A Systematic Review of the Efficacy of Acupuncture in the Treatment of Osteoarthritis of the Knee Joint

K.C. Turner & S. Igo

Abstract

Objectives: To evaluate the efficacy of acupuncture compared with sham acupuncture and standard care for improving pain and function in osteoarthritis (OA) of the knee.

Methods: A systematic review of randomised controlled trials comparing acupuncture with sham acupuncture or standard care for OA of the knee was conducted. Computerised databases were searched to February 2011. Studies were selected in which patients with OA knee were randomised to receive either acupuncture or a control consisting of sham acupuncture or standard care. Standard care interventions included advice, education, self management, standardised exercises or standard pharmacological management.

The internal validity of the studies was assessed independently by two reviewers using the PEDro scale. The adequacy of acupuncture within the studies was also assessed using pre-specified criteria. Short and long-term outcomes of pain and function were extracted from the trials and the mean change in outcome was calculated in each group for comparison.

Results: Eleven studies were included, of which eight utilised a sham acupuncture control and five used standard care. Two studies were three-armed trials incorporating both sham and standard care groups. All trials had PEDro scores of ≥6/10. Compared to sham acupuncture, there was inconclusive evidence to support the efficacy of acupuncture for improving pain or function at both short and long-term follow up. Compared to standard care, strong evidence was found to support the efficacy of acupuncture for the short and long-term improvement of pain and function. It was unclear whether the adequacy of the acupuncture within the trials had a bearing upon the outcomes.

Conclusion: Sham controlled trials reveal inconclusive evidence for the efficacy of acupuncture in improving pain and function in OA knee. There is strong evidence to support the efficacy of acupuncture over standard care for improving pain and function in OA knee. It is unclear to what degree these outcomes may be attributed to the adequacy of acupuncture treatment.
1: Introduction / Background

Osteoarthritis (OA) is a chronic joint condition involving cartilage degeneration, proliferation and remodelling of subchondral bone (Ernst, 1997). It is characterised by symptoms of pain, stiffness and functional disability, significantly impacting upon quality of life (Zhang et al. 2008) and ranking among the top 10 causes of disability worldwide (Gupta et al. 2005). Osteoarthritis most commonly affects the knee joint, and it follows that it is the most common reason for total knee replacement surgery (Felson & Zhang, 1998). Prevalence of OA knee in adults over 30 is estimated at 6%, with incidence increasing with age (Felson & Zhang, 1998). With an ageing population, the number of people with OA disability is expected to double by the year 2020, increasing the already significant economic burden of OA (Gupta et al. 2005).

The treatment of OA is directed towards reducing joint pain and stiffness, maintaining joint mobility, improving health-related quality of life and limiting the progression of joint damage (Zhang et al. 2008). Symptom management often involves the use of non-steroidal anti-inflammatory drugs (NSAID’s), as recommended in the National Institute for Health and Clinical Excellence (NICE) guidelines for the Care and Management of Osteoarthritis in Adults (NICE, 2008). However, NSAID’s have well-documented side effects such as gastrointestinal bleeding (Kwon, Pittler & Ernst 2006) resulting in approximately 2200 deaths and 1200 emergency admissions per year in the UK (Blower et al. 1997). Their value in the management of OA is therefore questionable when their side effects are considered alongside the short term pain relief they have been shown to provide in this patient group (Bjordal et al. 2004). If these drugs do not lead to adequate response, joint replacement surgery is often recommended (Witt et al. 2006). With increasing governmental pressure to contain health care expenditure and limit hospital stays, there is a growing requirement for effective conservative
treatment options (Gupta et al 2005). Effective non-pharmacological treatments are weight loss and exercise (Fransen, McConnell & Bell 2003). However, patients with OA may find exercise difficult due to poor pain control (Manheimer et al 2007). In the UK, up to 54% of OA patients therefore seek management with alternative therapies (Visser, Peters & Rasher 1992). Treatments such as acupuncture are consequently becoming increasingly popular for this condition (Rosemann et al 2007).

Acupuncture has been shown to have few serious side effects and is regarded as a largely safe treatment method (Ernst & White 2001), therefore showing potential advantages over NSAID use. A safety review of acupuncture used specifically for the treatment of OA knee has also confirmed the safety of this treatment modality (Yamashita et al 2006). Recent evidence based guidelines from Osteoarthritis Research Society International for the management of OA of the knee or hip include acupuncture as one of twelve non-pharmacological modalities recommended for use (Zhang et al 2008). Contrary to these recommendations, NICE (2008) report that available evidence suggests acupuncture can provide only short to medium term relief for the treatment of peripheral joint arthritis. In addition, electro-acupuncture was not recommended in light of poor cost-effectiveness outweighing a reliable evidence base to highlight its clinical benefits. Conversely, acupuncture for the treatment of OA of the knee or hip has been shown to be a cost effective treatment option when compared with routine care options including NSAID’s in a recent large scale randomised controlled trial (RCT) (Reinhold et al 2008).

Acupuncture involves the stimulation of specific points along meridians on the body with fine needles, to alter physiological functions including pain control (Stux, Berman & Pomeranz 2003). Although the exact physiological mechanisms of acupuncture analgesia remain
unclear, research suggests that acupuncture relieves pain through both the modulation of spinal reflexes, and the release of endogenous opioids in the central nervous system (CNS) (Andersson & Lundeberg 1995). Acupuncture has been shown to stimulate the release of beta-endorphin, enkephalin and endomorphin in the CNS (Lin & Chen 2008), which in turn activate opioid receptors which help to modulate inflammatory pain (Zhang et al 2005). The opioid peptide release and thus the analgesic benefits have been shown to be more pronounced with electroacupuncture than with manual acupuncture treatment (Ulett, Han & Han 1998).

It is widely recognised that obtaining the ‘DeQi’ response during needling is an important factor for achieving optimal analgesic benefits (Sandberg et al 2003). DeQi is translated as “to obtain Qi” or “the arrival of Qi” (Kong et al 2007). In Traditional Chinese Medicine (TCM) terms it is used to describe the connection between the acupuncture needles and the energy pathways in the body (Kong et al 2007). It describes a characteristic sensation, frequently expressed as a distended, sore, heavy, migratory sensation, experienced by the patient, and sometimes also by the practitioner (Mao et al 2007). DeQi is usually achieved by manual rotation, or lifting and thrusting of the needles, or by the addition of electrical stimulation by the practitioner (Stux, Berman & Pomeranz, 2003). Western medical explanations for this phenomenon suggest that the DeQi response relates to the activation of Aδ and C-fibres from free nerve endings in the skin and muscle, secondary to the noxious stimulation of the acupuncture needle (Andersson & Lundeberg 1995, Hui et al 2007). The Aδ and C-fibres send descending projections through the dorsolateral funiculus that terminate in the dorsal horn at all levels of the spinal cord (Le Bars et al 1991). Some neurones in the dorsal horn of the spinal cord are known to be strongly inhibited by the noxious stimulus produced by needling, in a phenomenon known as diffuse noxious inhibitory control (DNIC) (Le Bars et al...
This leads to the attenuation of pain, even in areas extra-segmental to the acupuncture needle (MacPherson et al 2008). Functional magnetic resonance imaging (fMRI) studies have also revealed the pain neuromatrix in the limbic system to be deactivated by achieving the DeQi sensation (Hui et al 2007). This deactivation of limbic structures has been associated with increased activity in the hypothalamus, resulting in the activation of sympathetic and pain inhibiting mechanisms (Napadow et al 2007).

There is an increasing body of evidence to support the use of acupuncture in the treatment of pain arising from peripheral joint OA (Manheimer et al 2010, Kwon, Pittler & Ernst 2006). However, early reviews of trials investigating the efficacy of acupuncture for OA knee reported insufficient and inconclusive evidence, reflecting the generally poor methodological quality and small scale of the included trials (Ezzo et al 2001). In recognition of the shortcomings of acupuncture trials and in an attempt to facilitate an improvement in their quality, the Standards for Reporting Interventions in Controlled Trials of Acupuncture (STRICTA) guidelines were published in 2001 (Macpherson et al 2001b). They specified items for researchers to include when conducting acupuncture RCT’s which, it was hoped would lead to more rigorous trial designs. Following their implementation, several higher quality studies were published, and more recent systematic reviews have reflected these improvements (White et al 2007, Selfe & Taylor 2008).

Many acupuncture RCT’s implement the use of ‘sham’ acupuncture in an attempt to incorporate a patient-blinded placebo intervention which separates the specific from the non-specific effects of treatment (Lund, Naslund & Lundeberg 2009). During ‘traditional’ acupuncture, needles are inserted into recognised meridian acupoints and are usually stimulated until DeQi is evoked (Lund, Naslund & Lundeberg 2009). In contrast, sham
methods may utilise superficial needling at non-acupoints or at true acupoints not indicated for the condition being treated (Moffet 2009). Other sham techniques utilise non-penetrating needles with a blunt tip which is affixed to the skin (Lund, Naslund & Lundeberg 2009). However, evidence suggests that such sham techniques are not physiologically inert, with many studies revealing that sham techniques may be as efficacious as traditional acupuncture (Moffet 2009). These techniques therefore cannot be considered a placebo intervention, but are nevertheless often utilised as a technique to attempt to minimise the specific effects of needling (Hammerschlag 1998) and to evaluate the validity of traditional acupoint indications and locations (Moffet 2009).

Two recent systematic reviews have evaluated the effects of acupuncture in OA of all peripheral joints, rather than the knee joint specifically. Kwon, Pittler & Ernst (2006) conducted a meta-analysis from which they concluded acupuncture to be superior to sham treatment in the short term, and recommended acupuncture to be an option worthy of consideration for OA knee in particular. More recent evidence is appraised by Manheimer et al (2010) in a Cochrane review of acupuncture for peripheral joint OA. Although the authors concluded statistically and clinically significant benefits of acupuncture in waiting list controlled trials, sham-controlled trials did not meet pre-defined thresholds for clinical relevance.

Reviews examining the efficacy of acupuncture for OA knee in isolation include a meta-analysis by White et al (2007), in which specified criteria for ‘adequate acupuncture’ was utilised, based upon clinical experience and empirical data. The authors concluded that acupuncture which meets the criteria for adequate treatment is significantly superior to sham acupuncture or no additional treatment in the management of chronic knee pain. However, in
light of the heterogeneity of the results, further research was recommended in the form of large scale, high quality trials. These positive findings were reiterated in the systematic review of Selfe & Taylor (2008) which concluded acupuncture to be an effective treatment for the pain and physical dysfunction associated with OA of the knee. However, shortcomings of the review are evident in the limited search strategy, in which trials from only two search engines are indexed (MEDLINE and CINAHL). This is contrary to the Cochrane musculoskeletal group guidelines, which recommend that a minimum database search of MEDLINE, EMBASE and CENTRAL is conducted in order to ensure the identification of as many relevant studies as possible (Maxwell et al 2006). In addition, included articles were limited to English only, introducing the limitation of language bias (Higgins & Green 2008).

Contrary to the findings of White et al (2007) and Selfe & Taylor (2008), a meta-analysis by Manheimer et al (2007) concluded that only clinically irrelevant short term benefits of acupuncture were found in sham-controlled trials of OA knee. However, some clinically relevant benefits were highlighted in waiting-list and standard care controlled trials. The above reviews report literature searches only as recently 2006. This search period is similar to that used in the NICE guidelines for OA (NICE 2008), which report upon RCT’s from pre-2006 in addition to the meta-analysis of White et al (2007). The most recent systematic review by Manheimer et al (2010) also contains literature searching only as recently as 2007. In addition, this review includes articles pertaining to OA of the knee, hip or hand, rather than the knee joint in isolation. As several RCT’s have been published in this area over more recent years (Suarez-Almazor et al 2010, Miller et al 2009, Jubb et al 2008, Itoh et al 2008a & b), it appears timely to conduct a further systematic review of the literature, which may provide additional insight into the current evidence base for acupuncture in the treatment of OA knee. When utilised within a comprehensive review of the literature, in which a larger
number of more recent trials are appraised than have been done previously, a future research agenda can be identified. This will be a valuable tool when highlighting the need for further research within the field of acupuncture for OA knee, in addition to summarising trends and clinical outcomes from the literature which could be transferred to clinical practice.

1.1: Research Question

Does the current evidence support the use of acupuncture over sham acupuncture or standard care for treating the pain and functional impairment associated with OA of the knee joint?

1.2: Objectives

- To systematically appraise the current evidence base examining the short and long-term efficacy of acupuncture for improving pain or functional level in individuals with OA of the knee in trials comparing acupuncture with sham acupuncture or standard care.
- To ascertain whether the adequacy of the acupuncture used within the included trials has a bearing upon the clinical outcomes.
- To ascertain the safety of acupuncture within the included trials by recording information on any adverse effects reported in the studies.
2: Methodology & Methods

2.1: Methodology

The pace of research production means that making sense of an often contradictory evidence base has become increasingly difficult (Olhson, 1994). Therefore, collation and systematic interpretation of the evidence is required in order to summarise the state of the actual knowledge (MacPherson et al 2008). In light of the extensive evidence base available surrounding acupuncture and OA knee, a systematic review is an appropriate research methodology to answer the research question. A systematic review “attempts to collate all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question” (Higgins & Green 2008: 6). When the findings from individual studies are synthesised in an unbiased way, a balanced and impartial summary of the findings can be given with due consideration of any flaws within the evidence base (Hemingway & Brereton 2009).

