Title

Assessing blinding in trials of psychiatric disorders: A meta-analysis based on blinding index

Journal

Psychiatry Research, 219(2)

ISSN

0165-1781

Authors

Freed, Brian
Assall, Oliver Paul
Panagiotakis, Gary
et al.

Publication Date

2014-10-01

DOI

10.1016/j.psychres.2014.05.023

Peer reviewed

Published in final edited form as:
Psychiatry Res. 2014 October 30; 219(2): 241–247. doi:10.1016/j.psychres.2014.05.023.

Assessing blinding in trials of psychiatric disorders: A meta-analysis based on blinding index

Brian Freed[a,*], **Oliver Paul Assall**[b], **Gary Panagiotakis**[a], **Heejung Bang**[c], **Jongbae J. Park**[d,e], **Alex Moroz**[a], and **Christopher Baethge**[b]

[a] The Center for Musculoskeletal Care and Department of Rehabilitation Medicine, New York University School of Medicine, New York, NY 10016, USA

[b] Department of Psychiatry and Psychotherapy, University of Cologne Medical School, Cologne, Germany

[c] Division of Biostatistics, Department of Public Health Sciences, University of California-Davis, Davis, CA 95616, USA

[d] Asian Medicine and Acupuncture Research, Department of Physical Medicine and Rehabilitation, UNC-Chapel Hill, School of Medicine, Chapel Hill, NC 27599-7200, USA

[e] Center for Pain Research and Innovation, UNC School of Dentistry, Chapel Hill, NC 27599-7455, USA

Abstract

The assessment of blinding in RCTs is rarely performed. Currently most studies that do report data on evaluation of blinding merely report percentages of correct guessing, not taking into account correct guessing by chance. Blinding assessment using the blinding index (BI) has never been performed in a systematic review on studies of major psychiatric disorders. This study is a systematic review of psychiatric randomized control trials using the BI as a chance-corrected measurement of blinding, a tool to analyze and understand the patterns of blinding across studies of major psychiatric disorders with available data. Of 2467 psychiatric RCTs from 2000 to 2010, 66 reported on blinding and 40 studies were found to have enough information on evaluation of blinding to be analyzed using the BI. The experimental treatment groups had an average BI value of 0.14 and the control groups had an average BI value of 0.00. The most common BI scenario was random–random, indicating ideal blinding. A positive correlation between effect size and more correct guesses was also found. Overall, based on BI values and the most common blinding scenario, the published articles on major psychiatric disorders from 2000 to 2010, which reported on blinding assessment for patients, were effectively blinded.

* Correspondence to: New York University School of Medicine, The Center for Musculoskeletal Care, 333 E, 38th Street, 5th FL Room 5-103, New York, NY 10016, USA. Brian.Freed@nyumc.org (B. Freed)..

5. Disclosure

The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institute of Health.

Appendix A. Supplementary materials

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.psychres.2014.05.023.

Keywords

Methodology; RCT; Double-blind trials; Blinded trials; Schizophrenia; Affective disorders; Therapy

1. Introduction

Blinding is an important method to reduce bias of many randomized controlled trials (RCT) whenever relevant and feasible. The ability to achieve its intended goal of an unbiased study is variable and affected by study methods, study execution, or participants' (and raters') individual beliefs. Blinding assessment can detect potential unblinding which may be indicative of limited trial validity.

A potential effect of blinding on study outcome has been indicated by various studies. With regard to subjective outcomes, for example, Wood et al. (2008) have shown smaller effects in blinded trials than in unblinded studies.

However, it is difficult – if not impossible – to know cause from effect in unblinding. Whether it is unblinding influencing the effects of a study or the effects of a study resulting in unblinding. Therefore, interpretation can be difficult and blinding assessment is not required as per the Consort 2010 guidelines (Schulz et al., 2010) – a change with respect to the 2001 guidelines (Moher et al., 2001); Consort only recommends the reporting of who was blinded and how blinding was attempted.

Recent systematic reviews have shown poor reporting of whether and how blinding was applied throughout RCTs in various fields of research (Rios et al., 2008; Greenfield et al., 2009; Reveiz et al., 2010; Borg Debono et al., 2012; Ghimire et al., 2012; Péron et al., 2012; Turner et al., 2012; Baethge et al., 2013). According to Baethge et al. (2013), over the past decade very few (2.5%) published RCTs on major psychiatric disorders have mentioned assessment of blinding, and only 54.1% of those reported details of blinding success.

Currently, blinding is measured by asking participants which treatment they believe they have been assigned (Fergusson et al., 2004; Boutron et al., 2005; Hróbjartsson et al., 2007; Baethge et al., 2013). In general, such analyses are based on dichotomous answers (e.g., verum or placebo), or they include a third option ("don't know"). However, merely reporting percentages so-derived does not take into account chance agreement. To overcome this disadvantage blinding indices (BIs) have been developed (James et al., 1996; Bang et al., 2004).

