Do Manual Therapies Help Low Back Pain?

A Comparative Effectiveness Meta-analysis

J. Michael Menke, DC, PhD

Study Design. Meta-analysis methodology was extended to derive comparative effectiveness information on spinal manipulation for low back pain.

Objective. Determine relative effectiveness of spinal manipulation therapies (SMTs), medical management, physical therapies, and exercise for acute and chronic nonsurgical low back pain.

Summary of Background Data. Results of spinal manipulation treatments of nonsurgical low back pain are equivocal. Nearly 40 years of SMT studies were not informative.

Methods. Studies were chosen on the basis of inclusion in prior evidence syntheses. Effect sizes were converted to standardized mean effect sizes and probabilities of recovery. Nested model comparisons isolated nonspecific from treatment effects. Aggregate data were tested for evidential support as compared with shams.

Results. Of 84% acute pain variance, 81% was from nonspecific factors and 3% from treatment. No treatment for acute pain exceeded sham’s effectiveness. Most acute results were within 95% confidence bands of that predicted by natural history alone. For chronic pain, 66% of 98% was nonspecific, but treatments influenced 32% of outcomes. Chronic pain treatments also fit within 95% confidence bands as predicted by natural history. Though the evidential support for treating chronic back pain as compared with sham groups was weak, chronic pain seemed to respond to SMT, whereas whole systems of clinical management did not.

Conclusion. Meta-analyses can extract comparative effectiveness information from existing literature. The relatively small portion of outcomes attributable to treatment explains why past research results fail to converge on stable estimates. The probability of treatment superiority matched a binomial random process.

From the Division of Psychology, School of Medicine, International Medical University, Kuala Lumpur, Malaysia.

Acknowledgment date: June 5, 2013. First revision date: October 21, 2013. Second revision date: December 22, 2013. Acceptance date: January 2, 2014. The manuscript submitted does not contain information about medical device(s)/drug(s).

No funds were received in support of this work.

No relevant financial activities outside the submitted work.

Address correspondence and reprint requests to J. Michael Menke, DC, PhD, Division of Psychology, School of Medicine, International Medical University, No. 126, Jalan Jilai Perkasa 19, Bukit Jilai, 57000 Kuala Lumpur, Malaysia; E-mail: michaelsmenke@imu.edu.my

DOI: 10.1097/BRS.0000000000000218

Spine
Pain is self-limiting; why treat it? Sixty percent to 70% of acute back pain resolves within 6 weeks without treatment; 90% within an year. Chronic back pain recovers 40% to 70% per year without treatment (Table 1). The extent of natural history effects within treatment arms must be subtracted to obtain treatment value.

**SPINAL MANIPULATIVE THERAPY**

According to a National Institutes of Health pamphlet, spinal manipulative therapy (SMT) is: “(performed by) practitioners...using their hands or a device to apply a controlled force to a joint of the spine...to relieve pain and improve physical functioning.”

A role for SMT in health care remains elusive as SMT outcomes shift from study to study irrespective of technique, provider, patients, diagnoses, or control variables. Perhaps SMT only works in a “whole system” context? A few studies in this analysis approximated whole systems where greater provider judgment and control was allowed.

**“SMOKE AND MIRRORS” OF NATURAL HISTORY AND NONSPECIFIC FACTORS**

To the hopeful stakeholder armed only with null hypothesis significance testing, 30+ years of indeterminacy suggests more research is needed. To others, continued ambiguity might imply ineffectiveness. Focusing on SMT to the exclusion of natural history and nonspecific components will trigger Type I errors, as described in the following text. Probability of recovery, pB, of 0.50 may seem respectable at first glance, but it is the same as the probability of a fair coin toss.

**GOALS OF A COMPARATIVE EFFECTIVENESS ANALYSIS**

The main objectives in this CER analysis were to (1) develop methods to capture CER information from existing data and (2) shed light on the role of SMT in the treatment low back pain. In meta-analyses, clinical trials are the units of study. For CER, trial arms are combined according to treatment.