The methodology of a systematic review lies within the positivist view of research. (Tranfield, Denyer & Smart 2003). The positivist researcher seeks cause and effect laws to ensure that knowledge of prior events enables a reasonable prediction of a subsequent event (Noblit & Hare, 1988). The critical rationalism approach to positivism attempts to acknowledge theoretical concepts and tests their adequacy by a process of attempted falsification (Sim & Wright 2000: 9). This process involves hypothetico-deductive logic, in which theories are rigorously tested against observations in a process of trial and error. Theories are accepted as an explanation of reality if they cannot be falsified through this method (Sim & Wright 2000). Systematic reviews sit comfortably within this paradigm, and the underpinning philosophy can be used to guide the research design.
2.2 Methods

2.2.1: Search strategy
CENTRAL, MEDLINE, EMBASE, AMED, CINAHL, PubMed and PEDro databases were searched from inception. Searching took place between January and February 2011. These databases are thought to be the most appropriate for the present subject, as CENTRAL, MEDLINE and EMBASE are recognised by the Cochrane collaboration as being the three databases considered the most important sources to search for reports of clinical trials (Higgins & Green 2008). In addition, the comprehensive searching of multiple databases has been shown to minimise publication bias (Akobeng 2005b). Subject-specific databases were also utilised, including AMED (Allied and Complementary Medicine), as this database focuses specifically on the subject of alternative medicine and is therefore likely to yield appropriate articles on the subject of acupuncture. PEDro is a Physiotherapy-specific evidence database, and CINAHL (the Cumulative Index to Nursing and Allied Health Literature) is a comprehensive resource for allied health literature. As such they are both likely to yield articles covering the subject matter of this search. PubMed was also searched, as it includes up-to-date citations not yet indexed for MEDLINE (Higgins & Green 2008).

The databases were searched individually by the author rather than performing combined database searches, in order to yield the maximum number of appropriate articles. When developing a search strategy, it is necessary to strike a balance between striving for comprehensiveness and maintaining relevance. Increasing the comprehensiveness (or sensitivity) of a search will reduce its precision and will retrieve more non-relevant articles (Higgins & Green 2008). For this reason, both MeSH (medical subject heading) descriptors (or alternative subject headings specific to the database searched, such as AMED Indexing terms) and key word searches were performed. Key words and appropriate synonyms were
searched for within the title and abstract of the articles. Hand searching of the reference lists of retrieved articles was also conducted and all RCT’s included in previous systematic reviews of acupuncture for OA knee were considered for inclusion. However, extensive hand searching of individual journals was not performed due to the resource implications of this thorough search method. In addition, non-English articles were not included (see exclusion criteria and justification in section 2.2.2). Grey literature, not formally published in sources such as books or journals has been shown to be an important source of studies for systematic reviews, in order to obtain the maximum number of appropriate articles and to minimise publication bias (Higgins & Green 2008). However, due to the resource implications of performing extensive grey literature searching, this has not been included. It is acknowledged that this, in addition to the exclusion of non-English articles, influences the publication bias of the review (Akobeng 2005b). This is discussed further in section 5.1.

The search strategy has been structured using the PICO method (Population, Intervention, Comparison, Outcome). However, the Cochrane group highlight that it is unnecessary and often undesirable to search for every PICO aspect, as these concepts may not be well described in the title or abstract of an article (Higgins & Green 2008). Therefore, MEDLINE searches for Cochrane reviews typically have three sets of search terms (Higgins & Green 2008: 129).

1. Terms to search for the Population of interest. In this instance, as OA knee is the group of interest, the two terms were searched for separately and then combined for a more comprehensive result.

2. Terms to search for the Intervention of interest. In this instance, the search aims to yield results for all types of acupuncture involving the insertion of needles.
3. A ‘filter’ for Controlled clinical trials or RCT’s. The Cochrane highly sensitive search strategy for identifying randomised trials in MEDLINE was utilised (Higgins & Green 2008). This filter was not utilised when searching CENTRAL, as this database contains only RCT’s or ‘controlled clinical trial’ publication types, indexed as human studies (Higgins & Green 2008).

The following search terms (table 1) were included for the MEDLINE search strategy, and adapted accordingly for the other search engines (appendix 1).
<table>
<thead>
<tr>
<th>Search No.</th>
<th>Search Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>(MH “Acupuncture Therapy+” )</td>
</tr>
<tr>
<td>S2</td>
<td>AB acupuncture or AB electroacupuncture or AB electro-acupuncture or AB acupoint or AB acu-point</td>
</tr>
<tr>
<td>S3</td>
<td>AB auriculotherapy or AB ear acupuncture</td>
</tr>
<tr>
<td>S4</td>
<td>AB medicine, oriental traditional or AB medicine, Chinese traditional</td>
</tr>
<tr>
<td>S5</td>
<td>S1 OR S2 OR S3 OR S4</td>
</tr>
<tr>
<td>S6</td>
<td>(MH “Arthritis+”)</td>
</tr>
<tr>
<td>S7</td>
<td>(MH “Osteoarthritis+”) OR (MM Osteoarthritis, Knee”)</td>
</tr>
<tr>
<td>S8</td>
<td>(MH “Joint Diseases+”)</td>
</tr>
<tr>
<td>S9</td>
<td>(MH “Arthralgia+”)</td>
</tr>
<tr>
<td>S10</td>
<td>AB joint pain or AB gonarthrosis or AB gonarthritis or AB osteoarthritis or AB degenerative arthritis or AB joint disease or AB arthralgia or AB arthritis</td>
</tr>
<tr>
<td>S11</td>
<td>S6 OR S7 OR S8 OR S9 OR S10</td>
</tr>
<tr>
<td>S12</td>
<td>(MH “Knee joint+”)</td>
</tr>
<tr>
<td>S13</td>
<td>(MM “Knee”) OR (MM “Arthroplasty, Replacement, Knee”) OR (MM “Osteoarthritis, Knee”)</td>
</tr>
<tr>
<td>S14</td>
<td>AB knee or AB knee joint or AB knees</td>
</tr>
<tr>
<td>S15</td>
<td>S12 OR S13 OR S14</td>
</tr>
<tr>
<td>S16</td>
<td>PT Clinical trial</td>
</tr>
<tr>
<td>S17</td>
<td>AB randomized or PT randomized controlled trial or PT controlled clinical trial</td>
</tr>
<tr>
<td>S18</td>
<td>AB placebo</td>
</tr>
<tr>
<td>S19</td>
<td>(MH “Clinical Trial+”)</td>
</tr>
<tr>
<td>S20</td>
<td>AB randomly</td>
</tr>
<tr>
<td>S21</td>
<td>TI trial</td>
</tr>
<tr>
<td>S22</td>
<td>S16 OR S17 OR S18 OR S19 OR S20 OR S21</td>
</tr>
<tr>
<td>S23</td>
<td>Animals [mh] NOT humans [mh]</td>
</tr>
<tr>
<td>S24</td>
<td>S22 NOT S23</td>
</tr>
<tr>
<td>S25</td>
<td>S5 AND S11 AND S15 AND S24</td>
</tr>
</tbody>
</table>
2.2.2: Study Selection

For the purposes of this review, the criteria for trial inclusion was addressed using the “PICO” framework. The inclusion and exclusion criteria have been summarised in table 2. Article screening against the eligibility criteria was performed by the author. Article titles and abstracts were screened initially. Full-text articles were then obtained for potentially relevant studies and studies with insufficient information within the abstract to allow screening against the eligibility criteria.

---

Medline Search Syntax

MM = exact major subject heading (MeSH term)
MH = exact subject heading (a MeSH term ‘exploded’)
+ (after search term) = MeSH term ‘exploded’
AB = a word in the abstract
TI = a word in the title
PT = a Publication Type term
<table>
<thead>
<tr>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population:</strong></td>
<td><strong>Population:</strong></td>
</tr>
<tr>
<td>Patient population with a diagnosis of primary OA knee from either clinical features and / or radiographical findings</td>
<td>Knee pathology other than primary OA</td>
</tr>
<tr>
<td>Studies of multi-joint OA included only if the results for the participants with OA knee are reported separately</td>
<td>Post-knee replacement patients</td>
</tr>
<tr>
<td>Randomised controlled trials</td>
<td>Non- randomised controlled trial design</td>
</tr>
<tr>
<td>Human studies</td>
<td>Animal studies</td>
</tr>
<tr>
<td>Articles published in English</td>
<td>Articles published in languages other than English</td>
</tr>
<tr>
<td><strong>Intervention:</strong></td>
<td><strong>Intervention:</strong></td>
</tr>
<tr>
<td>Intervention group with acupuncture techniques involving the insertion of needles</td>
<td>Non-needle penetration acupuncture approaches, e.g. non-needle penetrating moxa, acupressure, laser, electrical stimulation without needle insertion</td>
</tr>
<tr>
<td>Either manual or electro-acupuncture techniques</td>
<td>The addition of chemicals to the acupuncture procedure (e.g. bee venom acupuncture)</td>
</tr>
<tr>
<td>Traditional acupuncture points used (can be used in addition to Ashi or trigger points)</td>
<td>No traditional acupuncture points used</td>
</tr>
<tr>
<td>Observation period of at least 6 weeks (including combined intervention and follow up period)</td>
<td>Observation periods of less than 6 weeks</td>
</tr>
<tr>
<td><strong>Comparison:</strong></td>
<td><strong>Comparison:</strong></td>
</tr>
<tr>
<td>Comparison group with either sham acupuncture or standard care</td>
<td>Trials in which one form of acupuncture is compared only with another form</td>
</tr>
<tr>
<td><strong>Outcome:</strong></td>
<td><strong>Outcome:</strong></td>
</tr>
<tr>
<td>At least one validated primary outcome measures of knee pain and / or function</td>
<td>Trials in which acupuncture is compared only with waiting list controls or other interventions other than sham or standard care</td>
</tr>
<tr>
<td></td>
<td>Absence of any validated outcome measure of knee pain or function</td>
</tr>
<tr>
<td></td>
<td>No data reported</td>
</tr>
</tbody>
</table>
The above inclusion and exclusion criteria have been chosen in order to limit the trials included to those which are most likely to achieve the review’s objectives (Maxwell et al 2006). The exclusion criteria of ‘knee pathology other than primary OA’ and ‘post-knee replacement patients’, ensure specificity to trials examining acupuncture’s effects on OA of the knee. Knee arthroplasty and replacement are, however included in the search strategy in order to capture articles examining the effects of acupuncture on patients awaiting the procedure. The inclusion of RCT’s is specified, as such trials are likely to include the most rigorous methodology (Kunz, Vist & Oxman 2007). Although non-English articles would ideally be included in order to obtain the most comprehensive view of the existing research and to minimise publication bias (Akobeng 2005b), they have been excluded due to the resource requirements of obtaining article translation.

Excluding non-needle penetration approaches ensures specificity to trials examining needle acupuncture, which is shown to have different physiological effects than methods such as non-needle penetrating moxa or acupressure (Stux, Berman & Pomeranz 2003). Both manual and electroacupuncture techniques are included, as it has been shown that these methods produce sufficient intensity of needling to elicit DeQi (Sandberg et al 2003, Hui et al 2007). Trials in which only two types of acupuncture are compared are excluded, as are comparisons with only waiting list controls or other active interventions, as comparisons with sham acupuncture or standard care are required to address the review’s objectives. In addition, trials in which only non-traditional acupuncture techniques are used (such as trigger point or periosteal acupuncture at non-meridian points) are omitted, in line with the criteria of Manheimer et al (2010). This criterion helps to ensure that similar interventions are utilised in the included RCT’s and can therefore be compared more effectively. A minimum observation period of six weeks was chosen due to shorter observation periods seeming irrelevant to the
question of whether acupuncture is helpful in the management of a chronic condition such as OA (Manheimer et al, 2007). Multiple publications of the same study have been omitted in order that each study population is represented only once (Ezzo et al 2001). Trials in which no data is reported are excluded, as no analysis could subsequently be performed to meet the objectives of the review (Higgins & Green 2008). However, attempts were made to contact authors for missing data in these instances.

Only trials which utilise a validated outcome measure are included, as trials utilising non-validated outcome measures may report upon non-clinically relevant improvements (Tubach et al 2005). Many trials include multiple outcome measures, such as the Short-form 36 (SF-36) and the Western Ontario and McMaster Universities OA Index (WOMAC). Both are valid and reliable outcome measures for knee pain and function (Lingard et al 2001, Bellamy et al 1997). However, they have been shown to measure similar outcomes of function, pain and quality of life (Lingard et al 2001). Therefore, in line with Manheimer et al (2007), in instances where both outcomes are used in the same study, preference is given to reporting results of the WOMAC, as it is the most widely-used and thoroughly validated instrument for assessing patients with OA knee (Bellamy et al 1997).

