Title

Blinding, sham, and treatment effects in randomized controlled trials for back pain in 2000–2019: A review and meta-analytic approach

Journal

Clinical Trials, 18(3)

ISSN

1740-7745

Authors

Freed, Brian
Williams, Brian
Situ, Xiaolu
et al.

Publication Date

2021-06-01

DOI

10.1177/1740774520984870

Peer reviewed

Published in final edited form as:

Clin Trials. 2021 June ; 18(3): 361–370. doi:10.1177/1740774520984870.

Blinding, Sham and Treatment Effects in Randomized Controlled Trials for Back Pain in 2000–2019: A Review and Meta-analytic Approach

Brian Freed[1], **Brian Williams**[2], **Xiaolu Situ**[3], **Victoria Landsman**[4,5], **Jeehyoung Kim**[6], **Alex Moroz**[7], **Heejung Bang**[3,8,9], **Jongbae J Park**[10]

[1].Department of Pain Management, Summit Medical Group, Berkeley Heights, NJ, USA

[2].Departments of Physiatry and Orthopedic Surgery, Hospital for Special Surgery, New York, NY, USA

[3].Graduate Group of Biostatistics, Department of Statistics, University of California, Davis, CA, USA

Address for correspondence: Jongbae J. Park, KMD, PhD, LAc, Department of Anesthesiology, Center for Translational Pain Medicine, Duke University School of Medicine, Durham, NC 27710, USA, Jongbae.Jay.Park@duke.edu.

Author Contributions/Statement:
Brian Freed: Conceptualization, Data curation, Writing - Original draft preparation
Brian Williams: Conceptualization, Data curation, Writing - Original draft preparation
Xiaolu Situ: Investigation, Validation, Methodology
Victoria Landsman: Investigation, Validation, Writing- Reviewing and Editing
Jeehyoung Kim: Visualization, Methodology, Formal analysis
Alex Moroz: Conceptualization, Methodology, Supervision, Writing- Reviewing and Editing
Heejung Bang: Formal analysis, Software, Visualization, Writing- Reviewing and Editing, Guarantor
Jongbae J. Park: Supervision, Methodology, Writing- Reviewing and Editing, Guarantor

Conflict of Interest: None declared.

Disclosure: Authors have disclosed no conflicts of interest.

Reproducible Research Statement: Study protocol: See the Supplement. Statistical code and dataset: Available from H. Bang (e-mail, hbang@ucdavis.edu).

Supplemental materials:
Table S1. Study overview of 40 trials we identified and analyzed
Table S2. Models fitted by GEE for effect size as Response variable, and log(sample size), year, outcome type (primary vs. secondary) and modality as Covariates
Table S3. Effect size of a. Secondary outcomes and b. Primary outcomes after excluding outliers (effect size larger than 2), by modality and overall
Table S4. Some sensitivity analyses: a. Pooled Blinding Index using different underlying models and b. Blinding Index for 2×2 and 2×3 separately
Figure S1. Systematic Review Search and Selection of Eligible Trials with Blinding Data
Figure S2. Blinding Index: Active (a: upper), Sham (b: middle), Active+Sham arms (c: lower)
Figure S3. Least squares means from Table S2: right panel (using 79 observations) and left panel (excluding outliers, effect size larger than 2)
Figure S4. Figures 1–3 with Secondary outcomes: within Active arm (a: upper left), within Sham arm (b: upper right), between Active and Sham arms (c: lower left)
Figure S5. Caterpillar plots with paired data and correlations
Figure S6. Figures 1–3 with Investigators' guess data – 10 observations from 5 trials (with Primary and Secondary outcomes combined)
Figure S7. Figures 1–3 with timepoints for primary outcome and blinding assessment closely matched: within Active arm (a: upper left), within Sham arm (b: upper right), between Active and Sham arms (c: lower left)
Supplement 1: Data cleaning and standardization processes (a) and References of 40 trials (b)
Supplement 2: Protocol submitted to PROSPERO and PRISMA checklist

[4]Institute for Work and Health, Toronto, ON, Canada

[5]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

[6]Department of Orthopedic Surgery, Seoul Sacred Heart General Hospital, Seoul, Korea

[7]Department of Rehabilitation Medicine, New York University Grossman School of Medicine, New York, NY, USA

[8]Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, CA, USA

[9]Center for Healthcare Policy and Research & Clinical and Translational Science Center Davis School of Medicine, University of California, Sacramento, CA, USA

[10]Department of Anesthesiology, Duke University School of Medicine, Durham, NC, USA

Abstract

Background: Blinding aims to minimize biases from what participants and investigators know or believe. Randomized controlled trials, despite being the gold standard to evaluate treatment effect, do not generally assess the success of blinding. We investigated the extent of blinding in back pain trials and the associations between participant guesses and treatment effects.

Methods: We did a review with PubMed/OvidMedline, 2000–2019. Eligibility criteria were back pain trials with data available on treatment effect and participants' *guess* of treatment. For blinding, Blinding Index was used as chance-corrected measure of excessive correct guess (0 for random guess). For treatment effects, within-/between-arm Effect Sizes were used. Analyses of investigators' guess/blinding or by treatment modality were performed exploratorily.

Results: Forty trials (3,899 participants) were included. Active and sham treatment groups had mean blinding index of 0.26 (95% CI: 0.12, 0.41) and 0.01 (−0.11, 0.14), respectively, meaning 26% of participants in active treatment believed they received active treatment, whereas only 1% in sham believed they received sham treatment, beyond chance, i.e., random guess. A greater belief of receiving active treatment was associated with a larger within-arm effect size in both arms, and ideal blinding (namely, 'random guess', and 'wishful thinking' that signifies both groups believing they received active treatment) showed smaller effect sizes, with correlation of effect size and summary blinding indexes of 0.35 (p=0.028) for between-arm comparison. We observed uniformly large sham treatment effects for all modalities, and larger correlation for investigator's (un)blinding, 0.53 (p=0.046).

Conclusion: Participants in active treatments in back pain trials guessed treatment identity more correctly, while those in sham treatments tended to display successful blinding. Excessive correct guesses (that could reflect weaker blinding and/or noticeable effects) by participants and investigators demonstrated larger effect sizes. Blinding and sham treatment effects on back pain need due consideration in individual trials and meta-analyses.

Keywords

Back pain; blinding; clinical trial; guess; meta analysis; placebo; systematic review

Introduction

Back pain is a significant and costly health problem. According to the 2017 Global Burden of Disease Study, nearly 577 million people suffer low back pain globally, as a leading cause of years lost to disability.[1] While estimates vary, it is generally accepted that costs associated with back pain in the U.S. approximate 90 billion dollars annually for diagnosis and treatment.[2–4] Recently, the National Institutes of Health established "Back Pain Consortium as part of the Helping to End Addiction Long-term (HEAL) Initiative", and emphasized a need for Phase-2 clinical trials (https://heal.nih.gov/research/clinical-research/back-pain), citing that current chronic low back pain treatment options are ineffective, which has led to an increased use of opioids.