**MATERIALS AND METHODS**

**Selected Studies**

Fifty-six SMT studies published from 1974 through 2010 were classified into 5 categories: SMT, exercise, physiotherapy modalities, usual medical care, and control groups. A sixth category was separated from the SMT group as whole system case management. Trial arms included 95 SMT, 40 control, 40 medical care, 51 physiotherapy, and 31 of exercise. Study quality ratings were included.

**Within-Arm Effect Sizes: Treatment, Natural Histories, and Nonspecific Factors**

Within-arm standardized mean change effect sizes were converted to Hedges g to correct for bias from small samples. Hedges g of 0.20 is a small effect, 0.5 is a medium effect, and 0.8 is a large effect. All standardized mean change effect sizes were anchored only to first observations within trial arms to reflect strength and stability. R software was used for all analyses.

**Hedges g and pB**

Effect sizes, g, were derived from visual analogue scales, McGill-Melzack, Oswestry, Likert pain ratings, 12-Item Short Form Health Survey, and 36-Item Short Form Health Survey pain scales. Visual analysis of g distributions exhibited the expected negative skew in self-limiting conditions. Hedge’s g was converted to a beta-distributed probability of recovery, pB, ranging from 0 to 1. pB is less apt to be biased by large samples and distribution shapes than effect sizes. For example, a g of 2.93 may be significantly greater than g = 2.04, while the difference in pB is only 1.8%.

**Nested Model Comparisons**

Nested model comparisons (NMC) analyses isolate treatment from total model variance, to identify the net contribution of treatment to clinical outcomes. Nonspecific factors tracked were patient attrition, number of visits, random study effects, pain classification, group average intercepts (for regression to the mean), publication year, quality, natural history, dose, study effects, and attended versus unattended therapies. NMC analyses performed on g and pB were nearly equivalent. Hedges g correlated with pB with Pearson r = 0.91.

**Evidential Support**

Evidential support (ES) gauged quality of treatment evidence compared with sham effects. Sham conditions for each pain type were chosen as standards of comparison because they contain maximum nontreatment information. ES is interpreted as follows. Positive values favor sham; −1 is weak support for treatment; −2 is moderate; −3 is strong; and −4 or less is extremely strong evidence favoring treatment.

**RESULTS**

Of 265 study arms, 68 studied acute pain (0–4 wk since onset). Thirty-seven studies reported on chronic pain (>12 wk duration). Among the other studies were 2 arms of subacute pain (5–12 wk), 74 arms of subacute-acute, and 31 of subacute-chronic pain. The pain type of 48 study arms was unclassified or unreported.
Publication Year and Study Quality
A linear regression found that overall SMT study quality improved by an average 1.2 of 100 quality points each year from 1974. The mean SMT effect was a very large $g = 1.57$ (95% CI: −0.30 to 3.49, $p_B = 0.72$), independent of study quality, and stable for 36 years (Figure 1). Bias was not detected by differences in study quality of those published in medical versus nonmedical journals.

Acute Pain
There were 257 study arms with 6397 acute patient observations within the first 6 weeks, and 150 study arms and 7991 observations thereafter (see Supplemental Digital Content, Appendix I, Tables A and B). The first 6 weeks of acute pain treatment (see Supplemental Digital Content Appendix I, Table A available at http://links.lww.com/BRS/A854) were dominated by natural history ($g = 1.60$, $p_B = 65\%$). Within the first 6 weeks, treatments with high $p_B$ included SMT from mixed or unknown sources (see Supplemental Digital Content Appendix I, line 6, $p_B = 1.00$ available at http://links.lww.com/BRS/A854), back school classes (line 7, 0.86), and attended physiotherapy modalities (e.g., electrical stimulation, hot packs, etc., line 9, 0.75). No acute pain treatments exceeded sham conditions in ES in the short term, and $p Bs$ attenuated after 6 weeks (see Supplemental Digital Content Appendix I, Table B available at http://links.lww.com/BRS/A854).