The WOMAC index is a self-reported outcome measure consisting of 24 items divided into 3 subscales: 5 pain, 2 stiffness and 17 physical function items. These three subscale scores can be combined to give a total WOMAC index score, giving an overall representation of knee disability and function (Kersten, White & Tennant 2010). Higher scores on the WOMAC indicate more severe pain, stiffness and functional limitations (Ackerman 2009). The WOMAC index is available in 5-point Likert, 11-point numerical rating and 100mm Visual Analogue Scale (VAS) formats. Often, the score is transformed into a 0-100 scale for ease of
interpretation and comparison with other studies (Ackerman 2009). The WOMAC index has been shown to be a valid and reliable outcome measure for the disability and functional impairment associated with OA of the knee (Bellamy et al 1997, Faucher et al 2004). Total WOMAC index scores have therefore been categorised as functional outcome scores for the purposes of this review.

The Knee Society Score (KSS) is a duel rating system of knee pain and function, utilised by Miller et al (2009). This is a clinician derived outcome measure, consisting of two scores; one for knee pain and one for function (Reddy et al 2010). The KSS for pain and function have been found to be valid outcome measures for knee OA. The two scales have demonstrated strong correlations with the corresponding domains of the WOMAC index (Lingard et al 2001). Other functional outcome measures include the Oxford Knee Score (OKS), used by Williamson et al (2007). The OKS is a self-reported outcome measure consisting of 12 questions assessing pain and physical disability on a 5-point Likert scale. This yields a single score ranging from a best functional outcome of 12, to a worst functional outcome of 60 (Xie et al 2011). The OKS is considered a valid and reliable functional outcome measure for patients with OA knee (Xie et al 2011).

Other outcome measures examine pain rather than function. The pain VAS has been included in the trial of Vas et al (2004). The VAS 0-100mm score is a simple and frequently-used self-assessment method for variation in intensity of pain, which has been shown to be reliable, valid and appropriate for use in clinical practice (Williamson & Hoggart, 2005). The Joint-Specific Multidimensional Assessment of Pain (J-MAP) is also utilised by Suarez-Almazor et al (2010). The J-MAP includes 6-item pain sensory and 4-item pain affect subscales, with higher scores indicating more pain (O’Malley et al 2003). This has been shown to be a

Outcome measures such as the Profile of Quality of Life in the Chronically Ill, used by Vas et al (2004) are excluded due to a lack of validity and reliability studies on its use. The 6-minute timed walk, used by Berman et al (2004) has also been excluded. Although this instrument has been validated for a cardiac rehabilitation population (Hamilton & Haennel 2000), it is yet to be validated as a measure of knee pain or function.

2.2.3: Data Extraction

Data collection forms were piloted on a sample of studies before devising the final forms, in line with the Cochrane Musculoskeletal group recommendations (Maxwell et al 2006). Data pertaining to methodological design and quality, participants, acupuncture and control interventions and treatment outcomes, including adverse effects was then collected from studies and recorded on the data collection forms, as shown in appendix 2.

The majority of data extraction was possible through analysis of the study reports. However, authors were contacted for clarification on any specific aspects of the methodology which were unclear. Information on adverse effects was also recorded in line with Cochrane Adverse Effects Subgroup recommendations (Higgins & Green 2008). Assessment of the adequacy of acupuncture treatment in each trial was also assessed by the author using the criteria below, in order to ascertain whether this factor was influential upon the outcomes of the included trials.
2.2.4: Adequacy of Acupuncture

In clinical practice, the majority of acupuncture treatment is individualised to the patient’s changing symptoms. In contrast, clinical trials of acupuncture often use a prescribed formula of acupuncture points to allow for the best estimation of the specific effects of acupuncture (Manheimer et al 2010). There is no consensus on how best to assess the adequacy of acupuncture treatment within acupuncture trials (Birch 2004). However, it is acknowledged that examination of the adequacy of acupuncture treatments is important in order to correctly assess trials and to develop improved protocols for future research (White & Ernst 1998). Cochrane trials often utilise independent assessment of the adequacy of acupuncture within RCT’s by experienced acupuncture practitioners, incorporating subjective assessment of number and duration of treatments, choice of acupuncture points and needling technique (Manheimer et al 2010).

Acupuncture treatment ‘dose’ is a contentious issue, and no conclusive evidence has yet recommended an optimum needling duration or number of treatments in order to optimise clinical effects (White et al 2008). Lundeberg et al (1988) conducted a large-scale crossover study to establish optimum treatment parameters for acupuncture. Conclusions were drawn that a minimum of 30 minutes treatment duration over 10 sessions was optimal. However, the practicalities of clinical practice often dictate the use of 20 minute treatment durations over a minimum of 6 sessions. This criterion, in addition to a minimum of once-weekly treatment, has been utilised in the systematic review of White et al (2007) as a component of ‘adequate’ acupuncture. Manheimer et al (2010) also included the criterion of at least a 6 week minimum acupuncture course duration, through the consensus agreement of two experienced acupuncturists. Eliciting DeQi has also been shown to be clinically important for achieving acupuncture analgesic effects (Hui et al 2007). The experience of the acupuncture practitioner
has not been shown to have a significant impact on treatment outcomes (Witt et al 2010). There is little evidence as to the optimum needling locations for the treatment of specific conditions. However, the use of points local to the painful area, often in addition to distal points are thought to be useful in the treatment of specific joint pain (Deadman, Al-Khafaji & Baker 2007).

In light of these findings, criteria for adequate acupuncture within the included trials was defined as shown in table 3.

**Table 3: Criteria for Adequacy of Acupuncture**

<table>
<thead>
<tr>
<th>Criterion number</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A minimum ‘dosage’ of weekly treatments of 20 minutes duration each</td>
</tr>
<tr>
<td>2</td>
<td>A total course of acupuncture treatment of at least 6 weeks</td>
</tr>
<tr>
<td>3</td>
<td>Eliciting a DeQi response or strong needle sensation in the recipient</td>
</tr>
<tr>
<td>4</td>
<td>Acupuncture points local to the painful joint utilised, either in isolation or in addition to distal points.</td>
</tr>
</tbody>
</table>

**2.2.5: Internal Validity of Trials**

The internal validity of included trials was analysed using the PEDro scale (appendix 3). This is an 11 item scale, originally based on the Delphi list quality assessment criteria (Verhagen et al 1998). However, it has been adapted to include criteria to assess measurements of key outcomes and reporting of between-group statistical comparisons in RCT’s (PEDro 2010). The PEDro scale has been chosen due to its reliable and relatively unambiguous scoring system for RCT’s (Maher et al 2003). This appraisal system also includes a criterion for assessment of allocation concealment. This is a beneficial inclusion within a quality
assess the methodological quality of all papers was independently assessed by two reviewers; the author and a fellow student of the Acupuncture MSc course at Coventry University. Disagreements were resolved through consensus, and in the one instance where resolution could not be achieved, a third reviewer (a Chartered Physiotherapist qualified in the use of acupuncture), made an independent assessment of the relevant trial. This independent analysis of the studies was intended to minimise reviewer bias (Akobeng 2005b) and subjectivity which might affect the scoring system (Maxwell et al 2006). However, it is recognised that reviewer bias cannot be eliminated completely due to acupuncture being the specialised area of interest of both assessors. To minimise the risk of reviewer bias, it is also recommended that the assessors are blinded to the authors under review (Higgins & Green 2008). The second reviewer was therefore blinded to the RCT authors. However this was impractical for the first reviewer due to knowledge of the articles obtained during the article screening process.

The PEDro scale helps to identify which studies are likely to be internally valid (items 2-9) and have sufficient statistical information to make their results interpretable (items 10-11) (Sherrington et al 2000). It was decided that articles scoring less than 6/10 on the PEDro scale would not be included in the subsequent review as the findings of lower quality studies have been shown to exaggerate treatment efficacy and are less likely to be valid (Moher et al 1998).
The two articles scoring less than 6/10 were therefore excluded at this stage (Itoh et al 2008b, Christensen et al 1992).

2.2.6: Data Synthesis

The included studies were deemed to be heterogeneous, as they differ considerably in terms of the comparisons used and outcomes measured. It was therefore not considered appropriate to pool the results for meta-analysis (Higgins & Green 2008). A best evidence synthesis of the results of the included RCT’s was therefore performed (Slavin 1995). This analysis incorporates an outcomes assessment similar to that used by Ezzo et al (2001). This categorises outcomes as positive, neutral or negative, as shown in table 4. P-values of <0.05 were considered significant (Ezzo et al 2001).

Table 4: Outcomes Assessment

<table>
<thead>
<tr>
<th>Outcome assessment</th>
<th>Criteria</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>when acupuncture had a significantly better effect than the control group</td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>when the effect of acupuncture was not statistically different from that seen in the control group</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>When the effect of acupuncture was significantly less than that seen in the control group</td>
<td></td>
</tr>
</tbody>
</table>

Quality assessment and outcome assessment scores for studies were analysed within two primary categories:

1. Trials comparing acupuncture with sham acupuncture treatments
2. Trials comparing acupuncture with standard care
Sham acupuncture treatment was defined as a mock acupuncture intervention, intended to be virtually physiologically inert and credible as the active treatment (Manheimer et al 2007). Standard care was defined as the usual care recommended in the first tier of the core treatment options of the NICE Guidelines for the Care and Management of OA in Adults (NICE, 2008), including advice, education, standardised exercises and self management. Standard pharmacological management such as NSAID’s could also be included in this category. However other treatments such as Physiotherapy (including manual therapy, electrotherapy and progressive, individualised exercise sessions) and intra-articular injections were excluded, as they were considered active treatments rather than standard care. Waiting-list controls were also excluded from the review. This was defined as groups which received no other additional care while awaiting acupuncture treatment (Manheimer et al 2007). Assessment of the adequacy of acupuncture reported in each study was also assessed according to the previously specified criteria (table 3).

Finally, the results were weighted according to their methodological quality, using the four levels of scientific evidence described by the Cochrane Collaboration Back Review Group (van Tulder et al 2003), shown in table 5.

**Table 5: Levels of Evidence**

<table>
<thead>
<tr>
<th>Level of Evidence</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>Multiple, relevant, high quality RCT’s with generally consistent outcomes</td>
</tr>
<tr>
<td>Moderate</td>
<td>1 relevant, high quality RCT and 1 or more relevant, low quality RCT’s with generally consistent outcomes</td>
</tr>
<tr>
<td>Limited</td>
<td>1 relevant, high quality RCT or multiple relevant, low-quality RCT’s with generally consistent outcomes</td>
</tr>
<tr>
<td>Inconclusive</td>
<td>Only 1 relevant, low-quality RCT, no relevant RCT’s or RCT’s with inconsistent outcomes</td>
</tr>
</tbody>
</table>
For the purposes of this review, ‘relevant’ is defined as meeting the criteria for inclusion, ‘generally consistent’ is defined as two thirds or more of the results having the same outcome result (positive, negative or neutral), and ‘multiple’ is defined as two or more RCT’s (Ezzo et al 2001). The classification of high or low quality RCT’s was decided through a combination of analysing the PEDro scores in addition to critically appraising the articles individually.

3: Results

The literature search process generated 129 potentially relevant studies for screening. Following screening, 13 studies were obtained for PEDro scoring. Excluded articles which it would be plausible for the reader to assume might have been included in the review have been tabulated alongside reasons for exclusion (appendix 4). In one instance, an article was excluded as no outcome data was reported (Itoh et al 2008a). The authors were contacted requesting the missing data but no response was received. Studies which clearly do not meet the inclusion criteria, such as those published in other languages or those clearly not addressing the target population have not been tabulated due to the high volume of articles yielded, in line with the recommendations of the Cochrane Collaboration (Higgins & Green 2008). Articles achieving a PEDro score of less than 6/10 were then excluded (Christensen et al 1994, Itoh et al 2008b), leaving 11 studies involving a total of 5851 participants for inclusion in the review. The study selection process has been tabulated in figure 1 below:
3.1: Internal Validity