Although there is a wealth of research evaluating various treatment modalities for back pain, uncertainty among clinicians remains regarding which treatments are most likely to be effective among specific patient types. The first comprehensive protocol for acute low back pain in the U.S. was released in 1994,[5] after which increased research, systematic reviews, and meta-analyses became available such that the American College of Physicians and the American Pain Society established joint guidelines for diagnosis and treatment of back pain in 2007 and 2017.[6–8] Current guidelines for acute and chronic pain stress avoidance of diagnostic imaging and include an array of pharmacologic and non-pharmacologic approaches; however, they do not address the relative strengths of these treatment modalities. Further, while the Guideline for the Evaluation and Management of Low Back Pain identified 11 criteria to assess the quality of clinical trials–including whether blinding was employed–it does not address if blinding was reasonably well maintained.[6,7] Blinding can be particularly important to subjective outcomes.

CONSORT 2010 removed the item about assessment/test of blinding, and there was no consensus on topics relative to study design, data collection, analytic approach, and interpretation of results.[9–15] In contrast, assessing the success of blinding was considered by consensus to be essential, so reporting on tests for blinding has therefore been added to the TIDieR checklist in 2020, along with detailed guidance.[16] CONSORT offers separate guidelines for non-pharmacological treatments[17–20]–it is accepted that there are inherent difficulties in blinding non-pharmacologic interventions, supporting the need to evaluate blinding techniques.[21] While the majority of randomized controlled trials (RCTs) under-assessed/reported blinding assessment,[13,17,22] there have been advances in statistical framework and methodologies regarding blinding in RCTs.[23–36] For example, a blinding index, a *chance-corrected* measurement of potential unblinding, was developed as a proxy for disproportionate correct guesses in a trial. Blinding index can evaluate blinding success and patterns, and help elucidate possible underlying scenarios in a systematic and standardized manner.

Back pain in reviews offers the advantage of having multiple treatment modalities for a common medical condition so standardized comparisons can be attempted. Utilizing the blinding index to assess blinding in back pain trials may elucidate a correlation between participants' perception of treatment received and treatment effect sizes–pre- and post-treatment and between-treatment changes. By comparing blinding index values with within-

arm and between-arm effect size changes, it may be possible to systematically study their relationships.

Our aims were to evaluate the degree of potential (un)blinding in back pain trials in 2000–19 (overall; by modality) and to assess correlation between potential unblinding and treatment effects using empirical data and validated, standardized measures–blinding index and effect size. We hypothesized: 1) blinding is unsatisfactory in back pain trials overall; 2) feasibility of blinding differs by modalities (e.g., medications are easier); and 3) weaker blinding is associated with larger effect sizes.

Methods

Data sources and searches

Eligible studies were limited to RCTs in humans, written in English, and published in 2000–19. Our search was not limited to specific treatment type, area, or diagnosis. Of note, because blinding assessment data are rare, we kept clinical inclusion and search criteria very broad. The protocol and checklist are in the supplemental material.

Study selection

We attempted a meta-analysis in 2013–14, which was put on hold due to reasons unrelated to study results, such as limited resources, authors' relocations, and complexity in study identification, data retrieval, and independent verification. In 2013–14, studies were initially screened by authors (Freed/Williams) in 2000–13 (call 'period 1') via OvidMedline and PubMed using keywords "randomized clinical trials" and "back pain"; these searches resulted in the same results. In 2019–20 when the effort resumed, studies published in 2014–19 ('period 2') were further screened by authors (Park/Bang), using the following search terms for initial screening by PubMed: "back pain"; "blind or mask"; and "randomized clinical/controlled trials". Two authors (Park/Bang) also double- and cross-checked previously identified trials (in 2000–13) independently. Blinding data are almost never mentioned in title/abstract/keywords, which are relatively easy to search, we focused on finding blinding data within text/tables/figures.

Although our primary searches and recordkeeping may be less rigorous than standard systematic reviews or meta-analyses, treatment guidelines and previous review papers[6,26,27,37,38] were carefully read to find eligible trials, because these sources can provide useful hints for a targeted and efficient search, not generally captured by standard searches. Furthermore, we asked trialists/experts in the fields if they are aware of any blinding data.

In screening, we looked for data on patient guesses of treatment allocation by manual searching and browsing/skimming, e.g., search words such as 'blind,' 'mask,' 'allocate,' 'conceal,' 'believe/belief,' and/or 'guess,' because data are often located around these words. Studies that asked participants to guess treatment group (active, control, with or without "do not know") in an analyzable format (e.g., counts, %, or blinding index) were included. When data were not complete or analyzable, we contacted authors. We studied investigators' blinding for *exploratory* analysis from 5 trials.

We recorded/classified modality into 4 categories: Acupuncture; Medication (oral, topical, intravenous); Physical and Chiropractic therapy (abbreviated as PT/Chiro); and Procedure (interventional); see descriptions and details in Table S1 (supplemental material).

Of note, because this review was concerned with evaluating a component of study methodology (blinding), rather than treatment efficacy/effectiveness as in traditional meta-analysis, we did not assess quality or risk of bias. Blinding status is based on authors' *self-designation*,[39] single/patient-blind may be a norm in back pain trials, and single-blind may convey higher quality than poorly maintained double-blind. Suboptimal practice persists in the usage/reporting of blinding in RCTs.[40,41]

Data extraction

Among the eligible studies, data regarding patient guesses of treatment assignment were obtained for each arm and tabulated in 2×2 or 2×3 formats. Blinding data were generally collected once: at the end of study or treatment, or before cross-over.

To calculate effect size for clinical outcomes, we retrieved sample $Mean_{pre}/SD_{pre}/Mean_{post}/SD_{post}$, where pre/post denote pre/post-treatment and SD denotes standard deviation. We used primary timepoint as stated in original publications following our study protocol. We intended to retrieve two clinical outcomes from each trial–primary and secondary. For primary outcome, we selected "pain," prespecified as sole or co-primary outcome. For secondary outcome, we included original non-pain primary outcome or next important pain outcome, as detailed in Table S1 (supplemental material).

Four authors (Freed/Williams in period 1; Park/Bang in period 2) retrieved data, and two authors (Park/Bang) checked data accuracy of all studies. One author (Situ) double-checked everything again. Within and over two periods, two persons tried to work as independently as possible, but some dependence was unavoidable for timely discussions on hard-to-find data and accurate retrieval.

Data synthesis and analysis

The number of participants' guesses of treatment group allocation was used to calculate blinding index for active and sham treatment groups in each study, where verum (V) represents active, experimental, real or new treatment, and sham (S) represents control or placebo treatment. Blinding index can be interpreted as the proportion of correct guess of treatment assigned *beyond chance* in each arm. Values range from −1 to 1: the closer to 0, the more random the guesses (50:50 or perfect balance); the closer to 1, the more correct the guesses; and the closer to −1, the more opposite or incorrect the guesses. Here, negative values near −1 are rare and their interpretation might be challenging, but can be less concerning as they do not indicate unblinding. For interpretation, an *ad-hoc* cutoff of 0.2 for blinding index has been used: >0.2 indicates more (excessive) correct guesses beyond chance; between −0.2 and 0.2 random guess; and <−0.2 more opposite guesses beyond chance.[26,38,42] Additionally, we adopt 'summed blinding indexes' as a summary measure to detect severe imbalance in the same guess between two arms because not all correct guesses are undesirable–namely, "wishful thinking" is a common psychological phenomenon and

indeed ideal blinding in RCTs, in which *both* parties tend to believe they received active treatment, results in summed blinding indexes=0.[26,38,43–46]