The NMC analysis corroborated the lack of value in treating acute back pain within the first 6 weeks. The saturated acute pain model identified 84% of outcome variance and 81% remained after removing treatment effects. Thus, a relative 96% (81/84) of acute pain improvement in the first 6 weeks was unrelated to treatment. Attention placebo nearly doubled the $p_B$ as shown in the difference between attended and unattended physiotherapies (see Supplemental Digital Content Appendix I, Table A available at http://links.lww.com/BRS/A854).
Content Appendix I, Table A, lines 9 and 13, \( pB = 0.75 \) vs. 0.39), along with the low performance of medical treatments and unsupervised self-care (see Supplemental Digital Content Appendix I, Table A, lines 8 and 10, \( pB = 0.21 \) and 0.33). Acute pain treatment evidence never exceeded sham (Table 2).

**Chronic Pain**

The effectiveness of chronic pain’s natural history is \( g = 0.52 \) with \( pB = 0.70 \). Seventy-nine study arms of 2455 patient observations constituted the chronic pain treatments for the first 6 weeks and 82 study arms of 3103 observations afterward (see Supplemental Digital Content Appendix I, Tables C and D available at http://links.lww.com/BRS/A854). Waiting list control arms approximated chronic pain natural history. Wait-listed patients got worse, \( g = -0.62 \) before 6 weeks and \( g = -0.13 \) and afterward. By contrast, sham for chronic pain increased \( pB \) for the short term (see Supplemental Digital Content Appendix I, Table C, line 11, \( pB = 0.45 \) available at http://links.lww.com/BRS/A854) and long term (see Supplemental Digital Content Appendix I, Table D, line 11, \( pB = 0.46 \) available at http://links.lww.com/BRS/A854).

The NMC analysis of chronic pain established 98% of outcome variance, of which 32% was from treatment and the balance 66% from everything else. Furthermore, treatment evidence beat shams. Figure 2 illustrates the comparative effectiveness in \( g \) for 6 treatments of chronic pain.

**Best Providers of SMT**

Which SMT is best? SMT has little influence on acute pain outcomes. Nonetheless, Figure 3 summarizes a direct comparison of SMT providers. Five SMT provider types were summarized: osteopaths (15 studies/15 arms), physical therapists (22/27), chiropractors (37/39), allopathic medical physicians (9/10), and bonesetters (3/3 for chronic pain only). Seven studies with 9 arms were from mixed providers. Physical therapy SMT was most effective, and most variable, in both acute (\( g = 2.09, 95\% \text{ CI: } 1.17–3.01 \)) and chronic pain (\( g = 2.48, 95\% \text{ CI: } 1.68–3.29 \)). Analyses of variance revealed no significant differences.

**Figure 2.** Effect sizes in the 6 principle treatment categories divided into acute or chronic low back pain for short term (≤6 wk) and long term (>6 wk and up to 150 wk for some studies). The boxplots are displayed against a gray scatter gram of bootstrapped expected effect sizes attributable to the passage of time alone.
Does Formal Education Enhance SMT Outcomes?

Do SMT providers need classrooms, lecturers, and laboratories to become effective practitioners? Are SMT providers born or made? Bonesetting is learned today as it has been for centuries: through apprenticeship. The bonesetting studies here were only for chronic pain. Bonesetting (see Supplemental Digital Content Appendix I, Table C, line 14 available at http://links.lww.com/BRS/A854) was large for the first 6 weeks of chronic care ($p_B = 0.57$ with weak evidential support). A weighted average of the other SMT outcomes for chronic pain within 6 weeks was 0.54 for all SMT providers including osteopathic and chiropractic flexion-traction SMT, and 0.69 when excluding those. Therefore, bonesetting sits well within the range of schooled SMT provider effectiveness. Education may add value in recognizing and referring pathologies, however.

SMT Versus Whole Systems

Whole systems could appear more “effective” because of additional nonspecific elements inherent in whole system patient encounters. This was not supported from the several studies that served as whole system approximations. Recovery from acute pain under whole system care (3 studies with 6 arms) was actually worse than under SMT alone (25 studies with 30 arms). Neither chiropractic SMT nor clinical management was evidentially superior to sham for acute pain (see Supplemental Digital Content Appendix I, Tables A and B, line 3 vs. 20 available at http://links.lww.com/BRS/A854). For chronic pain, $p_B$ for SMT effectiveness was nearly double that of whole system management for short term and long term (see Supplemental Digital Content Appendix I, Tables C and D, lines 3 and 20 available at http://links.lww.com/BRS/A854). Chiropractic SMT beat sham evidence for chronic pain but chiropractic whole system management did not. The whole system hypothesis was rejected.