So far, BI analysis has not been applied to psychiatric RCTs. Blinding could be particularly critical in psychiatric research because soft or subjective outcomes often serve as study endpoints. Baethge et al. (2013) looked at the overall percentage of studies performing assessment of blinding and percentages of participants correct guesses of treatment allocation groups. This study, based on the work of Baethge et al. (2013), is intended to apply a BI in order to take chance agreement into account and analyze data in a more standardized fashion. The BI serves as an objective tool to estimate the percentage of

excessive correct guesses as a measure of the degree of potential unblinding beyond chance in a given arm of a study. That is, a chance controlled percentage of participant's correct guess of the treatment assigned (Bang et al., 2004). Using the BI it is possible to more rigorously evaluate and interpret the pattern of blinding of trials of psychiatric disorders (e.g., quantitatively based on BI values and qualitatively based on different blinding scenarios) and to quantify the extent of blinding in relation with effect size and types of intervention, i.e. pharmacologic or non-pharmacologic, in a systematic approach.

2. Methods

2.1. Data sources and searches

This study builds upon the systematic review by Baethge et al. (2013) that laid out methods in detail. In brief, the literature was searched using Medline and PubMed-Central databases. The search was limited to psychiatric randomized trials in either English or German published between 2000 and 2010, and it was narrowed with regard to affective disorders and to schizophrenia.

2.2. Study selection

Screening by Baethge et al. (2013) included full text searches for studies presenting single, double, or triple blinded RCTs of a psychiatric diagnosis predominantly, but not necessarily restricted to, schizophrenia and affective disorders as diagnosed by standard international criteria (DSM-IV or ICD-10).

In the current analysis one author (B.F.) separated the studies selected by Baethge et al. (2013) based on the reporting of treatment guess data. Independent verification of all treatment guess data was performed by another author (G.P.). A study was considered to have reported treatment guess data if participants were surveyed to guess whether they believed they were receiving Experimental (E) or Control (C) treatment, with or without the option of “Do not know” (DK), so that data could be tabulated in a 2 × 3 or 2 × 2 format. When blinding evaluation data was unclear or incomplete, contact with the primary author by E-mail was attempted twice. Studies with completed data were included, while non-responding authors and incomplete data were excluded.

If effectiveness of blinding (EOB) was measured multiple times within a study, only the EOB measured at the end of the study was included. If a study was determined to have two distinct experimental and control arms with independent EOB results, it was included as two individual studies. Studies with multiple experimental treatment arms but a single control arm were included as a single study combining experimental treatment arm results.

2.3. Data extraction

As the primary distinctive requirement of this meta-analysis, the quantity of subject guesses, E, C, and DK was extracted from each study. Also, it was recorded whether a study's experimental treatment group was pharmacologic or non-pharmacologic.

2.4. Data synthesis and analysis

The BI was used to statistically analyze and interpret EOB. BI values can range from 1 to –1. For example, if all guesses are correct, BI=1; if all guesses are incorrect, BI= −1; and if 50% of guesses are correct and 50% are incorrect then BI=0. Thus the more unblinded a study is based on participant guesses of treatment group the BI will be closer to −1, while if more participants are convinced they are in the opposite group the BI will be closer to 1. If guessing is completely random, as an ideally blinded study should be, BI will equal 0. With this ability to quantify correct/incorrect guesses within individual treatment arms of a study, while taking chance agreement into account, it is then possible to compare and further analyze blinding patterns within and across studies. Nine blinding scenarios have been proposed to evaluate and interpret the pattern of blinding (Table 1) (Park et al., 2008; Bang et al., 2010).

We then examined experimental BI (eBI) values, control BI (cBI) values, eBI+cBI from individual studies, and overall averages of eBIs and cBIs. Using both the calculated eBI and cBI values, blinding within a study was interpreted based on one of nine blinding scenarios as shown in Table 1 (Park et al., 2008). In order to separate studies into categories, BI values of ≥0.2 were considered more correct guesses beyond chance, −0.2<BI<0.2 were random guesses, and BI values ≤ 0.2 were opposite guesses. [Remark: these values are rules of thumb rather than definite cutoffs, and they are used here for exploratory classification purposes only. Also, we used the conventional term ‘unblinded’ for easier communication (Shapiro, 2003; Bang et al., 2010; Moroz et al., 2013) but it should be understood as “more correct guesses”.

eBI and cBI were computed from each study, along with 95% confidence intervals (Bang et al., 2004). For meta-analysis, an inverse-variance weighted average was calculated for both eBI and cBI. Individual BI values were then divided into two separate groups for comparison of pharmacologic and non-pharmacologic interventions. In order to use a single value for a study's overall EOB, eBI+cBI was also calculated as eBI+cBI≈0 can capture two relatively common ideal blinding scenarios where eBI≈0 and cBI≈0 (i.e., random guess in both arms) or eBI≈ cBI⩾0 (i.e., participants in both groups tend to guess they received active treatment). In congruence with parameters set for individual BI values, an eBI+cBI value of Z0.4 and r 0.4 was considered to have blinding that may call for some attention or action (e.g., secondary analysis accounting for blinding issues, providing lessons or caveats for future trials).