Regarding treatment effect, effect size was computed using (modified) Cohen-*d* for both within- and between-arm; within-arm effect size= $(Mean_{pre}—Mean_{post})/\sqrt{((SD_{pre}^2 + SDpost^2)/2)}$, and between-arm effect size=within-verum effect size—within-sham effect size.[47,48]

We computed Spearman correlation between blinding index and effect size values (within verum, within sham, between them) along with locally-weighted scatterplot smoothing (loess) fit to characterize nonlinear relationships between blinding index and effect size. We summarized blinding index and effect size by treatment modality, blinding index in caterpillar plots, and computed the 'pooled blinding index' with 95% confidence interval (CI) overall and by modality, via generalized estimating equations (GEE).[49]

We repeated the entire analyses for secondary outcomes. Multiple regression was fitted via GEE with effect size as a dependent variable and study characteristics (i.e., modality, year of publication, sample size, primary vs. secondary outcome) as independent variables. For sensitivity analyses, we did the following: 1) Pearson in place of Spearman; 2) blinding index under different statistical models;[49] 3) blinding index from 2×2 and 2×3 formats separately;[49] 4) investigators' blinding data instead for patients' (10 outcomes from 5 trials, with primary and secondary outcomes combined); and 5) exclude outliers in effect sizes (i.e., >2 for within or between). These analyses are reported with methods used for data processing/standardization in the supplemental material.

Finally, based on a reviewer's suggestion, we repeated our primary analysis with closely matching the timepoints of blinding assessment and outcome. Given that beliefs about group allocation can change over time,[15] analyses with primary outcomes might not capture the hypothesized mechanism: whether participant belief about their allocation (at that moment) influences clinical effects of the treatment. This could be important for self-reported outcomes because belief about group allocation could influence self-reported pain. SAS 9.4 and R were used for data analyses.

Results

Data search

Based on our search criteria, 1,643 articles published from 2000–13 were found via OvidMedline and PubMed when two authors conducted initial searches in 2013–14. After screening abstracts, 587 articles were found to potentially meet inclusion criteria. Of these, 22 included data on participant guesses of treatment allocation. OvidMedline and PubMed resulted in the same set of trials finally selected. One additional study was identified from a treatment guideline paper. In 2019, two authors independently resumed the project, and found 11 studies out of 296 candidates in 2014–19 with PubMed. Additionally, 6 studies in 2000–13 were identified in 2019 through checking systematic reviews and meta-analyses and contacting some trialists. For 2000–19, a total of 40 studies representing 3,899 participants were identified and included in analyses; see Figure S1 & Supplement.

Regarding modality, there were 10 acupuncture, 8 medication, 9 PT/Chiro, and 13 procedure studies. All 40 trials had primary pain outcomes data and 39 had data for secondary outcomes. Mean/median of sample size were 97/63, with the range of min=15 and max=692.

Blinding index values

Blinding index values for the 40 studies are displayed in caterpillar plots with study-specific 95% CIs for within-arm and between-arm analyses; see Table 1 & Figure S2. Blinding index showed a mean of 0.26 (95% CI: 0.12, 0.41) for verum, and of 0.01 (95% CI: −0.11, 0.14) for sham. This indicates 26% of participants allocated to the verum group thought they received active treatment, while only 1% allocated to sham thought they received sham treatment, beyond chance. These phenomena may be interpreted as “excessive correct guess” in verum and “random guess” in sham.[44] Summed blinding indexes showed the range of (−0.19, 1.34), away from “wishful thinking” of value 0, implying that less ideal blinding scenarios were common.

Effect size and its correlation with blinding index

Effect size showed a range of (−0.35, 3.1) within sham, (0, 4.7) within verum, and (−1.0, 2.9) between arms–there was a large effect size within each arm, where 0.5 is regarded as ‘medium’ and >0.8 as ‘large’.[47] Associations between blinding index and effect size was analyzed: 1) within verum; 2) within sham; and 3) between arms. Resulting correlations were: 0.31 (p=0.05) within verum; −0.21 (p=0.19) within sham; and 0.35 (p=0.03) between arms from 40 trials/observations. Thus, positive correlation within verum, negative correlation within sham, and positive correlation between-arms were observed, also confirmed by loess, suggesting that greater belief of receiving active treatment in either arm, or larger imbalance in same guess between arms, was associated with larger effect sizes. Conversely, ‘random guess’ (blinding index≈0 for sham and verum) and ‘wishful thinking’ (summed blinding indexes≈0) showed smaller effect sizes, as demonstrated in Figures 1–3.

Blinding index and effect size by modality

Table 1 presents the analyses by modality. Regarding effect size, largest within-arm effect sizes were seen in acupuncture studies (mean/median=1.59/1.47 within verum; 1.22/1.31 within sham; 0.37/0.16 between arms). In all modalities, there were moderate to large within-arm effect sizes (≥0.5) for sham, indicating substantial sham/placebo effect, contributing to low between-arm effect size (<0.25 in median). Medication trials showed lowest blinding index values for verum (<0.1) and largest for sham (>0.2), possibly due to low efficacy in verum and sham, contrary to our original hypothesis. In sham, PT/Chiro and procedure studies showed blinding index≈0, somewhat surprising, but with the *widest* 95% CI of (−0.53 to 0.53), possibly due to high heterogeneity/diversity in modalities. When we checked summed blinding indexes, PT/Chiro yielded largest value=0.62, farthest away from the “wishful thinking” scenario.

Secondary and sensitivity analyses

First, three regression models using effect size (within verum, within sham, between them) as a dependent variable, and log(sample size), year of publication, outcome type (primary

vs. secondary) and modality as covariates with n=79 observations (40 primary and 39 secondary outcomes) are presented. Year was not statistically significant in all 3 models ($p \geq 0.4$), while studies with larger enrollment showed larger effect size in sham ($p=0.01$), perhaps partly explained by 'small-study effects' documented in meta-analyses.[48] Second, between-arm effect sizes tended to be low, with PT/Chiro largest and medication smallest (≈ 0) with broadly overlapping CIs; see Tables S2–S3 & Figure S3 in the supplemental material.

Third, correlation analyses with secondary outcomes or Pearson correlation demonstrated qualitatively similar trends in direction of association between blinding index and effect size, although correlations were attenuated in between-arm analyses, as seen in Figure S4. Figure S5 shows correlations among blinding indexes, when studies were paired. Negative correlation between blinding indexes for sham and verum (correlation=−0.52, $p=0.0005$) might support the common "wishful thinking" scenario. Fourth, when we excluded outliers in effect size (2 acupunctures in primary and secondary outcomes, 1 procedure in primary outcome, and 1 PT/Chiro in secondary outcome), results were attenuated, while main findings were mostly unchanged; some results in Figure S3 & Table S3. When blinding index was computed with different statistical models[49] or data structures, main findings were similar, although 2×3 data yielded attenuation[27,30,49] (Table S4).

Intriguingly, investigators' blinding data showed larger correlations in magnitude, e.g., −0.67 ($p=0.034$) and 0.53 ($p=0.045$) in the small subset, which might imply investigators' correct guess was associated with larger effect sizes, compared to the patients' counterpart; see Figure S6. Finally, when we repeated the analyses with closely matching the timepoints of blinding assessment and outcome, we found qualitatively consistent results despite reduced statistical significance, similar trends to results for secondary outcomes (Figure S7).

