To what extent are surgery and invasive procedures effective beyond a placebo response? A systematic review with meta-analysis of randomised, sham controlled trials

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Published Version
doi:10.1136/bmjopen-2015-009655

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:23993516

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
ABSTRACT

Objectives: To assess the quantity and quality of randomised, sham-controlled studies of surgery and invasive procedures and estimate the treatment-specific and non-specific effects of those procedures.

Design: Systematic review and meta-analysis.

Data sources: We searched PubMed, EMBASE, CINAHL, CENTRAL (Cochrane Library), PILOTS, PsycINFO, DoD Biomedical Research, clinicaltrials.gov, NLM catalog and NIH Grantee Publications Database from their inception through January 2015.

Study selection: We included randomised controlled trials of surgery and invasive procedures that penetrated the skin or an orifice and had a parallel sham procedure for comparison.

Data extraction and analysis: Three authors independently extracted data and assessed risk of bias. Studies reporting continuous outcomes were pooled and the standardised mean difference (SMD) with 95% CIs was calculated using a random effects model for difference between true and sham groups.

Results: 55 studies (3574 patients) were identified meeting inclusion criteria; 39 provided sufficient data for inclusion in the main analysis (2902 patients). The overall SMD of the continuous primary outcome between treatment/sham-control groups was 0.34 (95% CI 0.20 to 0.49; p<0.00001; I²=67%). The SMD for surgery versus sham surgery was non-significant for pain-related conditions (n=15, SMD=0.13, p=0.08), marginally significant for studies on weight loss (n=10, SMD=0.52, p=0.05) and significant for gastroesophageal reflux disorder (GERD) studies (n=5, SMD=0.65, p=0.001) and for other conditions (n=8, SMD=0.44, p=0.004). Mean improvement in sham groups relative to active treatment was larger in pain-related conditions (78%) and obesity (71%) than in GERD (57%) and other conditions (57%), and was smaller in classical-surgery trials (21%) than in endoscopic trials (73%) and those using percutaneous procedures (64%).

Conclusions: The non-specific effects of surgery and other invasive procedures are generally large. Particularly in the field of pain-related conditions, more evidence from randomised placebo-controlled trials is needed to avoid continuation of ineffective treatments.

INTRODUCTION

Surgery and other invasive procedures such as endoscopy and percutaneous procedures are widely used in medicine but their specific efficacy and risk-benefit profile are rarely assessed in rigorous and systematic ways. The development of minimally invasive procedures has expanded the use of such interventions for treating a variety of conditions such as low-back pain, arthritis, endometriosis, Parkinson’s disease, gastro-oesophageal reflux and obesity.
Rarely are these procedures evaluated using rigorous research designs involving randomisation, allocation concealment and blinding or placebo controls, which are considered gold standards for medical interventions. In the absence of controls for common sources of bias, studies on these procedures may give a false impression of their true efficacy. Is it possible to test invasive procedures using rigorous methods? Blinding of outcome assessment is challenging since mimicking a complex, invasive procedure such as surgery, or insertion of a scope or a needle, requires an elaborate sham procedure. Moreover, there is significant controversy over the ethics of using sham procedures, even with carefully informed patients, further restricting the number of such studies being carried out. However, can we justify widespread use of these procedures without rigorous testing?

The use of blinded, sham procedures permits rigorous assessment of treatment efficacy by comparing the outcome in the treatment and sham groups. Specifically, sham procedures control for a variety of observed outcomes in the sham group that are distinct from the specific efficacy of the surgery or invasive procedure under investigation. These ‘non-specific’ outcomes include placebo responses (also sometimes called placebo effects), which we define here as the observed outcome changes in the sham groups. These changes are due to the natural history of the patient’s condition or regression to the mean and a response to the ritual of medical treatments. Such rituals include the type of procedure (pill, needle, knife or touch), the status, authority and communication style of the provider, the setting and context of the treatment and the patient’s and practitioners’ expectation about the outcome.