The internal validity of the studies was assessed using the PEDro scale and the detailed results are shown in appendix 5. Only two of the included studies scored less than 6/10 on the PEDro scale (Christensen et al 1992, Itoh et al 2008b). These studies scored 5/10 and 3/10 respectively. The total PEDro scores for each article are shown alongside the characteristics of the included studies in table 6.
<table>
<thead>
<tr>
<th>Study Author &amp; Year</th>
<th>Participants (n)</th>
<th>Mean age (years)</th>
<th>Study Design</th>
<th>Intervention &amp; control groups</th>
<th>Co-intervention</th>
<th>Outcome Measures of pain &amp;/or function</th>
<th>Measurement time points (weeks)</th>
<th>Acupuncture</th>
<th>Acupuncture ‘dose’</th>
<th>Adequate acu?</th>
<th>PEDro score /10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berman et al 1999</td>
<td>73</td>
<td>65.6</td>
<td>TCA + Standard care Vs Standard care alone</td>
<td>Standard Care</td>
<td>WOMAC index &amp; pain subscale</td>
<td>0, 4, 8, 12</td>
<td>Formula with EA</td>
<td>20 mins, 2x weekly for 8/52</td>
<td>Yes</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Berman et al 2004</td>
<td>570</td>
<td>65.5</td>
<td>TCA Vs Sham acu – combined insertion (at non-acu points) and non-insertion - both with mock EA) Vs education (Standard Care)</td>
<td>-</td>
<td>WOMAC pain &amp; function subscales</td>
<td>0, 4, 8, 14, 26</td>
<td>Formula with EA</td>
<td>20 mins 2X weekly for 8/52, 1X weekly for 2/52, Then 1X fortnightly for 4/52</td>
<td>Yes</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Foster et al 2007</td>
<td>352</td>
<td>63.2</td>
<td>TCA Vs Sham acu (Non-penetrating needling of same points) Vs Advice, education &amp; individualised exercises control group</td>
<td>Advice &amp; education</td>
<td>WOMAC pain &amp; function subscales</td>
<td>0, 2, 6, 26, 52</td>
<td>Flexible Formula</td>
<td>25-35 mins, up to 6X over 3/52</td>
<td>No</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Jubb et al 2008</td>
<td>68</td>
<td>65.1</td>
<td>TCA Vs sham acu (non-penetrating needles with mock EA at same acu points)</td>
<td>-</td>
<td>WOMAC pain subscale</td>
<td>0, 5, 9</td>
<td>Formula with EA</td>
<td>30 mins, 2X weekly for 5/52</td>
<td>Yes</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Miller et al 2009</td>
<td>55</td>
<td>71.3</td>
<td>TCA Vs Sham acu (non-penetrating acu at same points)</td>
<td>Standard care</td>
<td>KSS function and pain scores</td>
<td>0, 8, 12</td>
<td>Formula</td>
<td>20 min, 2X weekly for 8/42</td>
<td>Yes</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Scharf et al 2006</td>
<td>1007</td>
<td>62.8</td>
<td>TCA Vs Sham acu (minimal needling at non-acu points) Vs Conservative therapy (standard care)</td>
<td>Diclofenac and up to 6 physio sessions</td>
<td>WOMAC index</td>
<td>0, 13, 26</td>
<td>Flexible formula</td>
<td>20-30 mins, 10 sessions over 6/52</td>
<td>Yes</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>N</td>
<td>Age (years)</td>
<td>Group Description</td>
<td>Duration of Treatment</td>
<td>Outcome Measures</td>
<td>Randomisation Method</td>
<td>Quality Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----</td>
<td>-------------</td>
<td>------------------------------------------------------------------------------------</td>
<td>-----------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>-------------------------</td>
<td>--------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suarez-Almazor et al 2010</td>
<td>527</td>
<td>64.4</td>
<td>TCA Vs Sham acu (minimal needling at non-acu points) Vs Waiting list control</td>
<td>0, 4, 6, 12</td>
<td>WOMAC pain &amp; function subscales</td>
<td>Yes</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vas et al 2004</td>
<td>97</td>
<td>67.1</td>
<td>TCA Vs sham acu (non-penetrating needles with mock EA at same points)</td>
<td>0, 12</td>
<td>Formula with EA</td>
<td>Not stated</td>
<td>Unclear</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Williamson et al 2007</td>
<td>181</td>
<td>70.7</td>
<td>TCA Vs Physiotherapy (6 weekly group exercise circuits) Vs Standard care control</td>
<td>0, 7, 12 (and 3/12 post-operatively)</td>
<td>OKS</td>
<td>Yes</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Witt et al 2005</td>
<td>294</td>
<td>63.8</td>
<td>TCA Vs Sham acu (superficial needling at non-acu points) Vs Waiting list control</td>
<td>0, 8, 26, 52</td>
<td>Formula with EA</td>
<td>Yes</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Witt et al 2006</td>
<td>2627</td>
<td>61.5</td>
<td>TCA Vs Standard care control Vs nonrandomised acu group</td>
<td>0, 12, 24</td>
<td>Individualised</td>
<td>Unclear</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Key**
- TCA = Traditional Chinese Acupuncture
- EA = Electroacupuncture
- OKS = Oxford Knee Score
- Rx = Treatment
- KSS = Knee Society Score
- Acu = Acupuncture
- /52 = Weeks
- J-MAP = Joint-Specific Multidimensional Assessment of Pain
- WOMAC = Western Ontario and McMaster Universities OA Index
- TP = Trigger point
- VAS = Visual Analogue Scale
3.2: Acupuncture Protocol

The majority of studies utilised a set acupuncture point formula. Four studies utilised a flexible formula, where most points were mandatory, but some could be adapted or chosen at the discretion of the practitioner (Witt et al 2005, Scharf et al 2006, Foster et al 2007, Williamson et al 2007). Only one study (Witt et al 2006) conducted entirely individualised acupuncture treatments, with acupuncture points selected by the practitioner for each patient on an individual basis. Six studies utilised electro-acupuncture treatment (Berman et al 1999, Berman et al 2004, Vas et al 2004, Jubb et al 2008, Suarez-Almazor et al 2010).

3.3: Controls


education control group, involving six individualised, progressive exercise sessions and an advice leaflet.

3.4: Outcome Measures

Six studies utilised the WOMAC index as one of their outcome measures (Berman et al 1999, Berman et al 2004, Vas et al 2004, Witt et al 2005, Scharf et al 2006 and Witt et al 2006). The WOMAC function subscale is reported as a separate outcome measure in three studies (Berman et al 2004, Foster et al 2007, Suarez-Almazor et al 2010). This subscale is scored from 0-68, with higher scores indicating more severe functional limitations (Stratford & Kennedy 2004). The OKS is used as a functional outcome measure by Williamson et al (2007) and the KSS is used as a dual rating of knee pain and function by Miller et al (2009).

Other outcome measures examine pain levels only, including the WOMAC pain subscale, reported in seven of the included trials (Berman et al 1999, Berman et al 2004, Vas et al 2004, Witt et al 2006, Foster et al 2007, Jubb et al 2008, Suarez-Almazor et al 2010). This is usually scored from 0-20, with higher scores representing more severe pain (Stratford & Kennedy 2004). As studies differ regarding the format of the WOMAC pain subscale used (e.g. 0-20 Likert format, 0-100 or 0-500 VAS formats), the mean change scores of other scales have been converted into 0-20 scales for ease of comparison. The pain VAS is also utilised by Vas et al (2004).

3.5: Acupuncture Adequacy

The adequacy of acupuncture treatment used in each study was assessed according to the pre-defined criteria. If any criterion scored ‘no’, acupuncture was defined as being inadequate. Results of acupuncture adequacy are shown in table 7.
## Table 7: Acupuncture Adequacy

<table>
<thead>
<tr>
<th>Article (1st author &amp; year)</th>
<th>Min 20 min Rx</th>
<th>At least weekly?</th>
<th>At least 6 week Rx course?</th>
<th>DeQi / strong needle sensation?</th>
<th>Pts local to painful joint?</th>
<th>Acupuncture Adequate?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berman et al 1999</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Berman et al 2004</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Foster et al 2007</td>
<td>Yes</td>
<td>Yes</td>
<td>No (3 weeks)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Jubb et al 2008</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Miller et al 2009</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Scharf et al 2006</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Suarez-Almazor et al 2010</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Vas et al 2004</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Unclear</td>
</tr>
<tr>
<td>Williamson et al 2007</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Witt et al 2005</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Witt et al 2006</td>
<td>Unclear</td>
<td>Yes (assumed, as up to 15 sessions over 3/12)</td>
<td>Yes</td>
<td>Unclear</td>
<td>Unclear (specific points not recorded)</td>
<td>Unclear</td>
</tr>
</tbody>
</table>

**Key**
- Acu = Acupuncture
- Rx = Treatment
- /12 = months

One study did not meet the pre-defined ‘adequate’ criteria (Foster et al 2007). Two studies were not able to be fully assessed for adequacy of acupuncture due to insufficient documentation of the acupuncture procedure and dosage (Vas et al 2004, Witt et al 2005).
Unsuccessful attempts were made to contact both lead authors of these studies to clarify the acupuncture dosage used.

3.6: Safety
Nine authors described specific adverse or side-effects during their trials (Berman et al 2004, Vas et al 2004, Witt et al 2005, Witt et al 2006, Scharf et al 2006, Foster et al 2007, Williamson et al 2007, Jubb et al 2008, Suarez-Almazor et al 2010). Five authors reported the minor adverse or side effects within their trials to be treatment-related (Vas et al 2004, Witt et al 2006, Foster et al 2007, Williamson et al 2007, Suarez-Almazor et al 2010). Four authors recorded serious adverse effects (Berman et al 2004, Witt et al 2005, Scharf et al 2006, Jubb et al 2008). However, the frequency of serious adverse events was similar between acupuncture and control groups and no serious adverse effects were reported to be associated with acupuncture. Details of adverse and side-effects are shown in appendix 6.

3.7: Effects of Interventions
The effects of interventions are shown for acupuncture Vs sham and acupuncture Vs standard care in tables 8-11 below. Primary outcomes for pain and / or function have been extracted and short and long-term outcomes have been shown. For the purposes of this review, a short-term outcome measure has been defined as the measurement point closest to 6 weeks, but less than 12 weeks after randomisation. A long-term follow-up is 12 or more weeks after randomisation, and the measurement point closest to 24 weeks has been chosen (Manheimer et al 2007). P-values and confidence intervals are given where available. The reported mean difference between baseline and endpoint scores with standard deviation has been used where available, or otherwise mean differences have been calculated from baseline. Where the
standard deviation of the mean change in outcome measures is available, the effect size has
been calculated using Cohen’s D (Cohen 1988).

Outcomes: Acupuncture Vs Sham

Eight studies compare acupuncture with sham acupuncture. Of these, six utilise an outcome
et al 2009, Suarez-Almazor et al 2010). Four utilise the WOMAC pain subscale (table 8a),
including four short-term and three long-term measurement points. Other pain outcomes
include the KSS pain score (table 8b) used by Miller et al (2009), the VAS pain score (table
8c) used by Vas et al (2004), and the J-MAP (table 8d) used by Suarez-Almazor et al (2010).

Seven studies incorporate a functional outcome measure (Berman et al 2004, Vas et al 2004,
2010). Of these, three utilise the WOMAC index (table 9a), with one short term and three
long-term measurement points. Three studies reported the WOMAC function subscale (table
9b), each with short and long-term measurements and the KSS function score was used by
### Acupuncture Vs Sham: Pain Outcomes

#### Table 8a: WOMAC Pain Subscales (0-20 scale)

<table>
<thead>
<tr>
<th>Reference (1st Author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (wks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change sham group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foster 2007</td>
<td>352</td>
<td>6</td>
<td>2.83 (4.0)</td>
<td>3.02 (3.6)</td>
<td>-0.19</td>
<td>-0.19 (-1.18 to 0.79)</td>
<td>P=0.701</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
<tr>
<td>Jubb 2008</td>
<td>68</td>
<td>5</td>
<td>3.8 (3.84)</td>
<td>1.4 (4.84)</td>
<td>2.4</td>
<td>2.4 (0.2 to 4.64)</td>
<td>P=0.035</td>
<td>Positive</td>
<td>0.63</td>
</tr>
<tr>
<td>Suarez-Almazor 2010</td>
<td>527</td>
<td>6</td>
<td>3.27</td>
<td>3.22</td>
<td>0.05</td>
<td>Not given</td>
<td>P&gt;0.20</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
<tr>
<td>Berman 2004</td>
<td>570</td>
<td>8</td>
<td>3.15 (0.29)</td>
<td>2.66 (0.26)</td>
<td>0.49</td>
<td>0.49 (0.2 to 1.2)</td>
<td>P=0.18</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
</tbody>
</table>

#### Short-Term Outcomes (<12 weeks)

#### Long-Term Outcomes (≥ 12 weeks)

<table>
<thead>
<tr>
<th>Reference (1st Author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (wks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change sham group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foster 2007</td>
<td>352</td>
<td>26</td>
<td>2.32 (3.6)</td>
<td>2.53 (4.2)</td>
<td>-0.21</td>
<td>-0.21 (-1.25 to 0.83)</td>
<td>P=0.692</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
<tr>
<td>Suarez-Almazor 2010</td>
<td>527</td>
<td>12</td>
<td>2.5</td>
<td>2.76</td>
<td>-0.26</td>
<td>Not given</td>
<td>P&gt;0.20</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
<tr>
<td>Berman 2004</td>
<td>570</td>
<td>26</td>
<td>3.79 (0.33)</td>
<td>2.92 (0.30)</td>
<td>0.87</td>
<td>0.87 (-0.16 to 1.58)</td>
<td>P=0.003</td>
<td>Positive</td>
<td>2.64</td>
</tr>
<tr>
<td>Vas 2004</td>
<td>97</td>
<td>12</td>
<td>10.7</td>
<td>5.7</td>
<td>5</td>
<td>Not given</td>
<td>P&lt;0.001</td>
<td>Positive</td>
<td>-</td>
</tr>
</tbody>
</table>
**Acupuncture Vs Sham: Pain Outcomes cont.**

Table 8b: KSS Pain Scores (0-50 scale)

<table>
<thead>
<tr>
<th>Reference (1st author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (wks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change sham group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miller 2009</td>
<td>55</td>
<td>8</td>
<td>7.4</td>
<td>7.1</td>
<td>0.3</td>
<td>Not given</td>
<td>Not given for mean change between groups</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8c: VAS Pain Scores (0-100mm scale)

<table>
<thead>
<tr>
<th>Reference (1st author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (wks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change sham group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vas 2004</td>
<td>97</td>
<td>12</td>
<td>48.3</td>
<td>23.1</td>
<td>25.2</td>
<td>Not given</td>
<td>P&lt;0.001</td>
<td>Positive</td>
<td>-</td>
</tr>
</tbody>
</table>

Long-Term Outcomes (≥12 weeks)
### Acupuncture Vs Sham: Pain Outcomes cont.