Discussion

Previous reviews reported that a small percent of RCTs (2–8% in different fields) provided some evidence on blinding success.[22,27,50,51] Among studies that asked participants to guess treatment allocation included in our review, the pooled blinding indexes indicate participants in active treatment groups had a greater proportion of correct guesses beyond chance, while those in sham groups overall had random guessing, which may imply successful implementation of sham treatment. This may be promising because historically non-pharmacological interventions (such as PT/Chiro, acupuncture, interventional procedures) were perceived very difficult to blind.[21] Yet, we argue if real PT/Chiro is difficult to blind, then sham PT/Chiro should be difficult to blind as well. Therefore, potentially successful blinding of sham in PT/Chiro (blinding index for sham≈ 0, indicating perfect balance but least precise estimate, which we could not figure out) and procedure trials warrants more discussion. Despite these results, PT/Chiro showed least desirable blinding performance when we considered the *final* between-arm comparison (largest summed blinding indexes). Contrary to our hypothesis, medication studies–perceived to be easiest to blind–did not show best blinding performance, along with lowest effect sizes, which might be related to the complex nature of back pain, including psychosocial and structural components which may not be amenable to medications.[52]

In associations between blinding indexes and effect sizes, we observed that a stronger belief in receiving active treatment was associated with larger within-arm effect sizes in both arms. In between-arm comparison, an excess of correct or imbalanced guess (far away from 'wishful thinking') was associated with a larger between-arm effect size. Conversely, more ideal blinding scenarios ('random guess' or 'wishful thinking', mathematically speaking, allocation and guess are independent) showed smaller effect sizes.[43,45] These findings should be validated in future. Bi-directionality in cause-and-effect is likely, say, large treatment effect causing unblinding or small effect preventing unblinding.

When we reviewed previous comparable meta-analyses, Braithwaite et al. observed no evidence of a moderating effect of blinding index on pain in <30 dry needling trials (reporting summed blinding indexes of −0.2 to 0.5; blinding index of 0.37/−0.26 in verum/sham),[26,38] whereas Colagiuri et al.[27] found some moderators associated with worse blinding outcomes in 23 pharmacological trials for chronic pain (blinding index of −0.71 to 0.87 in verum vs. −0.64 to 0.32 in sham). Moroz et al.[23] reported blinding index of 0.34/−0.2 for verum/sham and the most common blinding scenario was that participants believed they received verum regardless of the treatment received in 54 acupuncture trials, namely, wishful thinking–similar to our subgroup analysis and Braithwaite et al.'s finding in comparable treatments. In contrast, Freed et al. and Baethge et al.[24,25,49] showed the overall blinding in 40 psychiatric trials is summarized as Random–Random (blinding index of 0.14/0.0 in verum/sham), and reported positive correlation between effect size and correct guesses.

We acknowledge limitations. First, with our not-state-of-art search skills and methods used (e.g., two time-periods, less uniform and systematic, not using specialized terminologies) and relatively low number of trials, our review may not be representative of entire back pain trials, introducing substantial selection bias. We may have missed studies, relying on manual screening and skimming/scanning of entire papers written in English identified by PubMed. The number of the trials we identified, however, is not very different from blinding index-based meta-analyses cited above. Language bias (non-English) and unsearchable *file drawers* may be a larger issue (aka, elephant in the room), beyond the scope of our study and capacity.[53,54] Although we tried to be as prospective as possible (e.g., statistical analysis protocol determined before screening/data extraction in period 1, which can be viewed as a strength), we could not achieve that goal fully (e.g., two periods; protocol registration attempted after period 1, and evolving statistical methodologies); we explained our own experiences about registration, prespecification and deviations in the supplemental material for interested readers. Prospective methods are nowadays recommended standard practice in pain-related research to maximize transparency.[55] Also, there is more scrutiny in some modalities than others, thus having more motivation in assessment and reporting of blinding data.[56]

Second, limitations on data available/used were unavoidable; for instance, participant blinding was studied more rigorously in our primary analysis, because of only 5 studies with clinician/investigator blinding data–must be less feasible or attempted in practice. Most studies reported blinding data at only one timepoint. Some pros and cons of single vs. repeated assessments have been discussed.[15,44]

Third, limitations on statistical measures should be noted. Blinding index serves as a numerical 'proxy' measure of potential unblinding *vis a vis* an excessive proportion of correct guesses; correct guess is not always undesirable because it can reflect noticeable therapeutic effects or hard-to-avoid adverse events, or natural psychological phenomenon, "wishful thinking". Also, RCT papers with continuous outcomes rarely report 'correlation of pre and post outcomes' or 'SD of their difference', which is needed to compute *correct* effect size; we hope trialists will report one of these quantities routinely for future meta-analyses. And we could not address causality or directionality, or provide meaningful insight.

Fourth, we did not assess quality/risk of bias although we provided some rationale. A custom risk of bias assessment to capture biases may provide useful information. For example, selective reporting bias–if successful blinding leads to smaller effect sizes for clinical outcomes, and outcomes with smaller effect sizes are less likely to be reported, then pooled effect size may be exaggerated. Blinding of other key parties (e.g. outcome assessors, clinicians) could also influence results; if participants were successfully blinded, but blinding status of other parties was unclear, this might weaken the relationship between effect size and blinding index because both variables could be moderated by other sources of bias. Future investigations could explore these issues or follow good examples of tailored bias assessments related to blinding.[26,38]

Notwithstanding, our study provides a review for two decades of literature and presents some interesting or unique findings. First, the notion of a blinded sham control in back pain appears feasible in non-pharmacologic treatments such as acupuncture, PT/Chiro and procedure. Interestingly or counterintuitively, they performed better than medications in terms of blinding, but we are unable to provide explanation(s). This is particularly important in the practice of back pain RCT because sham treatments demonstrated *uniformly high* within-arm effects. If real/active treatment is too difficult to blind, trialists should aim at creating active control or strong placebo effect, i.e., aiming at summed blinding indexes=0, instead of blinding index=0 for sham. Our finding may add more support to accumulating evidence toward potential benefits of placebo and sham treatments with complexity in research and clinical contexts/settings.[57,58]

Second, our study conveys some methodologic strengths: back pain is an ideal medical condition to study our aims because multiple treatment modalities are available, which made a comparative review possible. Third, we used standardized and validated metrics for quantifying the blinding (blinding index) and treatment effects (effect sizes). This contrasts with many previous reviews based on self-designation (especially in title) or authors/reviewers' subjective assessments of blinding success, or meta-epidemiological approaches; thus, without actual guess data.[59–61] Additionally, we checked internal consistency by analyzing primary and secondary outcomes, and matching timepoints for blinding and clinical data. To our knowledge, our study is the first study of its kind based on raw blinding/guess data from back pain trials with multi-modalities.

Finally, we may have validated some historical or conventional beliefs, such as blinding can exert greater influence on subjective outcomes such as pain, and a stronger belief of

receiving an active treatment (e.g., placebo effect) can show larger (within-group) treatment effects. Our findings may have implications on increasingly popular patient-reported outcomes research, device trials (e.g., digital health) and adapted designs (with advanced Bayesian analyses), in which subjective outcomes, self-reports and difficult or loosened blinding are common. With ignored/waived blinding (which will reduce the cost of trial design and operation, and make trialists' lives easier) and substantial sham effect, many treatments ineffective or with exaggerated effects could be unduly approved or widely used in practice (which may incur higher price to patients and society).[62–70]

Conclusion

Our study showed more correct guess about treatment received is associated with larger treatment effects in back pain trials. The main advantage of blinding assessment is an easy implementation in practice (minimally, one simple question once or twice), and the main disadvantages are still extra work involved and difficulty in interpretation (bi-directionality). Future meta-analyses could facilitate evaluation and interpretation–hopefully with more collection of blinding data (in a voluntary manner) and reporting–not only on treatment effect, but also for blinding itself toward the quality assurance and reliability of RCT evidence.[16]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement:

We thank Ms. Caron Modeas for her English editing service. We thank Editor, Associate Editor and Reviewers that helped greatly improve our manuscript and authors' learning.