Yet, invasive procedures are thought to incorporate many factors that may contribute to the placebo responses including use of a hospital-like setting; multiple, authoritative providers; frequent and repeated suggestions about expected outcomes; a physical invasion of the body; and an elaborate ritual of treatment delivery and recovery. Thus, one would expect a significant contribution from surgical ritual and other non-specific factors to the observed outcomes during invasive procedures in clinical practice and in randomised trials without sham control groups. Several high profile studies support this hypothesis in which sham procedures involving only superficial anaesthesia were compared to the more invasive true procedure. For example, Moseley et al. reported no greater pain improvement in patients with osteoarthritis of the knee that underwent arthroscopic knee surgery compared to a sham procedure in which a cut was made over the knee without introducing the arthroscope. Two more recent controlled studies of vertebroplasty for painful osteoporotic vertebral fractures reported similar degrees of pain relief from sham procedures involving only superficial anaesthesia compared to the more invasive active procedures. In contrast, a systematic review comparing surgical with non-surgical treatments for painful osteoporotic vertebral fractures came to the conclusion that vertebroplasty and kyphoplasty are superior to non-surgical treatments. Since invasive interventions frequently go along with larger non-specific effects than non-invasive treatments surgical trials that do not include a sham surgery arm may give biased results. Thus, the efficacy of invasive procedures, for example, for chronic pain conditions, remains controversial.

In addition, many invasive procedures involve the risk of anaesthesia and high cost. Therefore, it is important to estimate to what degree the observed outcomes from invasive procedures are due to specific efficacy of the treatments or to other factors.

To better understand these issues we conducted a systematic review and meta-analysis of studies on surgery and invasive procedures in which a parallel sham procedure was included for comparison. Our study aims were to: (1) assess the quantity and quality of such studies; (2) estimate the magnitude of specific effects over sham procedures; and, (3) estimate the contribution of the surgical ritual and other non-specific factors to outcomes from these procedures.

**METHODS**

**Identification of studies**

The following online databases were searched from their inception through January 2015: PubMed, EMBASE, CINAHL, CENTRAL (Cochrane Library), PILOTS, PsycInfo, DoD Biomedical Research, clinicaltrials.gov, NLM catalog, as well as NIH Grantee Publications Database. We used as our initial search terms: ‘Diagnostic Techniques, Surgical’ OR ‘Orthopedic Procedures’ OR ‘Specialties, Surgical’ OR ‘Surgical Procedures, Operative’ OR ‘surgery’ (Subheading) or ‘surgery’ (and) ‘Placebos’ OR ‘Placebo Effect’ or sham surg* or placebo surg* or mock surg* or simulated surg* or placebo proc* or sham proc* or mock proc* or simulated proc*). We restricted our search to humans and randomised controlled trials. Variations of these search terms were made for MESH terms, where necessary, and are available on request from the first author.

The ‘Grey literature’ was searched by looking for relevant dissertations, conference proceedings, Google Scholar and searching the internet using the keyword scheme as well as searching all relevant reference lists of identified articles and related reviews. We also contacted and consulted with leading experts in the fields of surgery and placebo, and shared databases that these experts have collected over the years relating to placebo to make sure we captured all the relevant literature.

**Eligibility criteria**

Studies were included in the systematic review if they: (1) were randomised controlled trials; (2) involved a population for which there was a symptom-driven medical condition for which an invasive procedure or
classical surgery as defined below was being performed; and (3) had a comparison group that used a sham procedure to mimic the real procedure.

Classical surgery was defined as a procedure that followed the typical surgical experience that uses preoperative preparation, anaesthesia, an incisional trauma (usually through muscle and fascia and into the peritoneum) and a postoperative recovery process. Invasive procedures were defined as when an instrument was inserted into the body (either endoscopically or percutaneously) for the purpose of manipulating tissue or changing anatomy. In all cases we selected studies where when these procedures were compared to a sham procedure that used the same surgical or invasive procedure, instrument and ritual, but eliminated the hypothesised active component of tissue manipulation. We excluded studies in which the procedure was used simply as a delivery mechanism for another ongoing active treatment such as a pacemaker, brain or cardiac stimulation, or delivery of a drug or biological product. Studies where an invasive procedure was implemented for prevention of a medical condition or there was no symptom-driven condition were also excluded.