In addition, BI values were correlated with effect size, where effect sizes were measured by Cohen's d (Hedges and Olkin, 1985). Effect size values were those that were used in the meta-analysis by Baethge et al. (2013).

3. Results

3.1. Data search

Based on the original search criteria from Baethge et al. 3732 studies were found on the initial search, but 1265 were excluded based on search criteria. Of the remaining 2467 studies that underwent full-text search, 61 were reported to contain information on blinding

assessment. Five additional studies were found by searching the reference lists of the 61 articles retrieved (Baethge et al., 2013). Of these 66 studies, 37 reported data on subjects' guesses as to which treatment they had received. After contacting the remaining primary authors twice, seven authors responded; two submitted the necessary data to include their studies.

One study (Holroyd, 2006) reported data for two distinct experimental and control arms with independent EOB results and was included as two separate studies. Four studies were found to have multiple experimental treatment arms with a single control arm and were included as a single study combining experimental treatment arms (Fitzgerald, 2003; Hall, 2002; Morely, 2006; Narushima, 2002). A total of 40 studies with 1907 patients were included in this meta-analysis (Fig. 1).

3.2. Blinding index calculations

BI values (point and interval estimates), including eBI and cBI, and effect size values computed from all 40 studies are presented in Table 2. The trends of BI values for experimental and control groups are plotted in Fig. 2. The 40 experimental treatment groups had an (inverse-variance weighted) average of 0.14, while the 40 control groups had an average of 0.00. While experimental treatment groups tend to have random guessing, control groups more often have random guessing. Comparisons of eBI with effect size (measured by Cohen's d) and eBI+cBI with effect size are visualized in Fig. 3.

Using the nine blinding scenarios based on the BI as shown in Table 1,0 27.5% (11/40) of studies were well blinded. On the other hand, including blinding scenarios and eBI + cBI values, 22.5% (9/40) possibly had problematic blinding. The remaining 50% (20/40) are included in less certain scenarios that cannot be considered either well or poorly blinded; however, they can still be interpreted for better understanding by the blinding scenarios. The largest groups in this ‘gray’ area are random-unblinded (15%), unblinded-opposite (15%), and unblinded-random (17.5%) for experimental-control.

We repeated the analyses for pharmacologic vs. nonpharmacologic interventions. The pharmacologic experimental treatment groups had an average BI of 0.18, while the nonpharmacologic experimental treatment groups had an average BI of 0.00. In contrast, the pharmacologic control treatment groups had an average BI of −0.02, while the non-pharmacologic control treatment groups had an average BI of 0.05. When using the nine blinding scenarios to evaluate pharmacologic and nonpharmacologic studies, 28% (9/32) of pharmacologic studies had acceptable blinding. Similarly, 25% (2/8) of non-pharmacologic studies had acceptable blinding.

4. Discussion

After the analysis of the 40 included studies with analyzable data, the overall blinding scenario/pattern was random in the experimental treatment group and random in the control group. This leads us to believe that overall, published articles on major psychiatric disorders from 2000 to 2010, which reported on blinding assessment for patients, were effectively blinded. Further supporting this idea, the most common blinding scenario was also random–

random (ideal from the scientific perspective), comprising of 27.5% of all studies included, as compared to 22.5% of studies having possibly problematic blinding. These figures support the finding of successful blinding in this group of studies as stated in Baethge et al.'s (2013) initial report. In comparison, a recent study evaluating the EOB using the BI in acupuncture studies found the overall blinding scenario as well as most common scenario to be unblinded-opposite, e.g., most participants thinking they were receiving experimental treatment regardless of their actual group (Moroz et al., 2013). Trial participants in psychiatric studies may have a different pattern of expectation. An average cBI of 0.00 lends support to this assumption.

Three blinding scenarios that cannot be considered either well or poorly blinded contain the majority of studies comprising this group. The first being random in the experimental arm and unblinded in the control arm. An interpretation could be that there was little treatment effect to influence guessing in the experimental arm and there was an inefficient control, also with no treatment effect, influencing guesses. The second scenario is more correct guesses in the experimental arm and more incorrect guesses in the control arm. It is possible that people have ‘wishful thinking’, hoping/believing the treatment they received is real treatment. This could also be seen as a strong placebo effect. In the third scenario the experimental arm is unblinded and the control arm is random. A possible explanation is that in the presence of treatment effect the experimental group becomes unblinded, while the control is successful in maintaining random guesses.

Of the 9 studies included in blinding scenarios that possibly have problematic blinding, 4 of them have eBI + cBI values that are ≤ 0.4. Two of those studies have cBIs that are very large negative values bringing the eBI+cBI into the possibly problematic scenario (Fitzgerald, 2005; Koran, 2009). It is possible that in these studies their controls actually have large treatment effects and were not inert controls. In the other two studies both cBIs and eBIs were large negative values (Hypericum Depression Trial Study Group, 2002; Frangou, 2006). This is an interesting scenario since patients in both arms appear to be convinced they are in the opposite group, which is expected to be rare in practice.

