CLINICAL TRIALS ARTICLE Clinical Trials 2007; **4**: 245–253

An exploratory test for an excess of significant findings

John PA Ioannidis[a, b] *and Thomas A Trikalinos*[a, b]

Background The published clinical research literature may be distorted by the pursuit of statistically significant results.
Purpose We aimed to develop a test to explore biases stemming from the pursuit of nominal statistical significance.
Methods The exploratory test evaluates whether there is a relative excess of formally significant findings in the published literature due to any reason (e.g., publication bias, selective analyses and outcome reporting, or fabricated data). The number of expected studies with statistically significant results is estimated and compared against the number of observed significant studies. The main application uses $\alpha = 0.05$, but a range of α thresholds is also examined. Different values or prior distributions of the effect size are assumed. Given the typically low power (few studies per research question), the test may be best applied across domains of many meta-analyses that share common characteristics (interventions, outcomes, study populations, research environment).
Results We evaluated illustratively eight meta-analyses of clinical trials with >50 studies each and 10 meta-analyses of clinical efficacy for neuroleptic agents in schizophrenia; the 10 meta-analyses were also examined as a composite domain. Different results were obtained against commonly used tests of publication bias. We demonstrated a clear or possible excess of significant studies in 6 of 8 large meta-analyses and in the wide domain of neuroleptic treatments.
Limitations The proposed test is exploratory, may depend on prior assumptions, and should be applied cautiously.
Conclusions An excess of significant findings may be documented in some clinical research fields. *Clinical Trials* 2007; **4**: 245–253; http://ctj.sagepub.com

Introduction

Research findings are sometimes considered more or less attractive based on their statistical significance [1–5]. Statistical significance is widely employed in hypothesis-testing. However, undue emphasis on significance levels may bias the accumulated scientific evidence [5]. The pursuit of statistical significance may manifest in various ways. Typical examples of such biases include publication bias, selective analysis and outcome bias, and fabrication bias.

In publication bias [6–10], the chances of publication of a study depend on its results, that is, studies with ‘positive’ (statistically significant) results have higher chances to be published than studies with ‘negative’ results. The latter may never be published or may be published with delay (time lag bias) [11]. We use here the terms ‘positive’ and ‘negative’ only for convenience without any connotation for study quality. Several tests have been developed to probe publication bias. Some tests examine whether results of large studies differ from results of smaller ones [12–16]. Another family of tests models with weight functions the selection process, that is, how the probability of publication may vary for different *P*-values [17–21]. However, all tests have limitations. Many sources of heterogeneity

[a]Clinical Trials and Evidence Based Medicine Unit and Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, [b]Institute for Clinical Research and Health Policy Studies, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, USA
Address for correspondence: John PA Ioannidis, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece. E-mail: jioannid@cc.uoi.gr

10.1177/1740774507079441

or other biases make large studies differ from smaller ones [22]. Weight function models are computationally complex and selection processes typically unknown. Importantly, no test is validated against real-life data where we knew that publication bias was or not present.

Selective analyses and selective outcome reporting may also result in spurious significant results being presented in the literature. This may be a major problem in medical research [23–25]. Data dredging is a common practice in epidemiology [26] and may be even more prevalent in modern discovery-oriented research [5]. Some 'positive' results may even represent fake data. Fabrication has been documented through real-life examples with major clinical impact [27]. The prevalence of fabrication bias is also supported by large-scale surveys of clinical researchers and biostatisticians [28,29] and indirect statistical testing of trial reports [30].

All of these biases eventually result in a relative excess of published statistically significant results as compared with what their true proportion should be in a body of evidence. This is impossible to detect in a single study for an outside observer who does not know what has happened to its design, database and analysis. However, one may examine a large body of studies on the same and similar questions and gain insights on the presence of an excess of statistically significant findings by examining these studies together.

Exploratory test

Main concept

We test in a body of n published studies whether the observed number of studies O with 'positive' results at a specified α level on a specific research question is different from the expected number of studies with 'positive' results E in the absence of any bias.

Suppose there is a true effect size that is being pursued by study i ($i = 1 \ldots n$) and its size is θ_i. In frequentist terms, we accept the alternative hypothesis for an effect θ_i. Then, the expected probability that a specific single study i will find a 'positive' result equals $1 - \beta_i$, its power at the specified α level. Power depends on θ_i. Assuming no bias, E equals the sum of the expected probabilities across all studies on the same question:

$$E = \sum_{i=1}^{n} (1 - \beta_i).$$

Effect size

If it can be safely assumed that the effect is the same in all studies on the same question, for example, all studies entered in a meta-analysis, an effect size to consider is the summary effect of all studies combined according to a standard meta-analysis procedure. Either fixed or random effects summary estimates may then be used [31], because they agree [32] in the absence of notable heterogeneity.

In the presence of considerable between-study heterogeneity, efforts should be made first to dissect sources of heterogeneity [33,34]. Applying the test ignoring genuine heterogeneity is ill-advised. If genuine heterogeneity can be properly dissected, then the test may still be applied considering different effects for each study or sub-sets of studies. However, between-study heterogeneity does not necessarily represent genuine diversity, but may reflect different amounts of bias among studies that otherwise should have had similar results.