Four investigators (CC, LC, KL and KM) screened titles and abstracts for relevance in two phases based on the inclusion criteria: phase one eliminated all clearly irrelevant studies, phase two applied all inclusion/exclusion criteria listed above for the remaining studies. Any disagreements about including a study were resolved through discussion and consensus, and approved by the first author (WJ). All reviewers were fully trained in systematic review methodology. At least two reviewers had to review each citation in order for it to progress to the next phase of the review. A Cohen’s κ on agreement was attained for both phases above 88%.

Quality assessment and data extraction
The methodological quality of the individual studies (sequence generation, allocation concealment, was assessed independently by three reviewers using the Cochrane Risk of Bias (ROB) tool. Descriptive data was independently extracted on the following items: population; condition for which surgery was performed; sample (population) entered; dropout rate; informed consent details; whether a power calculation was performed and achieved; intervention and sham procedure used; primary and secondary outcomes and the statistical data associated with these; whether expectation was reported; author conclusions; adverse events reported; funding source, and reviewer comments. We also extracted from each study, if available, a continuous and a dichotomous main outcome at two time points (intermediate and late), and a continuous and a dichotomous pain outcome (when applicable). The most important outcome measure (mOM) was defined as either: (1) the primary main outcome measure (pOM) at a time point as predefined in the trial; or (if not 1), (2) the only major outcome of a trial at the latest available time point; or (if neither 1 nor 2), (3) the clearly most relevant outcome determined by two independent reviewers at the latest available time point. Secondary outcomes were intermediate time points of the most important outcome measure; pain outcomes at the latest available time point; or, pain outcomes at the intermediate time point. All discrepancies were tracked by the review manager and were resolved by consensus and discussions during team meetings. Data were entered into a web-based, secure, systematic review management programme called Mobius Analytics SRS (Mobius Analytics Inc, Ottawa, Ontario, Canada).

Data synthesis and analysis
According to our analysis plan, the meta-analyses focused on continuous outcomes. The primary analysis was based on trials reporting a most important continuous outcome measure in sufficient detail to be included in the meta-analysis. Secondary analyses were based on trials reporting (1) a continuous outcome measure at an intermediate time point, (2) a pain measure at a late time point, (3) a pain measure at an intermediate time point. Trials reporting only a dichotomous outcome measure (responder data) are noted in online supplementary table 1, and a sensitivity analysis was computed for these outcomes (see below).

Within-group and between-group effect sizes were based on Cohen’s d for change within one group, and Cohen’s d for between-group effect measures, respectively, correcting for small-sample bias. In order to keep the effect size framework coherent for within-group and between-group designs, change from baseline was used throughout. When SD was not reported, it was calculated from pre-SD and post-SD, using r=0.5 for the product-moment correlation between pre and post measures.

Analyses of continuous data were performed with the generic inverse variance module of the Cochrane Collaboration’s Review Manager software (V.5.1), using standardised mean difference (SMD) as the effect size measure. As we expected heterogeneity, a random effects model was used. Within-group effect sizes were pooled in such a way that positive values indicate improvement, while positive values of between-group effect sizes indicate superiority (more pronounced improvement) of the intervention group over the control (sham) group. To estimate the relative contribution of non-specific outcomes to treatment effects, the per cent ratio of the pooled within-group treatment effects in the sham and the treatment groups was calculated. We used Cochrane’s Q test and calculated I² to examine statistical heterogeneity, with low, moderate and high I² values of 25%, 50%, and 75%. Egger’s test was used to assess funnel plot asymmetry. A p value of less than 0.05 was set as the level of significance.

Subgroup analyses were performed according to predefined categories of target diseases and types of
surgery. To check the robustness of results, we performed sensitivity analyses with four criteria: (1) studies specifying a primary main outcome measure (pMOM); (2) imputing 0.3 and 0.7 for pre–post correlation coefficient r, when missing; (3) studies with total sample sizes ≥100; and, (4) studies with low risk of allocation concealment. An additional sensitivity analysis was performed for dichotomous outcomes of 12 studies that provided no continuous outcome (see online supplementary figure 1).