The overall average for control arms showed excellent blinding results with a perfect BI of 0.00. The experimental arm average was still within a reasonable 0.14 of a perfect BI value. When effect size is compared to eBI, we observed a positive correlation of $r = 0.28$ between the two (Fig. 3a). It is possible that effective treatment leads to increased unblinding, which is natural or even ideal in some settings. There are other possibilities for increased unblinding including side effects of treatment and accidental unmasking, which could then lead to a falsely increased effect size so that bi-directionality is possible. This is also seen when effect size is compared to the overall combined BI value of eBI+cBI. When comparing effect size to eBI + cBI there is a positive correlation (r=0.4), but there is also a separation between two groupings of effect size values. As eBI + cBI values increase from 0.14 to 0.19 there appears to be an increase in the overall trend of effect size (Fig. 3b). One possibility is that as the BI value approaches approximately 0.2 there could be enough unblinding to cross a threshold artificially inflating effect size, or vice versa. Similar findings have been found in other studies showing that even if blinding is attempted but is then unsuccessful, it can

lead to overestimation of treatment effect (Jeong et al., 2013), though it is difficult to know cause from effect in unblinding.

The division into pharmacologic and non-pharmacologic studies further narrows the window of excellently blinded and well-blinded results. The pharmacologic control, non-pharmacologic experimental, and non-pharmacologic control arms were all excellently blinded, having BI absolute values less than or equal to 0.05. It is interesting to note the pharmacologic experimental treatment group with a BI of 0.18. This value is still within a cutoff of 0.2, but is closer to the cutoff. This could be due to greater side effects or noticeable improvements in patients taking pharmacological treatment. We tend to believe, however, that it is easier to blind a standard pharmacologic trial than a study involving psychotherapy. A patient usually knows whether or not he/she sat down with a therapist and may also be aware of the principles of the psychotherapy under study. The other type of nonpharmacologic intervention included in the studies used was transcranial magnetic stimulation. This intervention uses sham controls, which has the capability for successful blinding. The results of this current meta-analysis show the potential for successful blinding of non-pharmacologic interventions, but validation and elucidation are warranted.

Although we use the conventionally accepted term, unblinding, it has to be acknowledged that it is impossible to estimate blinding or unblinding. One can only know if the answered guess is correct or not (Roy, 2012). While this review is more quantitative and systematic in evaluating blinding methods for major psychiatric studies than in the past, it too has its limitations. With finite resources, Baethge et al. limited their literature search to Medline as the single most important literature database in Medicine, particularly with regard to pharmacopsychiatry. While it is likely that expanding the search to EMBASE or PsycInfo would have resulted in a limited number of additional RCTs. Also, it is conceivable that publication bias or selective reporting plays a role in explaining the results (Phillips, 2004). Our sample size of 40 studies only represents 1.6% of the studies meeting our initial search criteria. It is possible that the other 98.4% either chose not to evaluate blinding, selectively withheld blinding assessment data, or did not know/understand an accepted analytic method to evaluate and interpret blinding. Further valuable and representative data may have been found if blinding assessment was more widely performed (Rossner et al., 2007). As this review shows a feasible method for blinding evaluation, future studies in psychiatric research may follow this blueprint on an individual basis for the evaluation of blinding. Reporting of honest and unbiased data is strongly called for. Even in our report of voluntarily reporting studies more than one fifth showed some indication of possibly problematic blinding. There was also a small and unbalanced sample size in subgroup analyses (e.g., when comparing pharmacologic and non-pharmacologic treatments).

This systematic review's approach is intended to demonstrate the feasibility of BI evaluation of psychiatric RCTs and meant as an *empirical* effort for the evaluation of blinding, where blinding data are not commonly reported/available but some did provide sufficient data (so that we can format data in standardized fashion), which allow statistical analyses. This review could strengthen the blinding analysis in psychiatric RCTs and provide new insights on blinding patterns.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Jongbae J. Park was supported by the National Institute of Dental and Craniofacial Research of the National Institute of Health under Award no. K12DE022793.

References

Baethge C, Assall O, Baldessarini R. Systematic review of blinding assessment in randomized controlled trials in schizophrenia and affective disorders 2000–2010. Psychotherapy and Psychosomatics. 2013; 82(3):152–160. [PubMed: 23548796]

Bang H, Flaherty S, Kolahi J, Park J. 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.