Grant support: H. Bang was partly supported by the National Institutes of Health, through grant numbers UL1 TR001860 and R01 AR076088.

Funding: H. Bang was partly supported by the National Institutes of Health, through grant numbers UL1 TR001860 and R01 AR076088. The funding source had no role in the study design or implementation.

References

1. James SL et al. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018;392(10159):1789–1858. [PubMed: 30496104]
2. Martin B, Deyo R, Mirza S, et al. Expenditures and health status among adults with back and neck problems. JAMA 2008;299(6):656–664. [PubMed: 18270354]
3. Dagenais S, Caro J and Haldeman S. A systematic review of low back pain cost of illness studies in the United States and internationally. Spine J 2008;8(1):8–20. [PubMed: 18164449]
4. Luo X, Pietrobon R, Sun S, et al. Estimates and patterns of direct health care expenditures among individuals with back pain in the United States. Spine 2004;29(1):79–86. [PubMed: 14699281]
5. Agency for Health Care Policy and Research (AHCPR). Acute low back problems in adults. Clinical Practice Guideline. In: Department of Health and Human Services PHS, ed. Rockville, MD.1994.

6. Chou R, Deyo R, Friendly J, et al. Nonpharmacologic therapies for low back pain: a systematic review for an American College of Physicians Clinical Practice Guideline. Annals of Internal Medicine 2017;166(7):493–505. [PubMed: 28192793]
7. Chou R, Qaseem A, Snow V, et al. Diagnosis and treatment of low back pain: A Joint Clinical Practice Guideline from the American College of Physicians and the American Pain Society. Annals of Internal Medicine 2007;147(7):478–491. [PubMed: 17909209]
8. Qaseem A, Wilt T, McLean R, et al. Noninvasive treatments for acute, subacute, and chronic low back pain: A Clinical Practice Guideline From the American College of Physicians. Annals of Internal Medicine 2017;166(7):514–530. [PubMed: 28192789]
9. Schulz KF, Altman DG, Moher D, et al. CONSORT 2010 changes and testing blindness in RCTs. Lancet 2010;375(9721):1144–1146. [PubMed: 20338625]
10. Hopton AK and Macpherson H. Assessing blinding in randomised controlled trials of acupuncture: challenges and recommendations. Chin J Integr Med 2011;17(3):173–176. [PubMed: 21359917]
11. Schulz KF and Grimes DA. Blinding in randomised trials: hiding who got what. Lancet 2002;359:696–700. [PubMed: 11879884]
12. Kolahi J, Bang H, Park J, et al. CONSORT 2010 and controversies regarding assessment of blindness in RCTs. Dental Hypotheses 2010;1(2):99–105.
13. Greenfield M, Mhyre J, Mashour G, et al. Improvement in the quality of randomized controlled trials among general anesthesiology journals 2000 to 2006: a 6-year follow-up. Anesth Analg 2009;108(6):1916–1921. [PubMed: 19448222]
14. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. Journal of the American Medical Association 1996;276:637–639. [PubMed: 8773637]
15. Rees J, Wade T, Levy D, et al. Changes in beliefs identify unblinding in randomized controlled trials: a method to meet CONSORT guidelines. Contemporary Clinical Trials 2005;26:25–37. [PubMed: 15837450]
16. Howick J, Webster R, Rees J, et al. TIDieR-Placebo: A guide and checklist for reporting placebo and sham controls. PloS Med 2020;17(9):e1003294. [PubMed: 32956344]
17. Boutron I, Guittet L, Estellat C, et al. Reporting methods of blinding in randomized trials assessing nonpharmacological treatments. PLoS Med 2007;4:e61. [PubMed: 17311468]
18. Boutron I, Moher D, Altman D, et al. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Annals of Internal Medicine 2008;148(4):295–309. [PubMed: 18283207]
19. Boutron I, Estellat C, Guittet L, et al. Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: a systematic review. PloS Med 2006;3:e425. [PubMed: 17076559]
20. Borg Debono V, Zhang S, Ye C, et al. The quality of reporting of RCTs used within a postoperative pain management meta-analysis, using the CONSORT statement. BMC Anesthesiol 2012;12:13. [PubMed: 22762351]
21. Boutron I, Tubach F, Giraudeau B, et al. Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials. Journal of Clinical Epidemiology 2004;57(6):543–550. [PubMed: 15246122]
22. Kolahi J, Bang H and Park J. Towards a proposal for assessment of blinding success in clinical trials: up-to-date review. Community Dentistry and Oral Epidemiology 2009;37(6):477–484. [PubMed: 19758415]
23. Moroz A, Freed B, Tiedemann L, et al. Blinding measured: a systematic review of randomized controlled trials of acupuncture. Evidence-Based Complementary and Alternative Medicine 2013:708251. [PubMed: 23533515]
24. Freed B, Assall O, Panagiotakis G, et al. Assessing blinding in trials of psychiatric disorders: A meta-analysis based on blinding index. Psychiatry Research 2014;219:241–247. [PubMed: 24930582]
25. Baethge C, Assall O and Baldessarini R. Systematic review of blinding assessment in randomized controlled trials in Schizophrenia and Affective Disorders 2000–2010. Psychotherapy and Psychosomatics 2013;82:152–160. [PubMed: 23548796]