### RESULTS

#### Eligible studies

Our search identified a total of 7360 citations. After excluding clearly irrelevant references the full text of 113 publications were obtained. Of these, 46 were excluded, mainly for not including an instrumental or surgical intervention or a sham procedure as defined above. A total of 55 studies (in 67 publications) involving a total of 3574 enrolled patients met our inclusion criteria for systematic review (figure 1).26

#### Characteristics and quality of included studies

Characteristics of the included studies are summarised in online supplementary table 1. About half (25) of the studies were carried out on pain-related conditions with back pain (7) being the most frequent11 12 27–31 followed by arthritis (4),15 32–34 angina from coronary artery disease (4),35–39 abdominal pain (3),40–42 endometriosis (3),43–47 cholia (2)48 49 and migraine (2).50 51 The most frequently studied non-pain condition was obesity, especially when using balloon insertion (11).52–62 Other conditions that had more than one study included gastroesophageal reflux disease (GERD) (5),63–67 Parkinson’s Disease (2),68–74 sleep apnoea (2),75 76 dry eye (2)77 78 and asthma (2).79–81 Some other conditions were also studied (see online supplementary table 1).80–90 Many (22) of the studies involved endoscopic or percutaneous procedures in which tissue was removed or altered or a material (eg, dye, cement, balloon) was inserted.11–13 28 31 34 38–43 52 54 56 61 63 65 67 77–79 90 Some of these procedures used a catheter to reach an internal organ (such as the heart or gall bladder) or a needle to inject a material or cell (often into the lumbar spine or...
brain). Five studies evaluated more classical surgical procedures in which the body was opened with a scalpel or drill. In most studies, blinding was achieved using elaborate sham procedures. Those mimicking classical surgical procedures usually cut the body, leaving a scar but causing less damage than the real surgery. Sham percutaneous and endoscopic procedures often involved superficial insertion of a needle or a scope. For example, in the Parkinson’s studies on surgical interventions on the brain, sham procedures involved placing burr holes without penetration of the skull. Sham surgery for endometriosis would often involve ‘diagnostic laparoscopy’ with no internal tissue destruction. Sham balloon insertion for obesity treatment usually involved inserting the balloon but not inflating it.

Overall, the risk of bias was low in these studies, with some exceptions. Of the 55 studies (67 publications) included in the systematic review, 34 studies (62%) reported an adequate method for generating the allocation sequence, however only 23 (42%) had adequate concealment of allocation. Blinding of the patients and outcome assessors was adequate in 48 (87%) studies and incomplete data was adequately addressed in 52 (95%). Fifty-two (95%) of the studies were free from suggestion of selective outcome reporting and 53 studies were judged to be free of other sources of bias.

Overal analyses

Thirty-nine studies (2902 patients) with continuous data were included in the main analysis. The overall effect of surgery compared to sham surgery was highly significant (SMD 0.34, 95% CI 0.20 to 0.49; p<0.00001), while heterogeneity was large (I²=67%, p<0.00001). Excluding one outlier52 reduced I² to 57% (SMD, 0.30, 95% CI 0.17 to 0.43; p<0.00001), indicating moderate heterogeneity. Sensitivity analyses provided comparable effect sizes (figure 2), except for studies with overall sample sizes of 100 participants or more, for which the SMD was non-significant at 0.15 (n=10; 95% CI –0.02 to 0.32; p=0.09; I²=66%). Inspection of the funnel plot suggests the presence of biases in the meta-analysis, such as small study bias or publication bias (figure 3). Asymmetry in the funnel plot was confirmed by the Egger’s test (asymmetry coefficient 1.7, p=0.017).

Non-significant SMD were found when combining available data for the most important continuous outcome measure at an intermediate time point (n=14; SMD 0.12, 95% CI –0.05 to 0.29; p=0.17; I²=54%) as well as for specific pain outcomes at a late (n=14; SMD 0.12, 95% CI –0.03 to 0.27; p=0.11; I²=29%); or an intermediate time point (n=8; SMD 0.07, 95% CI –0.06 to 0.20; p=0.31; I²=0%).