Inferences

The expected number E is compared against the observed number O of 'positive' studies using the χ^2 statistic $\mathrm{A} = [(O-E)^2/\mathrm{E} + (O-E)^2/(n-E)] \sim \chi^2_1$. Alternatively, one may use a binomial probability test (preferable with small numbers).

The power to detect a specific excess, for a given O among n studies is easily derived as the power of a χ^2 test or a binomial probability test. Given that power will be low, especially with few 'positive' studies, we choose two-tailed $P < 0.10$, as in proposed publication bias tests [13,16]. One-tailed approaches [35] can be equally straightforward. The test can be applied regardless of whether the study outcome of interest is binary or continuous. Here, we focus on binary outcomes for parsimony and consistency.

Consideration of different effect sizes

The assumption that the observed meta-analysis summary effect size reflects the true effect may certainly be spurious. Most biases that increase the proportion of 'positive' results may also inflate the observed summary effect size, that is, E will be overestimated.

To address this issue, one may consider alternative effect sizes or a range of effect sizes. Alternative effect sizes may be derived from other external evidence [36,37] or topic-specific considerations. One may also routinely examine the 95% confidence interval (CI) of the observed summary effect as a range.

Let θ_{lim} be the effect size above which the excess of observed over-expected 'positive' studies becomes formally statistically significant ($P = 0.10$). If other (external) evidence is available for considering a range of effect sizes, θ_{lim} may be compared against this range. If θ_{lim} is within that range, significance-chasing bias is still possible.

Consideration of different levels of statistical significance

Usually the chase for statistical significance would entail the $P = 0.05$ threshold [1–4]. One may also calculate E and count O using different thresholds or a range thereof. The statistical significance of the difference $O - E$ can be plotted as a function of α. If significance-chasing is strongly guided by $\alpha = 0.05$ threshold, one expects to see the difference being significant in the vicinity of $\alpha = 0.05$, but not necessarily for smaller or larger α values: investigators would aim to pass the 0.05 threshold, but not necessarily lower thresholds (e.g., $P = 0.01$ or 0.001). Similarly, there is no strong reason to seek a P-value of say 0.12 rather than 0.20. However, this behaviour may sometimes not be seen despite strong pursuit of statistical significance. Possible explanations appear in Table 1.

Extension – domains

Bias in the pursuit of statistical significance may operate across studies on different research questions within the same domain. We define *domain* here as a field of research with similar, but not necessarily identical, research questions. There is an increasing interest in appraising biases in large-scale across many topics [38,39]. Domains may be defined according to common general theme, common type of interventions, common type of subjects, common methodology, common research environments, common language of publication or combinations of these factors. Defining the boundaries of a domain are a process similar to defining the boundaries of eligibility for studies to be included in a meta-analysis [40], but here the scale is one step larger.

Table 1 Potential explanations when bias seems to entail thresholds other than P =0.05

The 0.05 threshold may not be particularly more attractive than other thresholds
Investigators may reach close to 0.05 (say P = 5 0.06 or 0.08) and verbally make the leap that their results are ‘significant’
Investigators may try to bypass also multiple comparison corrections (e.g., Bonferroni) not accounted for in a meta-analysis
Bias may inadvertently drop the P-values to much lower levels than the desired 0.05 (‘over-kill’)
Meta-analysis may use standardized analyses [25,51] different from those selected by the original publication
Chance
Combinations of these reasons

In the domain setting, one may calculate the number of expected statistically significant results across all studies i ($i = 1,\ldots, n$) of all meta- analyses j ($j = 1,\ldots, k$) considered in the domain. Inferences would then be based on the statistic $\mathrm{A} = [(O-E)^2/\mathrm{E} + (O-E)^2/(n-E)] \sim \chi^2_1$, where.

$$E = \sum\nolimits_{i=1,j=1}^{n,k}(1-\beta_{i,j})$$

The binomial probability test may also be applied. The power of each study is estimated based on a plausible effect size, as above.

Consideration of a prior distribution for the effect size

Another approach is to consider a distribution for the prior probability for the effect θ_i with probability density function $f_i(\theta)$. There is also a probability, $u_{\text{null},i}$, that there is no effect at all. The probability of finding a ‘positive’ result in a study is equal to the probability of a finding a true positive result plus the probability of finding a false positive result [5,36,41,42]:

$$E = \sum_{i=1}^{n}\left(\int_{\theta\neq 0}((1-\beta_i(\theta))f_i(\theta)\,\mathrm{d}\theta) + au_{\text{null},i}\right).$$

One can use previous data to obtain an empirical prior. This may also be altered to account for bias (decreased mean effect) or different dispersion (larger variance) [36,43–45].