26. Braithwaite F, Walters J, Li L, et al. Blinding strategies in dry needling trials: systematic review and meta-analysis. Physical Therapy 2019;99(11):1461–1480. [PubMed: 31373369]
27. Colagiuri B, Sharpe L and Scott A. The Blind leading the not-so-blind: a meta- analysis of blinding in pharmacological trials for chronic pain. Journal of Pain 2019;20(5):489–500.
28. Arandjelović O A new framework for interpreting the outcomes of imperfectly blinded controlled clinical trials. PLoS One 2012;7(12):e48984. [PubMed: 23236350]
29. Arandjelović O Clinical trial adaptation by matching evidence in complementary patient sub-groups of auxiliary blinding questionnaire responses. PLoS One 2015;10(7):e0131524. [PubMed: 26161797]
30. Landsman V, Fillery M, Vernon H, et al. Sample size calculations for blinding assessment. Journal of Biopharmaceutical Statistics 2018;28(5):857–869. [PubMed: 29157126]
31. Zhang Z, Kotz R, Wang C, et al. A causal model for joint evaluation of placebo and treatment-specific effects in clinical trials. Biometrics 2013;69(2):318–327. [PubMed: 23432119]
32. Bang H, Ni L and Davis CE. Assessment of blinding in clinical trials. Controlled Clinical Trials 2004;25:143–156. [PubMed: 15020033]
33. Walter S, Awasthi S and Jeyaseelan L. Pre-trial evaluation of the potential for unblinding in drug trials: a prototype example. Contemporary Clinical Trials 2005;26(4):459–468. [PubMed: 16054578]
34. James KE, Bloch DA, Lee KK, et al. An index for assessing blindness in a multi-centre clinical trial: disulfiram for alcohol cessation - a VA cooperative study. Statistics in Medicine 1996;15(13):1421–1434. [PubMed: 8841652]
35. Houweling A, Shapiro S, Cohen J, et al. Blinding strategies in the conduct and reporting of a randomized placebo-controlled device trial. Clinical Trials 2014;11:547–552. [PubMed: 24902921]
36. Hertzberg V, Chimowitz M, Lynn M, et al. Use of dose modification schedules is effective for blinding trials of warfarin: evidence from the WASID study. Clinical Trials 2008;5(1):23–30. [PubMed: 18283076]
37. Machado L, Kamper S, Herbert R, et al. Imperfect placebos are common in low back pain trials: a systematic review of the literature. European Spine Journal 2008;17:889–904. [PubMed: 18421484]
38. Braithwaite F, Walters J, Li L, et al. Effectiveness and adequacy of blinding in the moderation of pain outcomes: Systematic review and meta-analyses of dry needling trials. PeerJ 2018;6:e5318. [PubMed: 30083458]
39. Moore S, Neylon C, Eve M, et a. “Excellence R Us”: university research and the fetishisation of excellence. Palgrave Communications 2017;3:16105.
40. Lang T and Stroup D. Who knew? The misleading specificity of “double-blind” and what to do about it. Trials 2020;21:697. [PubMed: 32758278]
41. Park J, White AR, Stevinson C, et al. Who are we blinding? A systematic review of blinded clinical trials. Perfusion 2001;14:296–304.
42. Park J, Bang H and Canette I. Blinding in clinical trials, time to do it better. Complementary Therapies in Medicine 2008;16:121–123. [PubMed: 18534323]
43. Bang H Random guess and wishful thinking are the best blinding scenarios. Contemporary Clinical Trials Communications 2016;3:117–121. [PubMed: 27822568]
44. Bang H, Flaherty SP, Kolahi J, et al. Blinding assessment in clinical trials: A review of statistical methods and a proposal of blinding assessment protocol. Clinical Research and Regulatory Affairs 2010;27(2):42–51.
45. Mathieu E, Herbert R, McGeechan K, et al. A theoretical analysis showed that blinding cannot eliminate potential for bias associated with beliefs about allocation in randomized clinical trials. Journal of Clinical Epidemiology 2014;67:667–671. [PubMed: 24767518]
46. Anderson T, Reid DB, Beaton GH. Vitamin C and the common cold: a double- blinded trial. Can Med Assoc J. 1972;107:503–508. [PubMed: 5057006]
47. Kadel R and Kip K. A SAS Macro to compute effect size (Cohen’s) and its confidence interval from raw survey data. Southeast SAS® Users Group (SESUG); 2012.

48. Schwarzer G, Carpenter J and Rucker G. Meta-Analysis with R. Switzerland: Springer; 2015.
49. Landsman V, Landsman D, Li C, et al. Overdispersion models for correlated multinomial data: Applications to blinding assessment. Statistics in Medicine 2019;38(25):4963–4976. [PubMed: 31460677]
50. Hrobjartsson A, Forfang E, Haahr MT, et al. Blinded trials taken to the test: An analysis of randomized clinical trials that report tests for the success of blinding. International Journal of Epidemiology 2007;36:654–663. [PubMed: 17440024]
51. Fergusson D, Glass KC, Waring D, et al. Turning a blind eye: The success of blinding reported in a random sample of randomised, placebo controlled trials. BMJ 2004;328:432. [PubMed: 14761905]
52. Ramond-Roquin A, Bouton C, Gobin-Tempereau A, et al. Interventions focusing on psychosocial risk factors for patients with non-chronic low back pain in primary care--a systematic review. Fam Pract 2014;31(4):379–388. [PubMed: 24632524]
53. Young S and Bang H. The file-drawer problem, revisited. Science 2004;306:1133–1134.
54. Phillips C Publication bias in situ. BMC Med Res Methodol 2004;4:20. [PubMed: 15296515]
55. Lee H, Lamb S, Bagg M, et al. Reproducible and replicable pain research: a critical review. Pain 2018;159(9):1683–1689. [PubMed: 29697535]
56. Coeytaux R and Park J. Acupuncture research in the era of comparative effectiveness research. Annals of Internal Medicine 2013;158(4):287–288. [PubMed: 23420237]
57. Kaptchuk T, Friedlander E, Kelley JS, et al. Placebos without deception: a randomized controlled trial in irritable bowel syndrome. PLoS One 2010;5(12):e15591. [PubMed: 21203519]
58. Schneider T, Luethi J, Mauermann E, et al. Pain response to open label placebo in induced acute pain in healthy adult males. Anesthesiology 2020;132:571–580. [PubMed: 31809325]
59. Moustgaard H, Clayton G, Jones H, et al. Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study. BMJ 2020;368:16802.
60. Godlee F Blinding may be unnecessary, but please divest. BMJ 2020;368:m255.
61. Armijo-Olivo S, Fuentes J, da Costa B, et al. Blinding in physical therapy trials and its association with treatment effects: a meta-epidemiological study. American Journal of Physical Medicine & Rehabilitation 2017;97(1):34–44.
62. Mahajan R and Gupta K. Adaptive design clinical trials: Methodology, challenges and prospect. Indian J Pharmacol 2010;42(4):201–207. [PubMed: 20927243]
63. Al-Lamee R, Thompson D, Dehbi H, et al. Percutaneous coronary intervention in stable angina (ORBITA): a double-blind, randomised controlled trial. Lancet 2018;391(10115):31–40. [PubMed: 29103656]
64. Beard D, Rees J, Cook J, et al. Arthroscopic subacromial decompression for subacromial shoulder pain (CSAW): a multicentre, pragmatic, parallel group, placebo-controlled, three-group, randomised surgical trial. Lancet 2018;391(10118):329–338. [PubMed: 29169668]
65. Cohen S, Wallace M, Rauck R, et al. Unique aspects of clinical trials of invasive therapies for chronic pain. Pain Reports 2019;4(3):e687. [PubMed: 31583336]
66. Wan M, Orlu-Gul M, Legay H, et al. Blinding in pharmacological trials: the devil is in the details. Arch Dis Child 2013;98(9):656–659. [PubMed: 23898156]
67. Shapiro S Risks of estrogen plus progestin therapy: a sensitivity analysis of findings in the Women's Health Initiative randomized controlled trial. Climacteric 2003;6:302–310. [PubMed: 15006251]
68. Garbe E and Suissa S. Issues to debate on the Women's Health Initiative (WHI) study. Hormone replacement therapy and acute coronary outcomes: methodological issues between randomized and observationals studies. Human Reproduction 2004;19(1):8–13. [PubMed: 14688150]
69. Prasad V and Cifu A. The necessity of sham controls. The American Journal of Medicine 2019;132(2):e29–30. [PubMed: 30076810]
70. Flum D In reply. Sham surgery in clinical trials. JAMA 2007;297:1546.