Subgroup analyses of most important outcome measures

Subgroups by condition

Figure 4 summarises the SMD and subgroup means for between-group changes and the 95% CIs for each condition. The overall test for subgroup differences was significant (χ²=10.26, p=0.04), indicating significant heterogeneity of SMD between subgroups. Fifteen studies (analysing 1584 patients) included in the meta-analysis investigated pain-related conditions, the overall SMD was non-significant at 0.13 (95% CI –0.01 to 0.28; p=0.08; I²=46%). Ten studies (287 patients) reported on weight loss, the SMD was marginally significant at 0.52 (95% CI 0.01 to 1.03; p=0.05; I²=76%). Excluding one outlier52 reduced I² to 14% (SMD 0.27, 95% CI 0.00 to 0.55; p=0.05). Most (nine) of these studies involved balloon and sham balloon insertion. Five studies (342 patients) involved GERD. They showed a significant SMD of 0.65 (95% CI 0.31 to 1.00; p=0.0002; I²=55%). One study on Parkinson’s (34 patients) showed an SMD of 0.36 (95% CI –0.37 to 1.09). Eight studies (655 patients) on other diseases yielded a pooled SMD of 0.44 (95% CI 0.14 to 0.74, p=0.004; I²=57%).

Subgroups by type of procedure

Between-group SMD did not differ significantly between classical surgery, endoscopic surgery and percutaneous procedures (χ²=1.10, p=0.58; results not shown).

Dichotomous outcomes

Twelve studies provided only a dichotomous outcome measure (see online supplementary table 1). Sensitivity analyses showed an overall effect of surgery compared to sham surgery (risk ratio 1.54, 95% CI 1.11 to 2.15; p=0.01), while heterogeneity was large (I²=59%, p=0.005). Subgroup analyses according to condition revealed a significant effect of surgery versus sham surgery for pain studies (n=9; risk ratio 1.60, 95% CI 1.11 to 2.30; p=0.01; I²=59%, p=0.01) but not for other studies (n=3; risk ratio 2.19, 95% CI 0.44 to 10.84; p=0.33; I²=60%, p=0.08; see online supplementary figure 1).

Changes from baseline within sham and active groups

The pooled SMD for changes from baseline was 0.61 in the sham groups (95% CI 0.47 to 0.75, p<0.00001, n=39, I²=76%) and 0.92 (95% CI 0.74 to 1.09, p<0.00001, n=39, I²=86%) in the treatment groups. Thus, on average, the changes in the sham groups accounted for 65% of the overall improvement from the treatments. This proportion of specific to non-specific treatment effects was larger in pain-related conditions (78%) and obesity (71%) than in GERD (57%) and other conditions (57%), and was considerably smaller in classical surgery trials (21%) than in endoscopic trials (75%) and those using percutaneous procedures (64%; figure 4). Changes in the sham groups accounted for 89% and 82% of overall improvement in intermediate and late pain outcomes.

DISCUSSION

This is the first comprehensive systematic review with meta-analysis estimating the magnitude of the specific
effects of surgery and invasive procedures for various conditions. While some high profile studies have reported no difference between treatment and sham procedures, we found a positive though modest overall effect size (Cohen’s d) from the invasive procedures included in the analysis. When only larger studies (≥100 participants) are taken, the specific effects invasive procedures disappears, indicating the current evidence is