Natural History, Regression to the Mean, and Patient Contact

Regression slopes correlating with intercepts is evidence of regression to the mean. Pearson’s $r_{acut}$ for slope-intercept correlation was 0.54, whereas $r_{chron} = 0.17$. As expected, regression to the mean was a bigger factor in acute pain, because self-report will be much more dynamic initially.

Sham effect sizes on acute pain were very large, $g = 1.40$ and 1.64 in the short term and long term, respectively. All acute effect sizes conformed to expectations from natural history as predicted by weeks of observation.
g = 0.46 × ln (wk) + 0.35

Chronic pain was, as expected, a “flatter” equation:

\[ g = -0.0027 \times \text{wk} + 1.09 \]

Figure 4 illustrates effect sizes, g, of treatment arms superimposed upon a background of expected outcomes from natural history, along with 95% confidence bands. Of 150 acute pain study arms, 5 (3%) were above the upper band, within a period of maximum regression to the mean. The outliers included: one SMT, one exercise, 2 clinical management, and one control condition. The coefficient of determination for natural history predicting acute effect sizes was \( R^2 = 99\% \).

Most of chronic pain also fell within 95% of effect sizes predicted by natural history alone. Three of 73 (4%) effect sizes fell above the 95% upper confidence bound: 2 for exercise and 1 for SMT. The coefficient of determination for natural history predicting chronic pain effect sizes was also \( R^2 = 0.99 \).

Studies of mixed diagnoses, listed at the end of Appendix II, \(^{49-56} \) were found in analyses to have effect sizes intermediate to acute and chronic pain. These studies were not included in Appendix I tables.

**Phantom Effectiveness From Undetected Randomness**

With strong natural history and nonspecific factors, treatments will seem superior at the same rate as control and comparison groups. However, clinical trials never include the count of times that controls beat treatment groups. Figure 5 and Table 2 corroborate that the pattern of SMT superiority is identical to a random draw from treatment arms.

**DISCUSSION**

From 1974 to 2010, 8400 SMT patients were observed at least 13,000 times in research costing from $32 to $80 million. Equivocal outcomes are unacceptable for the investment, leading some to propose a moratorium on future research.\(^ {56} \) Indeed, more and better clinical trials will only be more enigmatic when non-specific factors dominate clinical outcomes. Null hypothesis significance testing cannot detect placebo effects, natural history, and randomness. Meanwhile, SMT enthusiasts remain hopeful. More research is not the answer. That which is already known about SMT for back pain is quantifiably all that is worth knowing.

Fail-safe \( N \) calculations suggest it would take another 3000 (95% CI: 2100–3900) studies to change current effectiveness estimates.\(^ {57} \)

**TABLE 2. Confirmation of Spinal Manipulation Treatment Performance Compared With Expected Binomial Probabilities Based on Number of Treatment Arms**

<table>
<thead>
<tr>
<th>Count of Best Performing Treatment Arms by Method of Manipulation</th>
<th>Osteopathy</th>
<th>Chiropractic</th>
<th>Physical Therapy</th>
<th>Unspecified or Mixed</th>
<th>Medical</th>
<th>Bonesetters</th>
<th>Flexion Distraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 arms ( P ) (chance) = 50%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Better</td>
<td>11</td>
<td>21</td>
<td>10</td>
<td>3</td>
<td>7</td>
<td>*</td>
<td>5</td>
</tr>
<tr>
<td>Not better</td>
<td>11</td>
<td>21</td>
<td>10</td>
<td>3</td>
<td>9</td>
<td>*</td>
<td>5</td>
</tr>
<tr>
<td>Better/better + not better</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
<td>44%</td>
<td>*</td>
<td>50%</td>
</tr>
<tr>
<td>3 arms ( P ) (chance) = 33%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Not best</td>
<td>8</td>
<td>6</td>
<td>10</td>
<td>4</td>
<td>*</td>
<td>6</td>
<td>*</td>
</tr>
<tr>
<td>Best/best + not best</td>
<td>33%</td>
<td>33%</td>
<td>33%</td>
<td>33%</td>
<td>44%</td>
<td>*</td>
<td>33%</td>
</tr>
<tr>
<td>4 arms ( P ) (chance) = 25%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>*</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Not best</td>
<td>*</td>
<td>16</td>
<td>18</td>
<td>1</td>
<td>3</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Best/best + not best</td>
<td>*</td>
<td>27%</td>
<td>25%</td>
<td>50%</td>
<td>25%</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>5 arms ( P ) (chance) = 20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best</td>
<td>*</td>
<td>9</td>
<td>1</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Not best</td>
<td>*</td>
<td>36</td>
<td>4</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Best/best + not best</td>
<td>*</td>
<td>20%</td>
<td>20%</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