Bang H, Ni L, Davis C. Assessment of blinding in clinical trials. Controlled Clinical Trials. 2004; 25(2):143–156. [PubMed: 15020033]

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 Anesthesiology. 2012; 12:13. [PubMed: 22762351]

Boutron I, Estellat C, Ravaud P. Review of blinding in randomized controlled trials found results inconsistent and questionable. Journal of Clinical Epidemiology. 2005; 58(12):1220–1226. [PubMed: 16291465]

Fergusson D, Cranley Glass K, Waring D, Shapiro S. Turning a blind eye: the success of blinding reported in a random sample of randomized, placebo controlled trials. British Medical Journal. 2004; 328:432. [PubMed: 14761905]

Ghimire S, Kyung E, Kang W, Kim E. Assessment of adherence to the CONSORT statement for quality of reports on randomized controlled trial abstracts from four high-impact general medical journals. Trials. 2012; 13:77. [PubMed: 22676267]

Greenfield M, Mhyre J, Mashour G, Blum J, Yen E, Rosenberg A. Improvement in the quality of randomized controlled trials among general anesthesiology journals 2000 to 2006: a 6-year follow-up. Anesthesia and Analgesia. 2009; 108(6):1916–1921. [PubMed: 19448222]

Hedges, L.; Olkin, I. Statistical Methods for Meta-Analysis. Academic Press; Orlando: 1985.

Hróbjartsson A, Forfang E, Haahr M, Als-Nielsen B, Brorson S. 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(3):654–663. [PubMed: 17440024]

James KE, Bloch DA, Lee KK, Kraemer HC, Fuller RK. 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]

Jeong H, Yim HW, Cho Y, Park HJ, Jeong S, Kim H, et al. The effect of rigorous study design in the research of autologous bone marrow-derived mononuclear cell transfer in patients with acute myocardial infarction. Stem Cell Research & Therapy. 2013; 4:82. [PubMed: 23849537]

Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. Annals of Internal Medicine. 2001; 134:657–662. [PubMed: 11304106]

Moroz A, Freed B, Tiedemann L, Bang H, Howell M, Park JJ. Blinding measured: a systematic review of randomized controlled trials of acupuncture. Evidence-Based Complementary and Alternative Medicine. 2013; 12

Park J, Bang H, Cañette I. Blinding in clinical trials, time to do it better. Complementary Therapies in Medicine. 2008; 16(3):121–123. [PubMed: 18534323]

Péron J, Pond G, Gan H, et al. Quality of reporting of modern randomized controlled trials in medical oncology: a systematic review. Journal of the National Cancer Institute. 2012; 104(13):982–989. [PubMed: 22761273]

Phillips C. Publication bias in situ. BMC Medical Research Methodology. 2004; 4:20. [PubMed: 15296515]

Reveiz L, Chan A, Krleza-Jerić K, et al. Reporting of methodologic information on trial registries for quality assessment: a study of trial records retrieved from the WHO search portal. PLoS One. 2010; 5:8.

Rios L, Odueyungbo A, Moitri M, Rahman M, Thabane L. Quality of reporting of randomized controlled trials in general endocrinology literature. Journal of Clinical Endocrinology and Metabolism. 2008; 93(10):3810–3816. [PubMed: 18583463]

Rossner M, Van Epps H, Hill E. Show me the data. Journal of Cell Biology. 2007; 179(6):1091–1092. [PubMed: 18086910]

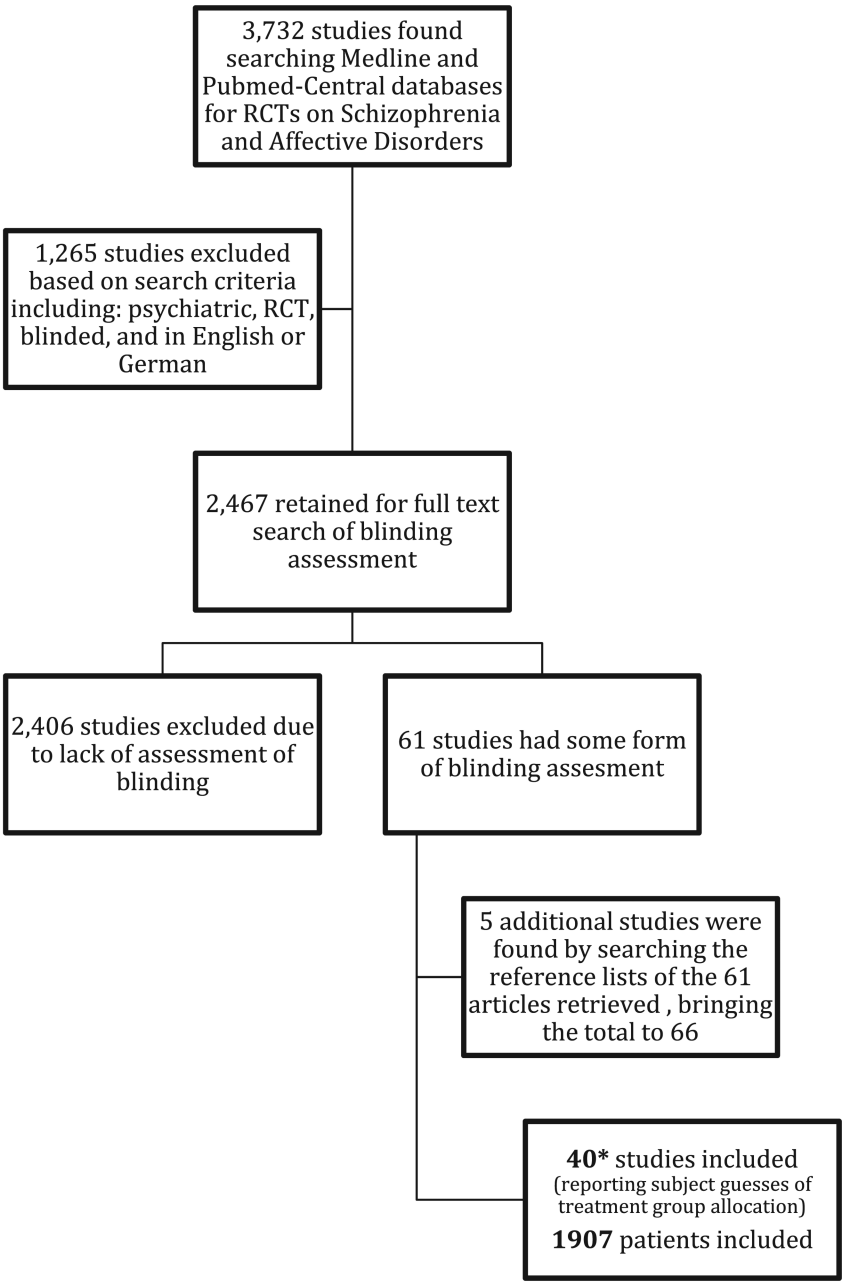
Roy J. Randomized treatment-belief trials. Contemporary Clinical Trials. 2012; 33(1):172–177. [PubMed: 21989161]

Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Annals of Internal Medicine. 2010; 152:726–732. [PubMed: 20335313]

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]

Turner L, Shamseer L, Altman D, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. Cochrane Database of Systematic Reviews. 2012; 11

Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. British Medical Journal. 2008; 336:601–605. [PubMed: 18316340]



*One article contained two separate studies.

Fig. 1.
Systematic review search and selection.

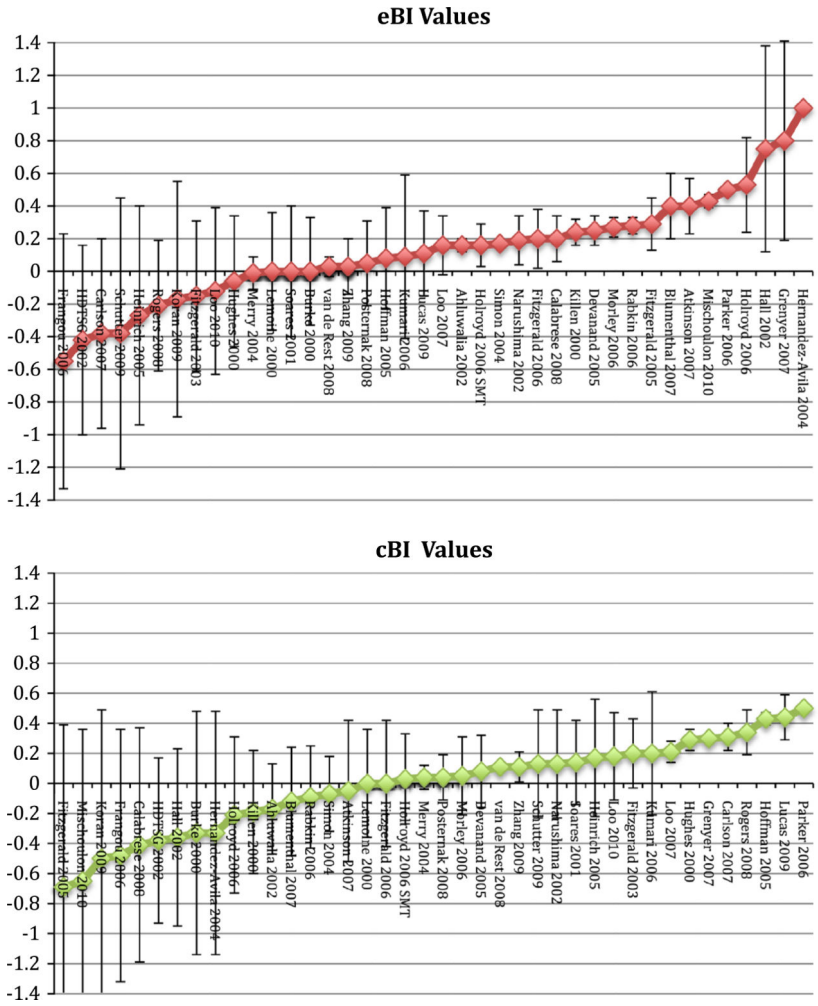


Fig. 2.

Experimental (a:upper) and Control (b:lower) BI values with confidence intervals. Confidence intervals (CIs) are 95% CIs, unadjusted for multiple testing. Those with missing error bars have values that exceed those shown in the scale of these figures.