**Table 8d: J-MAP (0-7 scale)**

#### Short-Term Outcomes (<12 weeks)

<table>
<thead>
<tr>
<th>Reference (1st author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (wks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change sham group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suarez-Almazor 2010</td>
<td>527</td>
<td>6</td>
<td>1.3</td>
<td>1.0</td>
<td>0.3</td>
<td>Not given</td>
<td>P &gt; 0.20</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
</tbody>
</table>

#### Long-Term Outcomes (≥ 12 weeks)

| Suarez-Almazor 2010           | 527     | 12             | 1.1                                    | 0.9                                    | 0.2                           | Not given              | P > 0.20 | Neutral           | N/S                   |

### Acupuncture Vs Sham: Functional Outcomes

**Table 9a: WOMAC Index (0-100mm VAS scale)**

#### Short Term Outcomes (<12 weeks)

<table>
<thead>
<tr>
<th>Reference (1st author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (wks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change sham group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Witt 2005</td>
<td>294</td>
<td>8</td>
<td>23.9</td>
<td>16.7</td>
<td>7.2</td>
<td>Not given</td>
<td>P &lt; 0.001</td>
<td>Positive</td>
<td>-</td>
</tr>
</tbody>
</table>

#### Long-Term Outcomes (≥ 12 weeks)

| Scharf 2006                   | 1007    | 26             | 23                                     | 21                                     | 2.0                           | 2.0 (1.0 to 5.0)        | P = 0.48 | Neutral           | N/S                   |
| Vas 2004                      | 97      | 12             | 47.6                                   | 24.3                                   | 23.3                          | Not given              | P < 0.001 | Positive          | -                     |
| Witt 2005                     | 294     | 26             | 20.4                                   | 16.2                                   | 4.2                           | Not given              | P = 0.063 | Neutral           | N/S                   |
Acupuncture Vs Sham: Functional Outcomes cont.

Table 9b: WOMAC Function Subscale (Likert 0-68 format)

<table>
<thead>
<tr>
<th>Reference (1st author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (wks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change sham group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berman 2004</td>
<td>520</td>
<td>8</td>
<td>10.77 (0.90)</td>
<td>7.84 (0.76)</td>
<td>2.93</td>
<td>2.9 (-0.8 to 5.0)</td>
<td>P&lt;0.01</td>
<td>Positive</td>
<td>3.26</td>
</tr>
<tr>
<td>Foster 2007</td>
<td>352</td>
<td>6</td>
<td>8.18 (11.5)</td>
<td>9.32 (11.4)</td>
<td>-1.14</td>
<td>-1.14 (-4.16 to -1.88)</td>
<td>P=0.459</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
<tr>
<td>Suarez-Almazor 2010</td>
<td>527</td>
<td>6</td>
<td>13.35</td>
<td>13.2</td>
<td>0.15</td>
<td>Not given</td>
<td>P&gt;0.20</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
</tbody>
</table>

Long-Term Outcomes (≥ 12 weeks)

<table>
<thead>
<tr>
<th>Reference (1st author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (wks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change sham group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berman 2004</td>
<td>520</td>
<td>26</td>
<td>12.42 (1.12)</td>
<td>9.88 (0.93)</td>
<td>2.54</td>
<td>2.5 (4.7 to 0.4)</td>
<td>P&lt;0.01</td>
<td>Positive</td>
<td>2.27</td>
</tr>
<tr>
<td>Foster 2007</td>
<td>352</td>
<td>26</td>
<td>6.25 (12.1)</td>
<td>7.13 (13.1)</td>
<td>-0.88</td>
<td>-0.88 (-4.25 to -2.49)</td>
<td>P=0.608</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
<tr>
<td>Suarez-Almazor 2010</td>
<td>527</td>
<td>12</td>
<td>11.65</td>
<td>12.5</td>
<td>-0.85</td>
<td>Not given</td>
<td>P&gt;0.20</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
</tbody>
</table>
**Table 9c: KSS Function Score (0-50 scale)**

<table>
<thead>
<tr>
<th>Reference (1&lt;sup&gt;st&lt;/sup&gt; author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (wks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change sham group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Short-Term Outcomes (&lt;12 weeks)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miller 2009</td>
<td>55</td>
<td>8</td>
<td>3.9</td>
<td>11</td>
<td>-7.1</td>
<td>Not given</td>
<td>Not given for change between groups</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Long-Term Outcomes (≥12 weeks)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miller 2009</td>
<td>55</td>
<td>12</td>
<td>6.3</td>
<td>6</td>
<td>-0.3</td>
<td>Not given</td>
<td>Not given for change between groups</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Outcomes: Acupuncture Vs Standard Care

Outcomes for acupuncture Vs standard care are shown in tables 10-11 below. Five studies compare acupuncture with standard care (Berman et al 1999, Berman et al 2004, Scharf et al 2006, Witt et al 2006, Williamson et al 2007). Of these, three report the WOMAC pain subscale, with two recording short term measurement points and all three reporting long term measurements (table 10). Three studies utilise the WOMAC Index as a functional outcome measure (table 11a), with one giving a short-term measurement (Witt et al 2006) and all three giving long-term end points. Berman et al (2004) use the WOMAC function subscale (table 11b) and Williamson et al (2007) use the OKS as a functional outcome measure, with both short and long term measurements reported (table 11c).
# Acupuncture Vs Standard Care: Pain Outcomes

Table 10: WOMAC Pain Subscale (0-20 scale)

<table>
<thead>
<tr>
<th>Reference (1st author &amp; year)</th>
<th>Pts (n)</th>
<th>Endpoint (weeks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change SC group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Short-Term Outcomes (&lt;12 weeks)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berman 2004</td>
<td>520</td>
<td>8</td>
<td>3.15 (0.29)</td>
<td>1.25 (0.30)</td>
<td>1.9</td>
<td>Not given</td>
<td>P&lt;0.001</td>
<td>Positive</td>
<td>6.55</td>
</tr>
<tr>
<td>Berman 1999</td>
<td>73</td>
<td>8</td>
<td>4.24</td>
<td>0.32</td>
<td>3.92</td>
<td>Not given</td>
<td>P&lt;0.001</td>
<td>Positive</td>
<td>-</td>
</tr>
<tr>
<td>Witt 2006</td>
<td>2627</td>
<td>24</td>
<td>3.47</td>
<td>4.1</td>
<td>-0.63</td>
<td>Not given</td>
<td>P= 0.090</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
<tr>
<td><strong>Long-Term Outcomes (≥12 weeks)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berman 2004</td>
<td>520</td>
<td>26</td>
<td>3.79 (0.33)</td>
<td>1.69 (0.33)</td>
<td>2.1</td>
<td>Not given</td>
<td>P&lt;0.01</td>
<td>Positive</td>
<td>6.36</td>
</tr>
<tr>
<td>Berman 1999</td>
<td>73</td>
<td>12</td>
<td>4.02</td>
<td>0.27</td>
<td>3.75</td>
<td>Not given</td>
<td>P&lt;0.001</td>
<td>Positive</td>
<td>-</td>
</tr>
</tbody>
</table>
## Acupuncture Vs Standard Care: Functional Outcomes

### Table 11a: WOMAC Index 0-100mm scale

<table>
<thead>
<tr>
<th>Reference (1st author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (wks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change SC group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berman 1999</td>
<td>73</td>
<td>8</td>
<td>20.61</td>
<td>0.76</td>
<td>19.85</td>
<td>Not given</td>
<td>P&lt;0.001</td>
<td>Positive</td>
<td>-</td>
</tr>
<tr>
<td>Long-Term Outcomes (≥ 12 weeks)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berman 1999</td>
<td>73</td>
<td>12</td>
<td>17.11</td>
<td>0.44</td>
<td>16.67</td>
<td>Not given</td>
<td>P&lt;0.001</td>
<td>Positive</td>
<td>-</td>
</tr>
<tr>
<td>Scharf 2006</td>
<td>1007</td>
<td>26</td>
<td>23</td>
<td>12</td>
<td>11</td>
<td>11 (14 to -8.0)</td>
<td>P&lt;0.001</td>
<td>Positive</td>
<td>-</td>
</tr>
<tr>
<td>Witt 2006</td>
<td>2627</td>
<td>24</td>
<td>15.3</td>
<td>17.3</td>
<td>-2.0</td>
<td>Not given</td>
<td>P=0.264</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
</tbody>
</table>

### Table 11b: WOMAC Function Subscale (0-68 Likert format)

<table>
<thead>
<tr>
<th>Reference (1st author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (wks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change SC group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berman 2004</td>
<td>520</td>
<td>8</td>
<td>10.77 (0.90)</td>
<td>5.30 (0.95)</td>
<td>5.47</td>
<td>Not given</td>
<td>P&lt;0.001</td>
<td>Positive</td>
<td>6.08</td>
</tr>
<tr>
<td>Long-Term Outcomes (≥ 12 weeks)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berman 2004</td>
<td>520</td>
<td>26</td>
<td>12.42 (1.12)</td>
<td>1.77 (1.07)</td>
<td>10.65</td>
<td>Not given</td>
<td>P&lt;0.001</td>
<td>Positive</td>
<td>9.53</td>
</tr>
</tbody>
</table>
Acupuncture Vs Standard Care: Functional Outcomes cont.

Table 11c: Oxford Knee Score (OKS) (12-60 scale)

<table>
<thead>
<tr>
<th>Reference (1st author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (weeks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change SC group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Williamson 2007</td>
<td>181</td>
<td>7</td>
<td>3.4</td>
<td>0.2</td>
<td>3.2</td>
<td>Not given</td>
<td>P=0.016</td>
<td>Positive</td>
<td>-</td>
</tr>
</tbody>
</table>

Long-Term Outcomes (≥ 12 weeks)

<table>
<thead>
<tr>
<th>Reference (1st author &amp; year)</th>
<th>Pts (n)</th>
<th>End-point (weeks)</th>
<th>Mean change acu group (SD where given)</th>
<th>Mean change SC group (SD where given)</th>
<th>Difference between Mean scores</th>
<th>95% Confidence interval</th>
<th>P-value</th>
<th>Outcome Assessment</th>
<th>Effect Size (Cohen’s D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Williamson 2007</td>
<td>181</td>
<td>12</td>
<td>2.1</td>
<td>-0.3</td>
<td>2.4</td>
<td>Not given</td>
<td>P=0.96</td>
<td>Neutral</td>
<td>N/S</td>
</tr>
</tbody>
</table>

Key (Tables 8-11)

<table>
<thead>
<tr>
<th>Wks</th>
<th>Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pts (n)</td>
<td>Number of participants</td>
</tr>
<tr>
<td>Acu</td>
<td>Acupuncture</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>N/S</td>
<td>Not significant</td>
</tr>
</tbody>
</table>

Where given, 95% Confidence interval and P-values in tables 8-11 relate to differences in mean change with treatment between groups.
4: Discussion

4.1: Acupuncture Vs Sham: Short-Term Pain Outcomes
Eight studies examined the effects of traditional versus sham acupuncture. All were rated as having high internal validity, scoring ≥ 7/10 on the PEDro scale. Of these, five included short-term outcomes for pain (tables 8a and b), with only one achieving a positive outcome (Jubb et al 2008). Table 8a indicates that four trials gave short-term outcomes for the WOMAC pain subscale, with three reporting neutral outcomes for acupuncture compared with sham (Berman et al 2004, Foster et al 2007, Suarez-Almazor et al 2010). This indicates no statistically significant difference between the outcomes of the acupuncture and sham groups at the short term measurement point in these studies.

Jubb et al (2008) report statistically significant improvements in short-term WOMAC pain subscale scores over the sham group (table 8a). However, statistical significance of a treatment does not always correspond to clinical relevance (de Vet et al 2006). Statistical significance shows that the effect occurs beyond some level of chance. Clinical relevance, however, is related more to the benefits derived from a given intervention or treatment that may be considered clinically meaningful (de Vet et al 2006). The minimal clinically important difference (MCID) on the WOMAC (0-20) pain subscale requires an improvement of 1.5 points (Angst, Aeschlimann & Stucki 2001). The results of Jubb et al (2008) therefore show clinically relevant improvements in pain for the acupuncture group (improvement of 3.8 points) at the short-term measurement point of 5 weeks, but not for the sham group (improvement of 1.4 points). An effect size of 0.63 has been calculated, which is considered ‘moderate’ (Cohen 1988). However, a relatively wide 95% confidence interval of 2.4 (0.2 to 4.84) was reported. This indicates a less precise estimate of the treatment effect and raises doubt as to the accuracy of the study in predicting the true size of the effect (Akobeng 2005b).
The study of Jubb et al (2008) was the only trial in the review which scored ‘no’ for the concealed allocation criterion on the PEDro scale. It has been demonstrated that trials in which allocation is not concealed may exaggerate treatment effects by up to 41% (Schulz 2000) and the selection bias which can be introduced in this way can invalidate the study design of an RCT, making the results no more reliable than an observational study (Akobeng 2005a). However, baseline characteristics revealed the two groups to be well-matched, thus decreasing the likelihood of selection bias. The only variable reaching statistical significance between groups was more baseline weight-bearing pain in the acupuncture group (Jubb et al 2008).