*No data available.
Limitations
Single-item outcome measures are psychometrically unreliable. Some authors fabricated pain assessments without proper vetting. Nonetheless, a disattenuation for unreliability item analysis would increase effect sizes reported here, and not affect conclusions.

Another anticipated criticism was that standardized SMT bear little resemblance to actual field practice and individual case management. Indeed, delivering unneeded SMT may harm. However, because more clinical freedom in approximate whole systems showed worse outcomes, this also is a moot issue.

Clinical and Policy Implications
Some providers with considerable knowledge about spines have claimed special back pain expertise. However, wide gaps remain between declarative and procedural knowledge. Future care guidelines should be based on results, not recitations. Results here support cautious observation, monitored exercise, and authoritative encouragement—services not requiring a licensed professional.

When all treatments seem equally effective but none stands out, more research will not help. Under these conditions, cheap treatments will always be the most cost effective. But for cost-effectiveness, you first need effectiveness. What decisions can be made when ineffective chiropractic care is more cost-effective than ineffective medical care? A correlation between cost-effectiveness and cost is a sign of weak or irrelevant treatment.

Other medical mysteries might be explained by within-arm dominance. Two examples include the effectiveness of “open label placebos” and that “most published research findings are false.” Homeopathy, laser therapy, energy medicine,
Spiritual healing, and many allopathic treatments with strong intentions, emotional investments, and therapeutic alliances between practitioner and patient probably exhibit within-arm dominance.

**Trick or Treatment?**

Complex biopsychosocial models do fail to predict and control back pain. Yet, biopsychosocial interventions can recalibrate patient expectations and alleviate anxiety. The acupuncturist’s Qi, chiropractic subluxations, and surgeon’s disc protrusions might serve as essential allegories in a therapeutic alliance. They promote perseverance and the vulnerability of self-regard as time and encouragement resolve 97% of acute and 67% of chronic pain.

Social support is the long ignored link between personal responsibility and professional care. For patients coping with pain and change, psychosocial support is necessary. The difference between sham effect size $g = 0.77$ and waiting list $g = -0.13$ for chronic pain illustrates the difference between attention and neglect. Patients in the sham group improved annually; waiting list controls got worse.

**CONCLUSION**

Physicians, payers, and policy makers should embrace psychosocial support without medicalizing or professionalizing a function that has thus far defaulted to alternative medicine. Medicalizing altruism with diagnoses, billing codes, reimbursement schedules, and explanations of benefits, only
adds opaque layers of complexity, emotional intensity, and professional defensiveness, along with inscrutable and costly maladies (wellness, Qi deficiency, and subluxations), taxing an already overburdened system. Internet illness communities may be the future of psychosocial support and personal resilience.

Key Points

- Spinal manipulation research produces equivocal results. Reasons may lie with inadequate analytics and methodology.
- Meta-analyses can produce comparative effectiveness information.
- All treatments of acute and chronic low back pain are much less effective than natural history and nonspecific factors.
- Back pain treatments serve to motivate, reassure, and calibrate patient expectations—features that might reduce medicalization and promote self-care.
- All conditions with predictable natural histories and richness in complex nonspecific factors should be reconsidered as to whether treatments are helpful or subordinate to the primary function of psychosocial support.

Supplemental digital content is available for this article. Direct URL citations appearing in the printed text are provided in the HTML and PDF version of this article on the journal's web site (www.spinejournal.com).

References