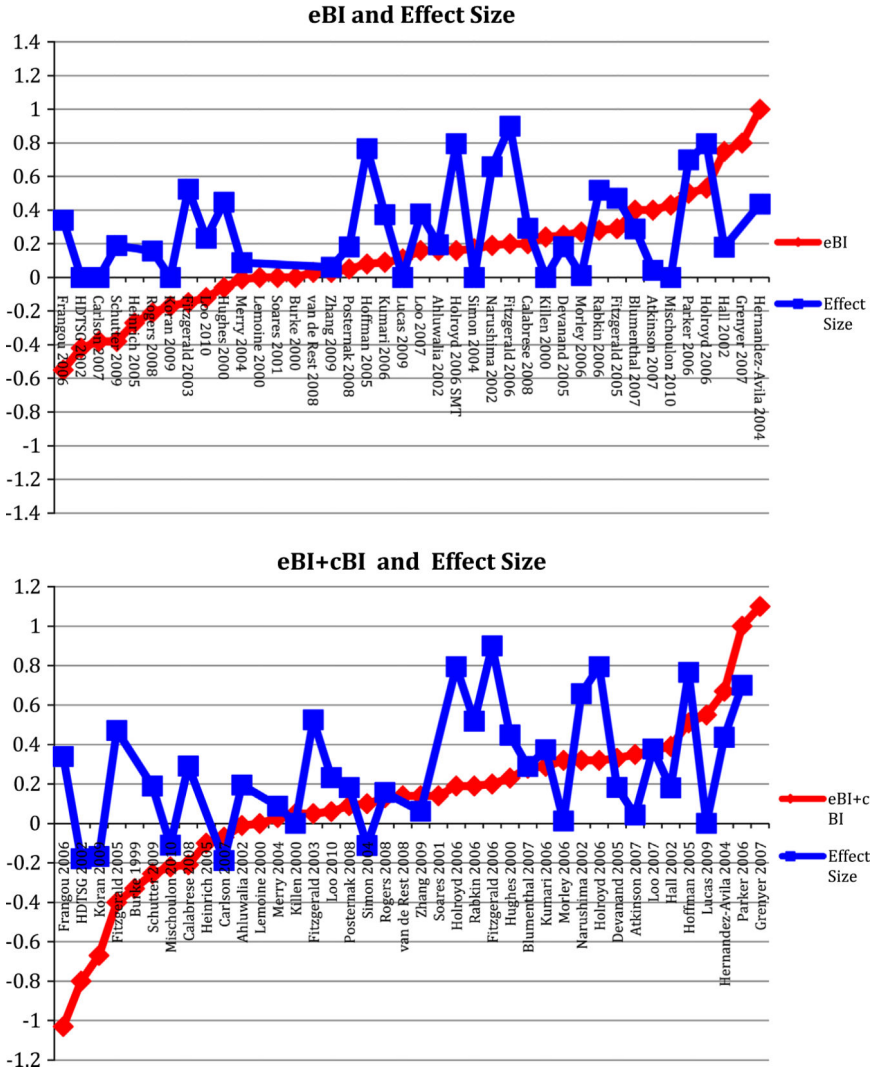


Fig. 3.

Experimental BI values (a:upper) and the sum of Experimental and Control BI values (b:lower) overlaid with Effect sizes. Higher eBI+cBI values indicate more excessive proportion of correct guesses while controlling for chance and ‘wishful thinking’. Effect Sizes are measured by Cohen's d.

Table 1

Blinding Scenarios (eBI, cBI) in 40 trials.

Experimental arm (Verum)	Control arm (Sham)	Possible blinding and clinical effectiveness interpretations	Trials number (%)
Random guess	Random guess	Ideal	11 (27.5)
Random guess	Opposite guess	Rare	2 (5)
Random guess	Unblinded	Possibly little treatment effect and completely no effect in control arm	6 (15)
Unblinded	Unblinded	Possibly problematic, strong treatment effect in experimental arm and no treatment effect in control arm (e.g. patients tend to know what to expect)	2 (5)
Unblinded	Opposite guess	Possible that patients tend to have wishful thinking, weak treatment and strong placebo effect, or any treatment administered is perceived as real treatment	6 (15)
Unblinded	Random guess	Possible treatment effect in experimental arm and no treatment effect in control arm (e.g. patients do not know what to expect in the absence of treatment)	7 (17.5)
Opposite guess	Opposite guess	Rare	2 (5)
Opposite guess	Random guess	Rare	2 (5)
Opposite guess	Unblinded	No treatment effect at all, or patients may have low expectations	2 (5)

These values are rules of thumb (based on 0.2 cutoff) rather than definite cutoffs, and used here only for exploratory classification purposes (Bang et al., 2010).

