcome point 1 SD**	**Outcome point 2 SD**	**Outcome point 3 SD**	**Baseline SD**	**Outcome point 1 SD**	**Outcome point 2 SD**	**Outcome point 3 SD**
Casanueva 2014	Acupuncture	0.322	0.453	0.404		0.311	0.310	0.340		0.407	0.405	0.444	
	control	0.315	0.318	0.304		0.286	0.282	0.292		0.374	0.368	0.381	
Witt 2006	Acupuncture	0.671	0.805	0.799		0.220	0.174	0.202		0.288	0.227	0.265	
	control	0.686	0.715	0.792		0.228	0.193	0.205		0.298	0.253	0.267	
Cho 2014	Acupuncture	0.515	0.564	0.611	0.603	0.031	0.045	0.068	0.060	0.099	0.143	0.214	0.187
	control	0.540	0.573	0.595	0.589	0.017	0.065	0.059	0.076	0.053	0.204	0.185	0.238
Birch 1998	Acupuncture	0.579	0.673			0.065	0.068			0.205	0.215		
	control	0.576	0.581			0.070	0.070			0.221	0.221		
Coan 1981	Acupuncture	0.543	0.616			0.060	0.078			0.188	0.244		
	control	0.564	0.562			0.078	0.076			0.247	0.238		
Schlaeger 2015	Acupuncture	0.554	0.646			0.063	0.059			0.199	0.187		
	control	0.551	0.570			0.077	0.097			0.241	0.306		

Appendix C: Combining intervention arms of 3 arm trials

Study		N	EQ-5D baseline mean	EQ-5D mean - outcome point 1	EQ-5D mean - outcome point 2	EQ-5D mean - outcome point 3	Baseline SD	Outcome point 1 SD	Outcome point 2 SD	Outcome point 3 SD
Cho (2014)	Acu	15	6.7	5	3.8	4.3	0.7	1.9	2.4	2
	Acu + NSAIDs	15	7.1	5.6	3.8	3.8	1.3	0.7	1.8	1.6
	COMBINED ARMS	30	6.9	5.3	3.8	4.05	1.05	1.44	2.08	1.80

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