Miller et al (2009) reported the KSS pain score as a short-term outcome measure (table 8b). Statistical differences for the change in outcome measure scores between baseline and the end-points were not reported in the study. The authors were contacted requesting this data but no response was received. The outcome can therefore not be accurately assessed, as statistical comparisons are required in order to ascertain whether the intervention group had a significantly different treatment effect to the control group. However, the data shows a larger mean change in KSS scores in the acupuncture group at 12 weeks (7.7) than in the sham group (3.8). This reveals a slight increase in the mean change in the acupuncture group from the 8 week (end of treatment) score of 7.4, and a substantial decrease in the sham group, from the 8 week score of 7.1. This accounts for statistically significant differences between the acupuncture and sham groups becoming apparent only after 12 weeks (Miller et al 2009). A change of ≥2.2 on the KSS knee score has been shown to correlate with clinically meaningful improvements in pain and function (Lingard et al 2001). Therefore, a clinically significant improvement occurred in both groups at both short and long term measurement points. A relatively small sample size of 55 was used in this study and a power calculation was not
performed. Small studies have been shown to overestimate treatment effects (Juni, Altman & Egger 2001) and this is therefore a limitation of this trial.

The results demonstrate that small improvements were seen in the acupuncture group over the 4 weeks following termination of treatment, while they declined in the sham group. A physiological explanation for the delayed effects of traditional acupuncture has been proposed in recent fMRI studies (Liu et al 2011). Both penetrating non-meridian sham and traditional acupuncture have been shown to bring about significant signal changes in the brain stem and cerebellum immediately following treatment (Liu et al 2011). However, delayed stronger functional connectivity was found in the limbic / paralimbic areas and brainstem with traditional acupuncture only (Liu et al 2011). As the limbic areas of the brain play a crucial role in pain processing and affective aspects of pain (Zhao 2008), the findings suggest a plausible explanation for the delayed effects of traditional acupuncture seen compared to sham techniques (Liu et al 2011).

When analysed against the ‘Levels of evidence’ criteria in table 5, there is inconclusive evidence that acupuncture is superior to sham for improving function in OA knee at the short term measurement point. Although the majority of included trials achieve high internal validity with PEDro scores of $\geq 7/10$, the study achieving positive outcomes (Jubb et al 2008) did not perform concealed allocation, potentially introducing selection bias (Akobeng 2005a). As the criteria for ‘generally consistent’ has been defined as two thirds of more of the results having the same outcome assessment (Ezzo et al 2001) and only one out of five of the outcomes within this review achieve a positive rating, this criterion cannot be fulfilled. Therefore an ‘inconclusive’ level of evidence is achieved.
4.2: Acupuncture Vs Sham: Long-Term Pain Outcomes

Long-term pain outcomes for acupuncture Vs sham reveal positive outcome assessments for three out of six long-term outcome markers (tables 8a-c). Neutral outcomes were found for two trials (Foster et al 2007, Suarez-Almazor et al 2010), indicating no significant difference between the acupuncture and sham groups at the long term measurement point.

Vas et al (2004) demonstrated positive outcome assessments at the long term measurement point of 12 weeks for both WOMAC pain subscales and VAS pain scores (tables 8a and 8c). The data reveals clinically relevant improvements in WOMAC pain subscale scores in both the acupuncture and sham groups (Angst, Aeschlimann & Stucki 2001). However, a statistically significant difference between the two groups is demonstrated, showing a significantly better effect in the acupuncture group over the sham group (table 8a). VAS pain scores also reveal positive outcomes (table 8c), showing a statistically significant improvement in the acupuncture group over the sham group. Although both acupuncture and sham groups showed improvements in VAS scores at 12 weeks, only the acupuncture group demonstrated a clinically relevant improvement (Bird & Dickson 2001).

Foster et al (2007) found no significant improvements in pain levels on the WOMAC pain subscale at either the short or long term measurement points (table 8a). A potential limitation of this study is the short treatment duration of three weeks. This does not meet the criteria of ‘at least 6 weeks total treatment duration’ specified in this review for acupuncture adequacy. Pain threshold has been shown to gradually increase with the application of acupuncture, indicating a delayed effect of acupuncture analgesia (Zhao 2008). It is therefore uncertain whether this shorter duration of treatment was sufficient to produce adequate analgesic
effects. This may therefore have contributed to the lack of statistically significant effects of acupuncture over sham.

When analysed against the ‘Levels of evidence’ criteria in table 5, there is inconclusive evidence that acupuncture is superior to sham for improving pain in OA knee at the long term measurement point. The trials scoring positively are considered to be of generally high quality with high internal validity scores of \( \geq \frac{8}{10} \) achieved on the PEDro scale. Unlike the short-term outcomes for pain, multiple relevant RCT’s (Berman et al 2004, Vas et al 2004) achieve positive outcomes for this measurement point. However, heterogeneity is identified within the results, as two large scale studies report no significant difference between pain outcomes for traditional and sham acupuncture at the long term measurement points (Foster et al 2007, Suarez-Almazor et al 2010). As two thirds or more of the results do not achieve a positive outcome, the criteria for ‘generally consistent’ outcomes cannot be fulfilled (Ezzo et al 2001) and an ‘inconclusive’ level of evidence is achieved.

### 4.3: Acupuncture Vs Sham: Short-Term Functional Outcomes

Five trials report short term functional outcomes for acupuncture compared with sham (table 9a-d). Of the six short-term functional outcomes reported in these studies, two reveal positive outcome assessment (Berman et al 2004, Witt et al 2005), indicating that statistically, acupuncture has a significantly better effect than sham at the short term measurement point in these studies.

Berman et al (2004) demonstrate a statistically significant difference in WOMAC function subscale scores between acupuncture and sham groups at 8 weeks (table 9b), with a large effect size of 3.26 (Cohen 1988). The MCID in WOMAC function subscale (0-20) scores has
been shown to correlate with a reduction of ≥5.37 points (Tubach et al 2005). Clinically relevant improvements were therefore seen in both acupuncture and sham groups at 8 weeks (mean change of 10.77 and 7.84 respectively). Similarly, Witt et al (2005) found that acupuncture treatment had statistically significant short-term effects on function compared to sham acupuncture, assessed with the WOMAC Index (table 9a). The MCID for the WOMAC (0-100) scale has been shown to be ≥7.9 points (Tubach et al 2005). Therefore, clinically significant improvements were seen in both acupuncture and sham groups (mean improvement of 23.9 and 16.7 respectively). Both trials were found to have high internal validity, scoring 8/10 and 9/10 on the PEDro scale respectively.

When analysed against the ‘Levels of evidence’ criteria in table 5, there is inconclusive evidence that acupuncture is superior to sham for improving function in OA knee at the short term measurement point. Included trials are considered to be of generally high quality. Trials scoring positively for outcome assessment achieved high internal validity scores of ≥8/10 on the PEDro scale and were found to have clinically meaningful results (Tubach et al 2005). However, as only two out of five short term functional outcomes achieve a positive outcome (Berman et al 2004, Witt et al 2005), the outcomes cannot be considered to be ‘generally consistent’ and an ‘inconclusive’ level of evidence is therefore achieved.

4.4: Acupuncture Vs Sham: Long-Term Functional Outcomes

Table 9a shows that notably positive results were shown in the trial of Vas et al (2004), reporting a mean change of 47.6 points in the acupuncture group on the WOMAC Index. This indicates a large clinically significant improvement (Tubach et al 2005). This study was rated as having high internal validity, scoring 9/10 on the PEDro scale. However, the validity of subject blinding was not assessed by the authors. Non-penetrating needles were utilised as a sham technique in this trial. Although the credibility of this technique has been previously demonstrated with acupuncture-naïve patients (Streitberger & Kleinhenz 1998), the authors did not specifically recruit acupuncture-naïve patients. Therefore, the validity of this sham technique is questionable in this instance. Participant knowledge of treatment received can be a source of bias, resulting in exaggerated treatment effects (Akobeng 2005a). This factor could therefore potentially have contributed to the positive results seen in this trial.

A further contributing factor to this study’s strongly positive results may have been that all four local points were given electrical stimulation. Interestingly, only one other study incorporated electrical stimulation to four or more local points (Jubb et al 2008), and this study also indicated positive results in favour of acupuncture. Electroacupuncture accelerates the release of enkephalin, beta-endorphin and endomorphin, which in turn stimulate the mu and delta opioid receptors to produce analgesic effects (Lin & Chen 2008). The increased stimulation when using electroacupuncture has been shown to produce stronger analgesic benefits than manual acupuncture alone (Ulett, Han & Han 1998).

The large-scale study of Berman et al (2004) was found to be the longest follow up period which scored positively for acupuncture versus sham (table 9b). A significant difference between the acupuncture and sham groups was maintained at 26 weeks, with a large effect size of 2.27 (Cohen 1988). However, a limitation of this study is the high dropout rate, which
was such that at 26 weeks, 43% of the participants in the education group, and 25% in each of
the true and sham acupuncture groups were not available for analysis (Berman et al 2004).
This high drop-out rate can lead to potential bias as the reader can no longer be confident that
important baseline prognostic factors in the groups are comparable (Akobeng 2005a).
Although an intention to treat analysis was performed to reduce this potential bias (Akobeng
2005a), the authors comment that they cannot be certain that the considerable attrition rate did
not affect their 26 week results (Berman et al 2004). Therefore, the findings must be
interpreted with caution.

A delayed treatment effect was demonstrated in the trial of Berman et al (2004), with
statistically significant improvements in pain scores not becoming evident until week 14.
Similarly, functional outcomes demonstrated statistical significance from baseline in the
traditional acupuncture group by week 8, which subsequently improved further by week 26.
The acupuncture protocol of Berman et al (2004) differed to that of the other trials,
incorporating a gradually tapering treatment regime, in which treatment was ongoing until the
final end point measurements were taken at 26 weeks. This tapering style has been supported
as an effective method for sustaining the treatment benefits of acupuncture, particularly in the
treatment of chronic pain (Helms 1998). Similar trials have demonstrated that by
administering monthly maintenance or ‘top-up’ treatments once a clinical improvement is
seen, the benefits of acupuncture can be sustained in the longer term (Christensen et al 1992).

Conversely, Scharf et al (2006) did not find a statistically significant difference in function,
measured with the WOMAC Index (table 9a) between traditional and sham acupuncture
groups at 26 weeks. A 6 week course of treatment incorporating 10 acupuncture sessions was
implemented. Therefore, unlike the trial of Berman et al (2004) treatment was not ongoing
until the final measurement point. Although there was no statistically significant difference between the two groups, the acupuncture group achieved a success rate of 53.1% and the sham group of 51% (Scharf et al 2006). The mean change in both groups achieved clinical significance (Tubach et al 2005). These findings could have been influenced by a number of factors, including the increased intensity of provider contact, where the increased therapist’s attention and presence in both the acupuncture and sham groups may have contributed to a beneficial treatment effect (Manheimer et al 2010). The potentially strong placebo effect of acupuncture may also have influenced the positive results seen in both true and sham acupuncture groups (Moerman & Jonas 2002). In addition, the physiological effects of minimal needle insertion, used as the sham acupuncture technique has been shown to have similar efficacy to true acupuncture in terms of pain relief (Moffet 2009). The physiological and clinical effects of sham acupuncture techniques are discussed further in section 4.10.

Miller et al (2009) performed both short and long term measurements using the KSS function score (table 9c). As with the KSS pain scores, statistical differences for the change in outcome measure scores between baseline and the end-points were not reported in the study and therefore outcome assessment could not be made. Data within this study reported upon the statistical difference between the acupuncture and sham groups at eight and twelve weeks. Although a statistically significant difference is reported between the groups at 12 weeks, it is unclear whether statistical differences for the change in outcome measure scores between baseline and the end-points also exist. Clinically significant improvements were seen in both acupuncture and sham groups at the short and long term end points (Lingard et al 2001). It is noted that the baseline data for KSS function scores appear to be considerably different (acupuncture group KSS function score = 61.1, sham group = 48.7). This is highlighted by the
authors as a baseline difference between groups, but they did not find this difference to be statistically significant (P=0.06).

When analysed against the ‘Levels of evidence’ criteria in table 5, there is inconclusive evidence that acupuncture is superior to sham for improving function in OA knee at the long term measurement point. The majority of trials in this category can be considered to be of high quality and all the included studies achieved a PEDro score of ≥ 7/10, indicating high internal validity. However, although two high-quality studies reported positive outcomes for acupuncture compared with sham at the long-term measurement point (Berman et al 2004, Vas et al 2004), these outcomes cannot be considered to be ‘generally consistent’ as less than two thirds of the results have positive outcomes. An ‘inconclusive’ level of evidence is therefore achieved.

4.5: Acupuncture Vs Standard Care: Short-Term Pain Outcomes

Three studies examine WOMAC pain subscale outcomes for acupuncture compared with standard care (table 10). Of these, two report short-term outcomes (Berman et al 1999, Berman et al 2004). Both studies indicate that acupuncture had a significantly better effect than the standard care at the short-term measurement point. The study of Berman et al (2004) also shows a large effect size of 6.55 (Cohen 1988). This study shows clinically significant improvements in WOMAC pain subscale scores for both acupuncture and standard care groups (Angst, Aeschlimann & Stucki 2001). Conversely, Berman et al (1999) show a clinically significant improvement only in the acupuncture group (Angst, Aeschlimann & Stucki 2001).
The standard care control differed between the trials of Berman et al (1999) and Berman et al (2004). A limitation the trial of Berman et al (1999) was the lack of equal therapist contact between the acupuncture and standard care groups. Participants in the standard care group were required to maintain their current level of oral analgesia but did not receive any additional treatment. Therefore, acupuncture participants experienced increased frequency and intensity of provider contact during their bi-weekly treatments over eight weeks, which may have contributed to a positive treatment effect (Manheimer et al 2010). Conversely, the standard care group of Berman et al (2004) consisted of six two-hour group sessions based on the Arthritis Self Management Programme (Lorig et al 1998). Participants in the standard care group were also periodically mailed educational materials in an attempt to equalise the amount of experimental contact time between the groups.