Table 2

BI values and effect sizes.

	eBI [CI 95%]	cBI [CI 95%]	Effect size	Exper. type
Frangou (2006)	–0.55 [–1.33 to 0.23]	–0.48 [–1.32 to 0.36]	0.34	P
HDTSG (2002)	–0.42 [–1.0 to 0.16]	–0.38 [–0.93 to 0.17]	–0.18	P
Schutter (2009)	–0.38 [–1.21 to 0.45]	0.13 [–0.23 to 0.49]	0.19	NP
Heinrich (2005)	–0.27 [–0.94 to 0.4]	0.17 [–0.22 to 0.56]	N.A.	P
Carlson (2007)	–0.38[–0.96 to 0.2]	0.31 [0.22–0.4]	–0.19	P
Rogers (2008)	–0.21 [–0.61 to 0.19]	0.34 [0.19–0.49]	0.16	P
Koran (2009)	–0.17 [–0.89 to 0.55]	–0.5 [–1.49 to 0.49]	–0.17	P
Burke (2000)	0[–0.33 to 0.33]	–0.33 [–1.14 to 0.48]	N.A.	P
Ahluwalia (2002)	0.16 [0.12–0.2]	–0.17 [–0.48 to 0.12]	0.20	P
Simon (2004)	0.17 [0.17–0.17]	–0.07 [–0.32 to 0.18]	–0.11	P
Lemoine (2000)	0[–0.36 to 0.36]	0 [–0.36 to 0.36]	N.A.	P
Holroyd (2006)	0.16 [0.03–0.29]	0.03 [–0.27 to 0.33]	0.80	P
Merry (2004)	–0.01 [–0.11 to 0.09]	0.04 [–0.04 to 0.12]	0.09	NP
Posternak (2008)	0.05 [–0.21 to 0.31]	0.04 [–0.11 to 0.19]	0.18	P
van de Rest (2008)	0.03 [–0.03 to 0.09]	0.11 [0.1–0.12]	N.A.	P
Zhang (2009)	0.03 [–0.14 to 0.2]	0.11 [0.01–0.21]	0.06	P
Narushima (2002)	0.19 [0.04–0.34]	0.13 [–0.23 to 0.49]	0.66	P
Soares (2001)	0[–0.4 to 0.4]	0.14 [–0.14 to 0.42]	N.A.	P
Loo (2010)	–0.12 [–0.63 to 0.39]	0.18 [–0.11 to 0.47]	0.23	NP
Fitzgerald (2003)	–0.15 [–0.61 to 0.31]	0.2 [–0.03 to 0.43]	0.53	NP
Kumari (2006)	0.09 [–0.41 to 0.59]	0.2 [–0.21 to 0.61]	0.37	P
Loo (2007)	0.16 [–0.02 to 0.34]	0.21 [0.14–0.28]	0.38	NP
Hughes (2000)	–0.06 [0.46 to 0.34]	0.29 [0.22–0.36]	0.45	P
Hoffman (2005)	0.08 [–0.23 to 0.39]	0.43 [0.39–0.47]	0.77	NP
Lucas (2009)	0.11 [–0.15 to 0.37]	0.44 [0.29–0.59]	0.00	P
Fitzgerald (2005)	0.29 [0.13–0.45]	–0.69 [–1.77 to 0.39]	0.47	NP
Mischoulon (2010)	0.43 [0.39–0.47]	–0.65 [–1.66 to 0.36]	–0.11	P
Calabrese (2008)	0.25 [0.11–0.39]	–0.41 [–1.19 to 0.37]	0.29	P
Hall (2002)	0.75 [0.12–1.38]	–0.36 [–0.95 to 0.23]	0.18	P
Hernandez-Avila (2004)	1[1.0–1.0]	–0.33 [–1.14 to 0.48]	0.44	P
Holroyd (2006)	0.53 [0.24–0.82]	–0.21 [–0.73 to 0.31]	0.80	P
Killen (2000)	0.24 [0.16–0.32]	–0.19 [–0.6 to 0.22]	0.00	P
Blumenthal (2007)	0.4 [0.2–0.6]	–0.12 [–0.48 to 0.24]	0.29	P
Rabkin (2006)	0.28 [0.23–0.33]	–0.09 [–0.43 to 0.25]	0.52	P
Atkinson (2007)	0.4 [0.23–0.57]	–0.05 [–0.52 to 0.42]	0.04	P
Fitzgerald (2006)	0.2 [0.02–0.38]	0 [–0.42 to 0.42]	0.9	NP
Morley (2006)	0.27 [0.21–0.33]	0.05 [–0.21 to 0.31]	0.01	P
Devanand (2005)	0.25 [0.16–0.34]	0.08 [–0.16 to 0.32]	0.18	P
Grenyer (2007)	0.8 [0.19–1.41]	0.3 [0.28–0.32]	N.A.	P

	eBI [CI 95%]	cBI [CI 95%]	Effect size	Exper. type
Parker (2006)	0.5 [0.5–0.5]	0.5 [0.5–0.5]	0.7	P

Experimental arm BI value (eBI), Control arm BI value (cBI), Pharmacologic (P), Non-Pharmacologic (NP), Hypericum Depression Trial Study Group (HDTSG). BI and effect sizewere measured for experimental and control arms of HDTSG not including the active comparator. Effect Size measured by Cohen's d. N.A.: Not available because data in original papers were not sufficient to calculate effect size using standard methods. See Supplementary table for individual study characteristics.