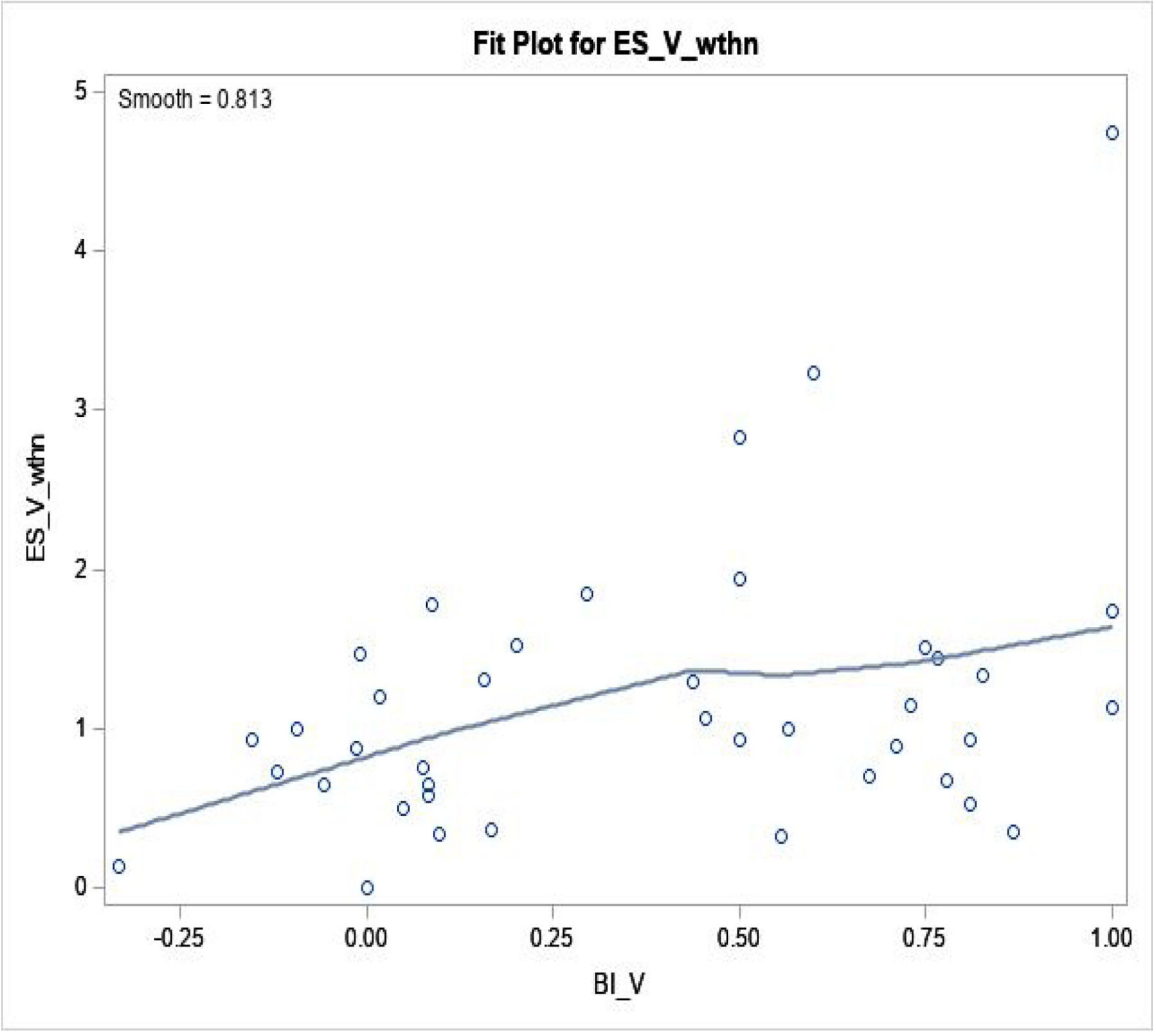

Figure 1. Blinding Index and Effect Size: within Active Arm

Regression fit was done by loess fit, in 40 trials.

Spearman correlation coefficient is 0.31 (p=0.05).

$ES_{V,\ wthn}$: effect size within active arm (Y-axis); BI_V: blinding index in active (X-axis).

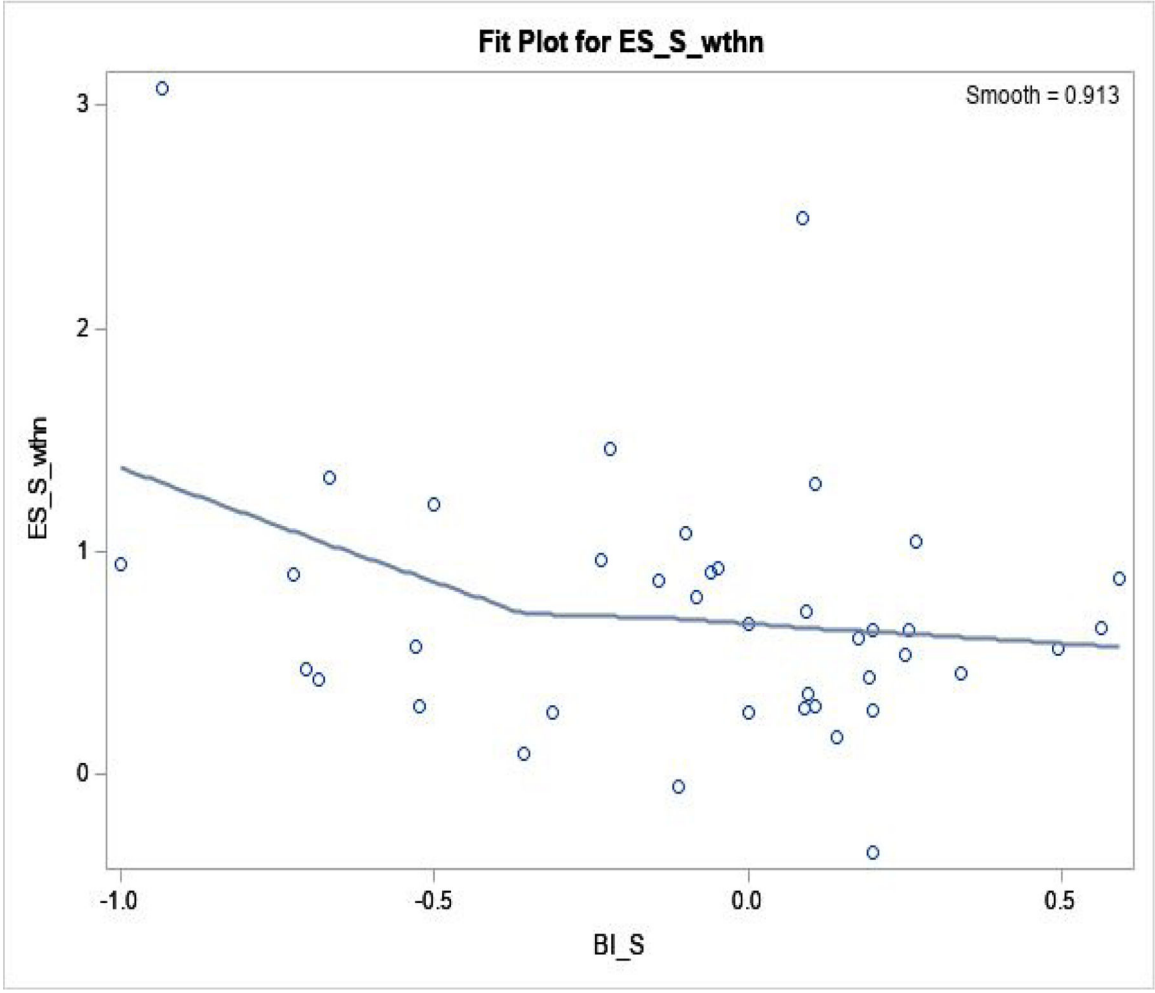

Figure 2. Blinding Index and Effect Size: within Sham Arm

Regression fit was done by loess fit, in 40 trials.