When analysed against the ‘Levels of evidence’ criteria in table 5, there is strong evidence that acupuncture is superior to standard care for improving pain in OA knee at the short term measurement point. The two studies included in this section (Berman et al 1999, Berman et al 2004), achieve PEDro scores of ≥ 7/10 and following critical appraisal, are considered to be high quality studies. Both trials score positive outcome assessment and consequently meet the criteria for ‘generally consistent’ outcomes (Ezzo et al 2001). A ‘strong’ level of evidence is therefore achieved.

4.6: Acupuncture Vs Standard Care: Long-Term Pain Outcomes

Three studies reported long term pain outcomes for acupuncture compared with standard care using the WOMAC pain subscale (table 10). Berman et al (2004) conducted follow up measurements at 26 weeks, and found that a statistically significant difference between the acupuncture and standard care groups had been maintained (table 10). A large effect size of
6.36 was calculated (Cohen, 1988). Berman et al (1999) also reported statistically significant improvements in WOMAC pain subscale scores in the acupuncture group compared with standard care, four weeks after cessation of treatment at 12 weeks. Clinically significant improvements were achieved in the acupuncture group but were not evident in the standard care group (Angst, Aeschlimann & Stucki 2001).

Witt et al (2006) conducted a large scale trial of acupuncture for OA of the knee or hip. The data for OA knee has been extracted for the purposes of this review. The standard care group were permitted to receive any additional conventional treatments as needed, excluding any form of acupuncture. The lack of specific treatments included in the standard care group may be considered a limitation from an experimental perspective as the effects of acupuncture against the specific effects of interventions in the standard care group cannot be analysed (Witt et al 2006). In addition, the lack of control over experimental contact time between the groups can be considered a limitation (Sim & Wright 2000). However, the study design was chosen to be representative of general medical practice, which is individualised to patient’s needs (Witt et al 2006). Results at 12 weeks reported a significant improvement in the acupuncture group compared with the standard care group. However, these between group differences were not maintained at the long-term measurement point of 24 weeks (12 weeks following cessation of acupuncture treatment). These findings are consistent with mechanistic studies of acupuncture analgesia which do not support such significantly long-lasting effects of acupuncture analgesia after acupuncture stimulation has been terminated (Cui et al 2005, Zhao 2008).

The mean change in WOMAC pain subscale scores at 24 weeks for the study of Witt et al (2006) are shown in table 10. Both groups show improvements compared with baseline which
are reported to be significant on within-group analysis (Witt et al 2006) and which fulfil the criteria for MCID (Angst, Aeschlimann & Stucki 2001). A slightly larger reduction is noted in the standard care group, however, there is no statistically significant difference between the groups (P=0.090). Witt et al (2006) achieved the lowest PEDro score (6/10) of those included in the review. This was primarily due to inadequate blinding of subjects, therapists and assessors, which has been shown to exaggerate treatment effects by an average of 17% (Akobeng 2005a). Inadequate follow up was also identified as a limitation, as the PEDro criteria specifies that at least one key outcome should be obtained for more than 85% of the subjects initially allocated into groups (PEDro 2010). Witt et al (2006) achieved 81.3% follow up at 12 weeks, therefore not achieving this criterion.

When analysed against the ‘Levels of evidence’ criteria in table 5, there is strong evidence that acupuncture is superior to standard care for improving pain in OA knee at the long term measurement point. Two out of three trials scored positive outcome assessment in this category (table 10). The trials with positive long-term outcomes (Berman et al 1999, Berman et al 2004) showed high internal validity, scoring ≥7/10 on the PEDro scale and revealed clinically relevant outcomes (Angst, Aeschlimann & Stucki 2001). They were also considered to be of generally high quality following critical appraisal. The trial scoring neutral outcome assessment (Witt et al 2006) was also considered to be of high quality and was noted to be the largest scale study within the review. However, as two thirds of the outcomes are positive, they can be considered to be ‘generally consistent’ according to the specified criteria (Ezzo et al 2001). A ‘strong’ level of evidence is therefore achieved.
4.7: Acupuncture Vs Standard Care: Short-Term Functional Outcomes

Three studies report short term functional outcome measures for acupuncture compared with standard care (Berman et al 1999, Berman et al 2004, Williamson et al 2007). All three studies report positive outcome assessment at their short term measurement points, indicating a more significant improvement in the acupuncture group compared with standard care (tables 11a-c). Berman et al (1999) show a large difference in WOMAC Index mean change scores between the acupuncture and standard care groups at 8 weeks (table 11a). The acupuncture group shows a clinically significant improvement in WOMAC Index scores, whereas the standard care group does not (Tubach et al 2005). This improvement correlates with a 34% reduction in WOMAC Index score from baseline at week 4 and a 42% reduction at week 8 (Berman et al 1999). Berman et al (2004) also demonstrate a statistically significant improvement in WOMAC function subscale scores in the acupuncture group compared with the standard care group at 8 weeks (table 11b). A large effect size of 6.08 has also been calculated for this measurement point (Cohen 1988).

Williamson et al (2007) report a significant difference in OKS scores between acupuncture and standard care groups at 7 weeks follow-up (table 11c). There is no well-established value for the MCID in OKS score in the literature (Alzahrani et al 2011). Therefore it is unclear whether the mean change in the acupuncture group is clinically significant. However, it appears unlikely that the change of 0.2 seen in the standard care group would be clinically meaningful. The standard care group received an exercise and advice leaflet to read at home, while the acupuncture group were seen weekly over 6 sessions. This was not a comparable level of therapist contact between the groups and this is therefore a limitation of the study (Manheimer et al 2010).
When analysed against the ‘Levels of evidence’ criteria in table 5, there is strong evidence that acupuncture is superior to standard care for improving functional level in OA knee at the short term measurement point. All three studies reporting short term functional outcome measures achieve positive outcomes (Berman et al 1999, Berman et al 2004, Williamson et al 2007). The change in outcome measure scores were also found to be clinically significant in the studies reporting WOMAC Index or function subscales (Tubach et al 2005). These studies were considered to be of generally high quality and achieved high internal validity with PEDro scores of ≥7/10. As the criteria is fulfilled for ‘multiple, relevant, high quality RCT’s with generally consistent outcomes’ (table 5), a ‘strong’ level of evidence is achieved.

4.8: Acupuncture Vs Standard Care: Long-Term Functional Outcomes


Scharf et al (2006) report a statistically significant improvement in the WOMAC Index at 26 weeks in the acupuncture group compared with the standard care group (table 11a). The authors attempted to legislate for the variable of different experimental contact time between the groups by allowing the standard care group the same number of therapist contacts as the acupuncture and sham groups. However, although equal in number, it is acknowledged that the patient-provider contacts were generally less intense in the standard care group (Scharf et al 2006). Significant improvements in the WOMAC index were demonstrated in the traditional and sham acupuncture groups compared with the standard care group at week 26. This large-scale trial was found to have high internal validity, scoring 9/10 on the PEDro
A limitation of utilising standard care controls may be the factor of patient preferences and expectations affecting outcomes (Kalaukalani et al 2001). It has been demonstrated that patient’s attitudes and expectations of the treatment they receive can significantly influence treatment response (Linde et al 2010). Interestingly, in the trials of Berman et al (2004), Scharf et al (2006) and Williamson et al (2007), it is noted that more patients discontinued treatment in the standard care groups, compared to the traditional acupuncture and sham groups. In addition, in the cases of Berman et al (2004) and Scharf et al (2006), the majority of these dropouts occurred immediately after treatment group assignment. This may reflect some patient’s pre-randomisation preferences for acupuncture which could influence the later assessment of outcomes in favour of the acupuncture and sham groups (Kalaukalani et al 2001). However, these trends were not reflected in the trials of Berman et al (1999) or Witt et al (2006), who reported comparable dropout rates for each of their treatment groups.

When analysed against the ‘Levels of evidence’ criteria in table 5, there is strong evidence that acupuncture is superior to standard care for improving functional level in OA knee at the long term measurement point. Three out of five long term functional outcomes achieve a positive outcome assessment, meeting the criteria for ‘generally consistent’ outcomes (Ezzo et al 2001). The trials scoring positive outcomes were considered to be of high quality and achieved high internal validity, with PEDro scores of ≥7/10. Clinically significant outcomes were demonstrated in these trials (Tubach et al 2005). Two of these studies were large scale,
multi-centre trials (Scharf et al 2006, Berman et al 2004). It is, however noted that one of the trials scoring neutral outcome assessment was the largest scale study within the review and was also considered to be of high quality (Witt et al 2006). Nevertheless, the specified outcomes fulfil the criteria specified in table 5 for a ‘strong’ level of evidence.

4.9: Adequacy of Acupuncture

No consensus exists on how to best assess treatment adequacy in acupuncture RCT’s, and no criteria have yet been validated (Birch 2004). Therefore, the criteria specified in table 3 for acupuncture adequacy has been formulated from the limited existing evidence base surrounding acupuncture treatment ‘dose’ as described in section 2.2.4.

On assessment of the adequacy of acupuncture according to the criteria in this study, only one study was scored as inadequate (Foster et al 2007). The study of Foster et al (2007) did not fulfil the criteria of a total course of acupuncture of at least six weeks. Instead, up to six acupuncture treatments were performed over three weeks. Empirical evidence suggests that due to the cumulative effects of acupuncture, an initial course of 6 weekly treatments should be conducted to ascertain whether a treatment response has been elicited (Helms 1998). It is thought that a course of up to 12 sessions is then optimum in order to obtain a more enduring response (Helms 1998). Other research suggests that a minimum of 30 minutes duration over 10 sessions is optimal (Lundeberg et al 1988). Pain outcomes of the trial of Foster et al (2007) were neutral at both short and long term measurement points, as no significant differences were found between the traditional acupuncture and sham groups.

Two studies (Vas et al 2004, Witt et al 2006) were scored as ‘unclear’ for adequacy of acupuncture (table 7). In both cases, unsuccessful attempts were made to contact the authors.
for clarification on the acupuncture treatment and dosage. In the trial of Vas et al (2004), the duration of each acupuncture treatment is not documented. It is also unclear whether treatments were conducted weekly, as the authors reported:

“The treatment lasted 12 weeks, starting with visit 0 and ending with visit 11” (Vas et al 2004: 2).

It is unclear from this statement how frequently over this period acupuncture treatments were performed. As discussed in sections 4.2 and 4.4, this trial scored notably positive results for long term outcomes of pain and function with acupuncture compared to sham.

Witt et al (2006) also scored ‘unclear’ for acupuncture adequacy (table 7). This study utilised individualised acupuncture point prescription for each patient, in which the number and location of points were chosen at the treating physician’s discretion. This method is more representative of TCM practice than the ‘formula’ approach (Ezzo et al 2001). However, in research, formula acupuncture protocols are more frequently used as they control better for unknown confounders (Ezzo et al 2001). Due to the individualised acupuncture protocol, it is unclear whether this trial meets the criteria for a minimum of 20 minutes treatment duration and the use of points local to the painful joint, as specified in table 3. The authors also did not document whether the treating physicians were instructed to illicit DeQi. As discussed above, this trial showed significant improvements in pain and functional outcome measures between acupuncture and standard care groups at 12 weeks. However, these between-group differences were not maintained at the long term measurement point of 24 weeks (Witt et al 2006).

Eight trials within the review were found to have used adequate acupuncture according to the criteria specified in table 3. It is uncertain whether the shorter duration of treatment in the trial of Foster et al (2007) may have contributed to the lack of statistically significant effects of
acupuncture over sham. Considering the variation in outcomes found in the eight studies which scored positively for acupuncture adequacy, it appears unlikely that the specified criteria for adequacy had a strong bearing upon the clinical outcomes seen.

4.10: Sham Acupuncture

Many of the included studies utilise a sham treatment but techniques of delivering sham acupuncture can vary widely, making it difficult to draw comparisons between studies (Dincer & Linde 2003). Within this review, four studies utilised non-needle penetration at the same acupuncture points as those used in the true acupuncture group (Vas et al 2004, Foster et al 2007, Jubb et al 2008, Miller et al 2009). Three studies utilised minimal stimulation, superficial needling at non-acupuncture points (Witt et al 2005, Scharf et al 2006, Suarez-Almazor et al 2010), and one study utilised a combined insertion and non-insertion sham method (Berman et al 2004). The use of minimal stimulation, superficial needling at non-acupuncture points has been shown to have similar physiological effects to true acupuncture (Lund, Naslund & Lundeberg 2009) and to have similar efficacy in terms of pain relief (Moffet 2009). Therefore this method cannot be viewed as a valid control from a physiological perspective (Lund, Naslund & Lundeberg 2009).