<table>
<thead>
<tr>
<th>Analysis, subgroup or study</th>
<th>N analysed</th>
<th>I²</th>
<th>ES [95% CI]</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall analysis</td>
<td>2902</td>
<td>67%</td>
<td>0.34 [0.20, 0.49]</td>
<td>&lt;.00001</td>
</tr>
<tr>
<td>Sensitivity analyses</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre–post correlation coefficient r=0.3</td>
<td>2902</td>
<td>65%</td>
<td>0.33 [0.19, 0.47]</td>
<td>&lt;.00001</td>
</tr>
<tr>
<td>Pre–post correlation coefficient r=0.7</td>
<td>2902</td>
<td>70%</td>
<td>0.37 [0.22, 0.52]</td>
<td>&lt;.00001</td>
</tr>
<tr>
<td>Primary outcome clearly defined</td>
<td>1325</td>
<td>66%</td>
<td>0.29 [0.07, 0.50]</td>
<td>0.009</td>
</tr>
<tr>
<td>Sample size at least 100</td>
<td>1741</td>
<td>66%</td>
<td>0.15 [0.02, 0.32]</td>
<td>0.09</td>
</tr>
<tr>
<td>Low risk of allocation concealment</td>
<td>1539</td>
<td>71%</td>
<td>0.34 [0.14, 0.54]</td>
<td>0.0008</td>
</tr>
<tr>
<td>Low risk of unblinding</td>
<td>2861</td>
<td>71%</td>
<td>0.36 [0.21, 0.51]</td>
<td>&lt;.00001</td>
</tr>
</tbody>
</table>

Subgroup analyses

Pain

| Abdominal pain | Swank 2003 | 100 | 0.17 [−0.22, 0.56] | 0.40 |
| Angina pectoris | Leon 2009 | 298 | 0.06 [−0.18, 0.30] | 0.62 |

Arthritis

| Moseley 2002 | 163 | −0.14 [−0.47, 0.19] |
| Bradley 2002 | 177 | −0.15 [−0.44, 0.14] |
| Silhoven 2013 | 146 | −0.09 [−0.47, 0.24] |

Endometriosis

| Abbott 2004 | 39 | 0.10 [−0.53, 0.73] | 0.75 |

Back pain

| van Keef 1999 | 31 | 0.90 [0.19, 1.61] |
| Leclaire 2001 | 66 | 0.14 [−0.35, 0.63] |
| Friedman 2005 | 55 | −0.43 [−1.17, 0.31] |
| Nath 2008 | 40 | 0.56 [−0.07, 1.19] |
| Buchbinder 2009 | 73 | 0.22 [−0.23, 0.67] |
| Kallmes 2009 | 131 | 0.02 [−0.33, 0.37] |
| Patel 2012 | 51 | 0.60 [0.01, 1.19] |

Migraine

| Dowson 2009 | 139 | 0.28 [−0.05, 0.61] |
| Guyuron 2009 | 75 | 0.68 [0.17, 1.19] |

GERD

| Corley 2003 | 52 | 0.68 [0.11, 1.25] |
| Deviere 2005 | 63 | 0.63 [0.34, 1.32] |
| Montgomery 2006 | 43 | −0.14 [−0.75, 0.47] |
| Rothstein 2006 | 144 | 0.76 [0.43, 1.09] |
| Schwartz 2007 | 40 | 1.07 [0.40, 1.74] |

Obesity

| Lindor 1987 | 21 | 0.58 [−0.30, 1.46] |
| Meshkinpour 1988 | 42 | 0.03 [−0.58, 0.64] |
| Mathus–Vliegen 1990 | 27 | −0.25 [−1.01, 0.51] |
| Rigaud 1995 | 20 | −0.05 [−0.93, 0.83] |
| Mathus–Vliegen 1996 | 17 | 0.16 [−0.80, 1.12] |
| Mathus–Vliegen 2002 | 42 | 0.38 [−0.23, 0.99] |
| Mathus–Vliegen 2003 | 28 | −0.03 [−0.79, 0.73] |
| Martinez–Brocca 2007 | 21 | 0.47 [−0.41, 1.35] |
| Gencio 2006 | 32 | 3.46 [2.33, 4.60] |
| Gersin 2010 | 37 | 1.03 [0.34, 1.72] |

Parkinson’s

| Olanow 2003 | 34 | 0.36 [−0.37, 1.09] | 0.33 |

Other

| Winters 1984 | 16 | 0.00 [−1.00, 1.00] |
| Larson 1998 | 155 | 0.48 [0.11, 0.86] |
| Slusser 1998 | 56 | 0.52 [−0.01, 1.05] |
| Friedman 2008 | 55 | 1.37 [0.78, 1.96] |
| Geids 2008 | 22 | 0.21 [−0.63, 1.05] |
| Koutsourelakis 2008 | 49 | 0.09 [−0.48, 0.66] |
| Castro 2009 | 288 | 0.17 [−0.06, 0.42] |
| Rodriguez 2009 | 14 | 0.77 [−0.45, 1.99] |