Spearman correlation coefficient is −0.21 (p=0.19).

$ES_{S,\ wthn}$: effect size within sham arm (Y-axis); BI_S: blinding index in sham (X-axis).

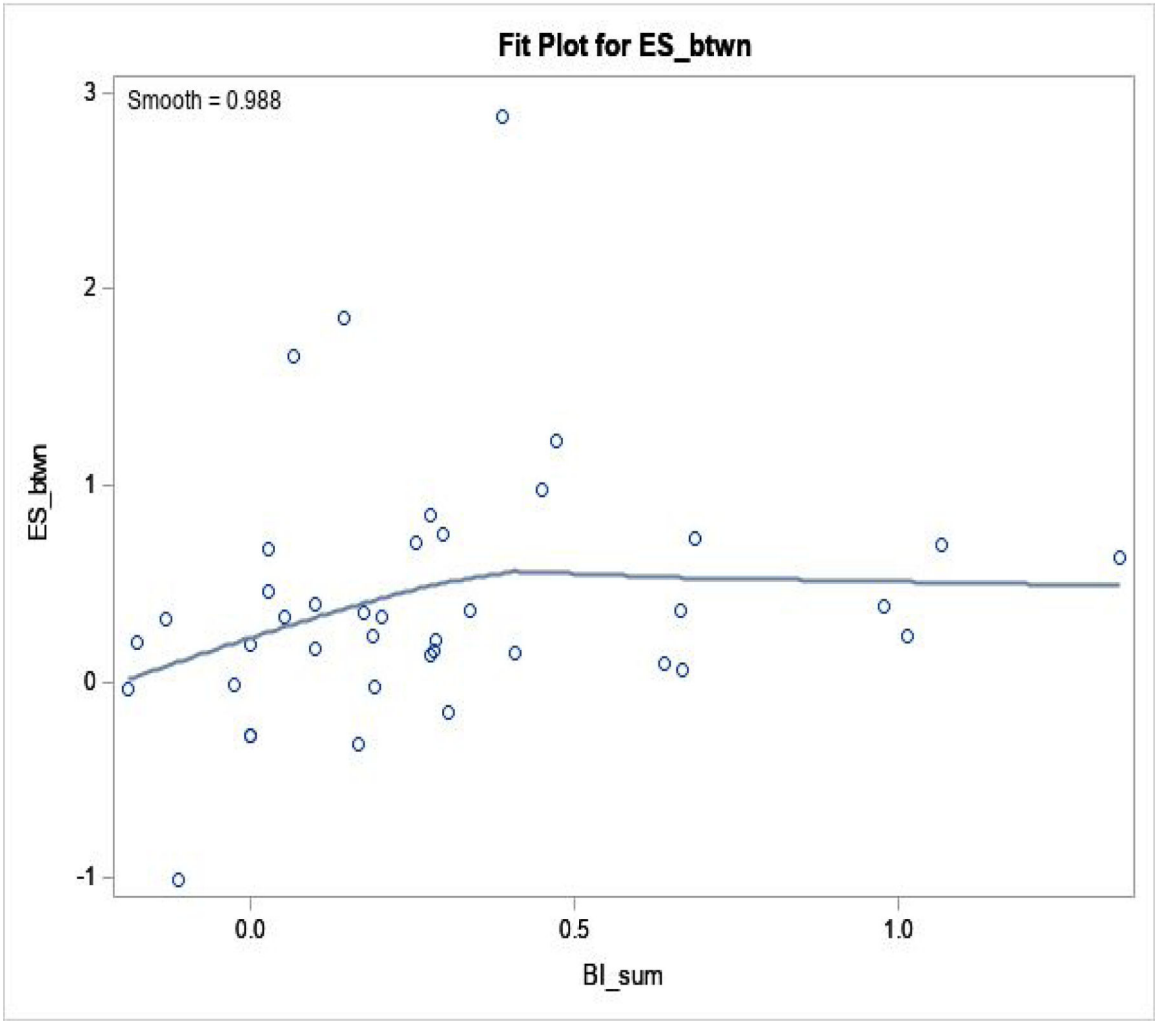

Figure 3. Blinding Index and Effect Size: between Active and Sham Arms

Regression fit was done by loess fit, in 40 trials.

Spearman correlation coefficient is 0.35 (p=0.03).

When we excluded 3 outliers (with ES>2), 0.35 (p=0.03) was unchanged.

ES_{btwn}: effect size between active and sham arms (Y-axis); $BI_{sum}=BI_V+BI_S$ (X-axis).

Table 1.

Blinding Index Values and Treatment Effect Sizes Overall and by Modality

Modality (n=number of trials)	Blinding index Mean (95% confidence interval)			Effect size Mean/Median		
	Active arm	Sham arm	Sum*	Active arm	Sham arm	Between arms
Acupuncture (n=10)	0.25 (−0.10, 0.60)	−0.09 (−0.40, 0.21)	0.13	1.59/1.47	1.22/1.31	0.37/0.16
Medication (oral/topical/intravenous) (n=8)	0.09 (−0.14, 0.32)	0.24 (0.07, 0.41)	0.32	0.76/0.88	0.57/0.65	0.19/0.23
Physical & Chiropractic Therapy (n=9)	0.62 (0.38, 0.86)	0.00 (−0.53, 0.53)	0.62	0.95/0.75	0.63/0.46	0.32/0.24
Procedure (n=13)	0.18 (0.01, 0.35)	−0.01 (−0.19, 0.18)	0.17	0.98/0.93	0.75/0.62	0.23/0.14
Overall (n=40)	0.26 (0.12, 0.41)	0.01 (−0.11, 0.14)	0.27	1.14/1.00	0.85/0.74	0.29/0.20

*Summed blinding indexes.

Confidence interval was estimated via Generalized Estimating Equations (GEE) with working independence correlation (Landsman et al. 2019).

Effect size weighted by sample size.

Blinding index (range in −1 to 1) can be interpreted as the proportion of correct guess within a given arm beyond chance: 0 means “random guess” (50:50 or perfect balance); >0 means more correct guesses; and <0 means more incorrect/opposite guesses.

Summed blinding indexes measure the difference in proportions in the *same guess*: ‘Sum=0’ means an equal proportion of participants in both arms believed they received active treatment (or sham), so called “wishful thinking” (or “negative thinking” for sham, which is rare); ‘Sum>0’ means more participants in active treatment arm believed they received active treatment, compared to those in sham arm who believed they received active treatment.

Value of 0 for 1) both blinding indexes for active and sham, or 2) summed blinding indexes is the two most ideal blinding scenarios, implying ‘random guess’ or ‘wishful thinking’, respectively. Effect size may be interpreted with Cohen’s d; 0.2 small; 0.5 medium; >0.8 large as a rule of thumb.

See Table S3 for effect size calculations after excluding 3 studies with very large effect size (>2), which are 2 acupuncture and 1 procedure trials.