Non-penetrating sham methods also have methodological limitations, as it has been demonstrated that any method incorporating light touch of the skin will stimulate C-tactile afferents. This stimulation induces a “limbic touch” response resulting in emotional and hormonal reactions (Olausson et al 2002). It is therefore likely that all of the sham methods described will activate the C-afferent response and can therefore not be regarded as “placebo”, as they are not inert treatments (Lund & Lundeberg, 2006). Hence it may be
inappropriate for authors to draw estimates of the clinical effectiveness of acupuncture from the observed difference between true and sham acupuncture in RCT’s (Cummings, 2008).

These non-inert effects of sham acupuncture may explain the statistically significant improvements in both acupuncture and sham groups over standard care or waiting list controls seen in some studies (Scharf et al 2006, Suarez-Almazor et al 2010). Both Scharf et al (2006) and Suarez-Almazor et al (2010) reported improvements over a third treatment arm when utilising minimal depth and minimal stimulation needling at non-acupuncture points as a sham technique. This raises the question of whether specific point selection is significant, and whether deep needle insertion with stimulation of DeQi is superior to shallow needling. However, the effects of more intensive practitioner-patient contact (Kaptchuk 2002), and the placebo effects of true and sham acupuncture (Moerman & Jonas 2002) in these two studies may also account for these observations.

In addition, although both sham regimes utilised non-acupuncture points, Scharf et al (2006) highlight that they could not be completely sure that no active points had been needled. This also appears true for the trial of Suarez-Almazor et al (2010), which utilised points only 1.0 cun away from true acupuncture points in their sham protocol, raising the question of whether true acupuncture points may have been needled in some instances. In addition, it has been demonstrated that superficial needling within the same dermatome as the knee joint may have the same physiological effects and therefore cannot be considered a valid placebo (Lund, Naslund & Lundeberg 2009). The sham technique of Suarez-Almazor et al (2010) also incorporated a short duration of TENS electrical stimulation of the needles. Although a short duration of electrical stimulation was used, this cannot be viewed as an inert placebo
technique, as fMRI imaging has demonstrated significant areas of effect in the pain centres of the brain with durations of acupuncture up to only a few seconds (Li et al 2006).

4.11: Comparison with Systematic Reviews

The results of this study are similar to the meta-analysis of Manheimer et al (2007), who reported clinically relevant short and long-term effects of acupuncture compared with standard care controls. However, comparison with sham acupuncture revealed clinically irrelevant, short term benefits only. Conversely, White et al (2007) found acupuncture to be superior to sham over both short and long-term follow up periods. Acupuncture was also found to be superior to standard care in short and long term follow up periods, however these results were weakened by heterogeneity. The consistent findings between the two previous meta-analyses and the current review regarding the effectiveness of acupuncture over standard care controls are promising, particularly when the lack of long-term data to support the use of other interventions such as oral or topical NSAID’s for OA knee is considered (Bjordal et al 2004). However, the extent to which the positive results of acupuncture over standard care can be attributed to the specific effects of acupuncture compared with the non-specific effects such as placebo remains unclear (Moerman & Jonas 2002).

The criteria for defining short and long-term measurement points differed between this study and the two previous meta-analyses. For the purposes of this review, the short-term measurement point has been defined as the measurement point closest to 6 weeks, but less than 12 weeks post-randomisation. The long-term measurement point is 12 or more weeks post-randomisation, and the measurement point closest to 24 weeks has been chosen. Conversely, White et al (2007) defined short-term as the point closest to 12 weeks, but up to 25 weeks from randomisation. The long-term end point was taken as the last reported
measurement between 26 and 52 weeks. Manheimer et al (2007) defined short-term as the point not exceeding 12 weeks, but closest to 8 weeks after randomisation. The long-term end point was defined as the point exceeding 12 weeks but closest to 6 months post-randomisation. There is no agreed criteria as to what measurement points should constitute short and long-term in reviews such as this. However, a minimum follow up period of 12 weeks has been suggested to constitute adequate long-term follow up, with periods less than this examining only short-term effects (Birch 2004).

Variations in inclusion and exclusion criteria have also resulted in different study selection than the two previous meta-analyses, contributing to the variation in findings. Both authors included waiting-list controlled trials and White et al (2007) did not specify a minimum observation period or that traditional acupuncture points should be used. In addition, more recent studies have been included in the current review (Foster et al 2007, Williamson et al 2007, Jubb et al 2008, Miller et al 2009, Suarez-Almazor et al 2010). Although meta-analyses were conducted by White et al (2007) and Manheimer et al (2007), the included studies in the current review were deemed to be heterogeneous, as they differ considerably in terms of the comparisons used and outcomes measured. It was therefore not considered appropriate to pool the results for meta-analysis (Higgins & Green 2008).

4.12: Safety

Acupuncture has been shown to have few serious side effects and is regarded as a largely safe treatment method (Ernst & White 2001, Macpherson et al 2001a). Serious adverse effects related to acupuncture are reported to be rare when performed by a competent practitioner (Macpherson et al 2001a, Yamashita et al 2006).
Details of adverse and side effects reported in the included RCT’s are shown in appendix 6. Nine authors described specific minor adverse or side-effects during their trials (Berman et al 2004, Vas et al 2004, Witt et al 2005, Scharf et al 2006, Witt et al 2006, Foster et al 2007, Williamson et al 2007, Jubb et al 2008, Suarez-Almazor et al 2010). Of the adverse or side-effects thought to be related to the acupuncture treatment, none were reported to be serious, with bruising and minor bleeding or post treatment pain reported most frequently (Vas et al 2004, Witt et al 2006, Foster et al 2007, Williamson et al 2007, Suarez-Almazor et al 2010). However, one incidence of post-treatment infection at the needle site was reported (Suarez-Almazor et al 2010), which could be regarded as a serious adverse event (MacPherson 1999) although it was not reported as such in the trial.

Many studies document ‘adverse events’ as any unfavourable medical event occurring during or after the study period, regardless of any causal relationship with acupuncture (Yamashita et al 2006). Subsequently, serious adverse events including death due to myocardial infarction are reported by two authors (Witt et al 2005, Scharf et al 2006), and heart disease, cancer and stroke by another (Berman et al 2004). A flare-up of knee synovitis was also reported in one study (Jubb et al 2008). No serious adverse effects were reported to be associated with the acupuncture intervention and the frequency of serious adverse events was similar between acupuncture and control groups.
5: Limitations

5.1: Risk of Bias

A limitation of this review includes language bias, as due to resource implications only articles published in English were included. Although it is acknowledged that accessing Chinese databases would increase the number of articles yielded, it is estimated that this omission will have a conservative effect on the results of a review, as published Chinese studies of acupuncture are known to yield invariably positive results (Vickers et al 1998). Nevertheless, a full search without language restriction would have provided the most comprehensive view of the existing research in addition to minimising publication bias (Akobeng 2005b).

Publication bias is a further limitation due to the included articles being obtained from published sources only (Wieseler & McGauran 2010). It has been demonstrated that trials with statistically significant results are preferentially indexed in Medline (Akobeng 2005b) and may overestimate the effects of interventions (Wieseler & McGauran 2010). As extensive efforts were not made to identify unpublished data and grey literature sources, this is a limitation of the review (Wieseler & McGauran 2010). It is also acknowledged that omitting extensive hand and grey literature searching may have limited the number of appropriate articles yielded (Higgins & Green 2008).

The use of two reviewers to independently assess the quality of the included studies on the PEDro scale minimised the risk of reviewer bias (Akobeng 2005b). However, it is recognised that reviewer bias cannot be eliminated completely due to acupuncture being the specialised area of interest of both assessors. Although the second reviewer was blinded to the authors of the RCT’s, the risk of bias could have been limited further by blinding all reviewers to the
authors and article titles under evaluation (Higgins & Green 2008). Independent assessment of the adequacy of acupuncture within the trials may also have reduced reviewer bias, in line with the reviews of Ezzo et al (2001) and Manheimer et al (2007). Selection bias could also have been limited further by involving two reviewers in independent article screening against the predefined inclusion criteria (Maxwell et al 2006).

5.2: PEDro Scoring

Items on the PEDro scale have equal ‘weighting’ in terms of the score achieved despite individual scale items reflecting different importance in terms of the risk of bias of an RCT (Sherrington et al 2000). It may therefore be misleading to take into account only the PEDro score when examining the validity of a study. This is noted to be the case in the trial Jubb et al (2008), which scores 7/10 on the PEDro scale, but does not include allocation concealment. This trial therefore cannot be considered as valid as a study such as that of Berman et al (1999), also scoring 7/10, but with the omitted items of patient, therapist and assessor blinding which, although limitations, have been shown to have a less significant impact on the validity of a trial (Akobeng, 2005a).

Although the majority of discrepancies in PEDro scores between the two independent assessors could be resolved by discussion and consensus, there was one instance in which resolution could not be achieved and a third reviewer was required to make an independent assessment of the trial. This discrepancy occurred when determining whether item 7 of assessor blinding had been fulfilled in the trial of Miller et al (2009). In line with PEDro guidelines, this criterion was marked as ‘no’, as assessor blinding was not explicitly stated in the report (Maher et al 2003). This highlights the inevitable element of subjectivity and potential reviewer bias which can occur when utilising scales such as this (Slavin 1995).
All of the included trials scored ‘no’ on the PEDro scale for criterion 6: therapist blinding (appendix 5). Double blinded studies in which both participants and health care professionals are blinded to the assigned treatment have been shown to reduce the risk of bias in RCT’s (Akobeng 2005a). However, depending upon the nature of the RCT, therapist blinding is not always appropriate or possible (Akobeng 2005a). Due to the nature of acupuncture, the practicalities of blinding therapists in acupuncture RCT’s are very difficult (White, Filshi & Cummings 2001). All the included RCT’s have therefore have achieved a lower score on the PEDro scale due to this criterion.

5.3: Heterogeneity

Results within the review were not pooled for meta-analysis as they were considered to differ considerably in terms of the comparisons used and outcomes measured. However, a more accurate assessment of the level of heterogeneity could have been assessed using I² tests (Higgins & Green 2008). An I² value of <40% may represent acceptable level of heterogeneity for data to be pooled (Higgins & Green 2008). Pooling the data for meta-analysis in these instances would have allowed an overall effect size to be calculated and provided additional insight into the overall clinical effectiveness of the interventions (MacPherson et al 2008).

5.4: Recommendations for Further Research

It is unclear from this review whether the adequacy of acupuncture had a bearing upon clinical outcomes. However, as there is no consensus on how best to assess the adequacy of acupuncture treatment within acupuncture trials (Birch 2004), it is unclear whether valid criteria were used. It is acknowledged that examination of the adequacy of acupuncture treatments is important in order to correctly assess trials and to develop improved protocols
for future research (White & Ernst 1998). As there is currently limited evidence as to what constitutes adequate acupuncture treatment (Vas & White 2007), further research in this area would be valuable.

Many acupuncture RCT’s use a sham comparison with the intention of separating the specific from the non-specific (placebo) effects of acupuncture (Lund & Lundeberg 2006). However, it has been demonstrated that the variety of sham methods described cannot be regarded as “placebo”, as they are not inert treatments (Lund & Lundeberg, 2006). Although it is accepted that the physiological effects of sham acupuncture cannot be diminished altogether, further research leading to recommendations for the most appropriate sham technique to minimise the specific effects of acupuncture would be beneficial (Linde et al 2010). In addition, there are few studies comparing acupuncture with other active non-pharmacological interventions recommended in the NICE guidelines for the Care and Management of Osteoarthritis in Adults (NICE, 2008) such as exercise or manual therapy and therefore, future research should address this.
6: Conclusion

There is inconclusive evidence to support the efficacy of traditional acupuncture over sham acupuncture in the short or long-term improvements of pain or function in OA knee. However, there is strong evidence to support the efficacy of acupuncture over standard care in the short and long-term improvement of pain and function in OA knee and the outcomes to support this have been shown to be clinically significant (Angst, Aeschlimann & Stucki 2001, Tubach et al 2005). Considering its relative safety in comparison to other standard care interventions such as NSAID’s, the evidence appears to be sufficiently robust to support the consideration of acupuncture as a clinical treatment option within a multidisciplinary approach for treating OA knee. It is unclear whether the adequacy of acupuncture within the included trials had a bearing upon the clinical outcomes. Further research into appropriate sham techniques and the criteria for adequate acupuncture would be beneficial.
References


Akobeng (2005a) ‘Understanding randomised controlled trials’ Archives of Disease in Childhood 90 (8) 840-844

Akobeng (2005b) ‘Understanding systematic reviews and meta-analysis’ Archives of Disease in Childhood 90 (8) 845-848


Angst F, Aeschlimann A, Stucki G (2001) ‘Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities’ Arthritis Care and Research 45 (4) 384-391


changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement’ *Annals of the Rheumatic Diseases* 64:29-33


Williamson L, Wyatt MR, Yein K, Melton JK (2007) ‘Severe knee osteoarthritis: a randomized controlled trial of acupuncture, physiotherapy (supervised exercise) and standard management of patients awaiting knee replacement’ *Rheumatology* 46 (9) 1445-1449


Acknowledgements

I would like to express thanks to the following people:

Steve Bailey for his independent assessment of PEDro scores;
James Turner and Jonathan Dove for their advice and constructive criticism;
Val Hopwood and Lynn Pearce for their help and support throughout the Acupuncture MSc course.