Subtotal (95% CI) | 655 | 57% | 0.44 [0.14, 0.74] | 0.004 |
not strong and could be changed with more and better research. In addition, the contribution of non-specific effects is even more substantial for certain conditions and procedures. While non-specific effects accounted for approximately 65% of the effects from all invasive procedures, they made up to 78% of the active treatment effects in chronic pain conditions and 71% of the active treatment effects in obesity. These percentages are substantially higher than those observed in non-surgical trials, namely 40% for chronic pain conditions and 33% for obesity. The higher contribution of non-specific effects in surgical trials could well be the result of higher placebo effects. However, the lack of no-treatment groups in our data set (and other data set) allows no firm conclusion. Our subgroup analyses indicate that the current evidence does not support the specific efficacy of invasive procedures for chronic pain conditions (p=0.08) and was borderline for obesity (p=0.05), but does support these procedures for GERD (p=0.0002). However, please note that the analysis of dichotomous outcomes showed a somewhat larger specific effect for pain studies (see online supplementary figure 1). There is insufficient data to make recommendations about the other conditions examined.

Strengths and weakness of this study
This study has several limitations. First, both the central strength and limitation of our study is that we pooled effect estimates of the included studies. We consider this a strength as it allows us to: (1) make an estimate of the overall effects of invasive procedures in sham-controlled surgical studies, (2) estimate the strength of confidence in the currently available data as to the specific efficacy of those procedures; and, (3) empirically investigate to what extent results differ between conditions and procedures. Obviously, it is not reasonable to expect that surgery has similar specific effects across conditions and outcomes so our subgroup estimates should not be interpreted clinically without considering how the interventions and outcomes varied. This is also indicated by the moderate-to-large heterogeneity in our meta-analyses, indicating more variation of effect sizes than would be expected by chance. Second, it is difficult to fully double-blind invasive procedures. While most studies successfully blinded patients and outcome assessors, physicians doing these procedures could not be blinded. Thus, it is possible that they communicated information to patients that biased the studies. Price and others have shown that physician expectations can influence pain outcomes even when restrictions are placed on verbal communication. Third, publication bias may play a role in the accuracy of our estimates. It is known that negative studies (in this case, studies showing no difference between real and sham procedures) are not published as frequently as positive studies. However, our search strategy was comprehensive and the study selection process was reliable. We also conducted a thorough search of the grey literature, as described above, and had input by experts in placebo research, increasing the likelihood of capturing all studies in this area. This activity allowed for a cross-check in the end to ensure we captured most of the relevant published randomised controlled trials for this review. We did not find any unpublished reports that met our inclusion criteria appropriate for this review, however there were some publications that were not readily accessible through the search engines commonly accessed that we were able to
capture through these methods. Our sensitivity analyses on study quality factors did not change our primary findings, except restricting the analyses to large studies with 100 participants and above, revealed a considerably smaller, non-significant SMD at 0.15 (95% CI −0.02 to 0.32; p=0.09). Egger’s test for funnel plot asymmetry, however, suggested a small study bias in our data set. While our combined estimates of effect size must be considered crude for the overall meta-analysis, they are reasonable estimates for the pain, GERD and obesity subgroups. Meta-analyses of placebo-controlled drug studies in pain, depression, hypertension, ulcer treatment and other areas often report a similar magnitude of specific treatment effects compared to non-specific effects.95–98 Those studies, however, usually have much larger sample sizes, increasing confidence in their estimates. Finally, we found only one three-armed study that included no treatment, active and sham groups.67 Therefore, it is not possible to estimate the contribution that the ritual and context make to outcomes in invasive procedures compared to no treatment. Especially in the field of pain and obesity such three-armed studies would seem to be essential for making good evidence-based decisions.

Our findings are consistent with a systematic review published in the *BMJ* in 2014.92 That study, however, used vote count and reported that 74% of 55 trials showed improvement in the placebo arm with 51% reporting no difference between surgery and placebo and 49% reporting surgery was superior to placebo. We have built on that study by doing a more comprehensive literature search and meta-analysis which allowed us to estimate the magnitude of surgical effects, the confidence in the current findings and to examine that magnitude across various quality parameters, conditions, procedures and outcomes. We can now conclude that at least chronic pain conditions lack clear evidence for the efficacy of the explored surgical interventions (eg, classic surgery and endoscopic procedures). Since these conditions represent a high public health burden worldwide we need to obtain better evidence for the use of these procedures. In addition, it is clear that the evidence from placebo controlled trials in the field as a whole is poor.

**Implications for practice, research and policy**

These results have a number of implications for practice, research and policy. The evidence from available sham-controlled trials indicates that invasive procedures are not clearly more effective than sham procedures for various chronic pain conditions including endometriosis, back pain, arthritis, angina and migraine. There is evidence to support surgical interventions for GERD and limited evidence to support the use of balloon insertion for obesity.

Given the large number of invasive and surgical procedures being performed, it is noteworthy that we could identify only 55 sham-controlled studies in the literature. Certainly, not all invasive procedures warrant sham-controlled comparisons; for example, when results demonstrate indisputable changes in objective parameters the risks of sham procedures would be excessive. However, given that non-specific factors make a large contribution to the effects from invasive procedures for conditions like pain, more rigorous evaluation is needed before their widespread use is recommended for these conditions. A recent survey of surgeon’s attitudes about sham surgery may provide an opportunity to conduct more such research. Surgeons generally agreed that a placebo component to surgical intervention might exist.99 Furthermore, results of a recent systematic review indicate that the risks of adverse effects associated with sham surgical procedures are small.92 Thus, more well-designed sham-controlled surgical trials are warranted to avoid the continued use of ineffective invasive treatments.

**Author affiliations**

1Samueli Institute, Alexandria, Virginia, USA
2Department of Pain and Translational Symptom Science, School of Nursing, University of Maryland, Baltimore, Maryland, USA
3Department of Anesthesiology, School of Medicine, University of Maryland, Baltimore, Maryland, USA
4Program in Placebo Studies, Beth Israel Deaconess Medical Center, Harvard Medical School Boston, Massachusetts, USA
5The Methodist Hospital, Houston, Texas, USA
6Department of Bioethics, Clinical Center, National Institutes of Health, Bethesda, Maryland, USA
7Department of Medical Psychology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
8Institute of General Practice, Technische Universität München, Munich, Germany
9Institute of Medical Psychology, Ludwig-Maximilians-Universität Munich, Munich, Germany

**Acknowledgements** The authors would like to thank Ms LaDonna Johnson, Research Assistant at Samueli Institute for assistance with article retrieval and tracking, and Ms Viviane Enslein for assistance with manuscript preparation.

**Contributors** WBJ served as the PI on this project and was responsible for the conception and design of the project, obtaining funding, acquisition of data and interpretation of the data, drafting and final revision of the article, and final approval of the version submitted. CC served as the project manager and reviewer, and contributed to the conception and design of the systematic review, acquisition of data and analysis and interpretation of data, drafting the article, and approval of the version to be submitted. LC and KM served as study quality reviewers and contributed to the conception and design, acquisition of data, analysis and interpretation of data, drafting the manuscript and approval of the version to be submitted. KL served as the statistical expert on the project and was involved in the design and conduct of the meta-analysis, acquisition of data, analysis and interpretation of the data, revising the manuscript critically for important intellectual content and approval of the version to be submitted. LK served as the project manager and reviewer, and contributed to the conception and design of the systematic review, acquisition of data and analysis and interpretation of data, drafting the article, and approval of the version to be submitted. LK served as the statistical expert on the project and was involved in the design and conduct of the meta-analysis, acquisition of data, analysis and interpretation of the data, revising the manuscript critically for important intellectual content and approval of the version to be submitted.