Application

We illustrate empirically this exploratory test in different meta-analyses and in a whole domain. First, we examine large meta-analyses with over 50 included studies each where low power would not be an issue. Secondly, we evaluate a sample of meta-analyses with numbers of studies that are in the typical range seen in most research fields. These meta-analyses all pertain to a common group of interventions, the same type of outcome and the same disease; a domain analysis is also presented using these data. Analyses were performed in Intercooled Stata 8.2 (Stata Corp., College Station, Texas). Asymptotic power calculations used the sampsi command in Stata; simulation based power calculations were programmed as described in the Appendix

Comparison against tests for publication bias

We compared the results of the exploratory test against traditional ‘publication bias’ tests: the nonparametric (tau) correlation of the effect size (natural logarithm of the odds ratio (OR)) against its variance (Begg-Mazumdar test) [13], the regression of standardized effect size against the inverse of the standard error [14] and a typical selection model test based on weight functions that links the probability that a study will be published with the study's effect size only through it's *P*-value (assuming a two-step selection process with *P*-value cutoffs 0.025 or 0.975) [20].

Meta-analyses with many studies

We perused the Cochrane Library, issue 3, 2003 to identify all meta-analyses with at least 50 included studies and binary outcomes. Studies with zero event counts on both arms were excluded from all calculations. For meta-analyses that used similar or overlapping studies (e.g., same comparison, different outcomes), we retained the one with largest number of studies. Whenever several meta-analyses on the same review had the same number of studies, we selected the one with largest number of events.

Eight meta-analyses qualified (732 studies, range 55–155 per meta-analysis) (Table 2). The proportion of observed 'positive' studies ranged from 5 to 55% across meta-analyses. The proportion of expected 'positive' studies ranged from 3 to 46%. The observed number of 'positive' studies always exceeded the expected, except for tamoxifen in early breast cancer. This is the only meta-analysis based on individual-level data and performed under the guidance of a central secretariat. Hence, it is expected that publication and reporting biases would be minimized. All other meta-analyses used published reported information. The ratio of observed over expected 'positive' studies in these seven meta-analyses ranged from 1.13 to 1.80.

In three meta-analyses, the difference between observed and expected was beyond chance ($P < 0.10$) with an excess of O over E. Our test suggested bias in one meta-analysis where neither the correlation nor the regression test gave a signal; the reverse situation was seen in another case (Table 2). The selection model gave a signal only in one of the three meta-analyses with identified bias in our test.

Calculation of $\theta_{\lim}$

$\theta_{\lim}$ was calculated iteratively, using Newton's algorithm. Starting from a value of θ away from the observed, we estimated whether E was significantly different from O. Sequential θ values were derived with linear interpolation and were assessed iteratively until the P-value for bias reached 0.10 ± 0.005.

Besides the three meta-analyses with clear signals, in another three meta-analyses $\theta_{\lim}$ was very close to the observed θ. Bias would be inferred,

Table 2 Meta-analyses including over 50 studies

					P-values for various tests of bias				
					Current test		Publication bias[b]		
Topic (outcome)	Studies (pts)	OR (95% CI)[a]	*O*	*E*	χ^2	Binomial	tau-b	Regression	Selection
Nicotine replacement (not smoking at 6–12 months)	98 (37760)	1.79 (1.65–1.93)	37	32.6	0.34	0.34	0.01	0.001	0.22
Tamoxifen for early breast cancer (recurrence)	55 (37099)	0.71 (0.67–0.76)	20	22.2	0.54	0.58	0.76	0.64	0.29
Antibiotic prophylaxis for C-section (endometritis)	80 (11715)	0.30 (0.26–0.34)	40	35.0	0.26	0.26	0.16	0.10	nc
Antibiotics for appendicectomy (wound infection)	70 (8808)	0.34 (0.29–0.41)	27	18.6	0.02	0.03	<0.001	<0.001	0.11
Aprotinin (allogeneic blood transfusion)	75 (7680)	0.37 (0.31–0.44)	36[c]	27.4	0.04	0.04	0.012	0.10	0.08
Single-dose aspirin for acute pain (relief > 50%)	64 (4922)	3.75 (3.16–4.44)	35	29.6	0.17	0.12	<0.001	<0.001	0.38
TCA versus SSRI discontinuation (drop-out)	135 (16151)	1.19 (1.08–1.30)	11	6.1	0.04	0.06	0.18	0.75	0.78
Amitriptyline versus other drug in depression (drop-out)	155 (13871)	0.96 (0.88–1.05)	7[d]	4.1	0.16	0.20	0.36	0.57	0.94

[a] Summary OR is derived by random effects calculations (DerSimonian and Laird model [52]); for all presented examples fixed effects estimates are very similar or even identical; four meta-analyses (nicotine replacement, tamoxifen, aprotinin, TCA versus SSRI) have significant between-study heterogeneity ($P < 0.10$ for the Q statistic), but the amount of heterogeneity is not excessive ($I^2 = 20$–47% in all four cases), while the other four have no significant heterogeneity ($I^2 = 5$–15%). I^2 is usually considered large for values of 75% or higher [53].

[b] tau-b: tau correlation test [13]; regression: weighted regression of standardized effect on inverse of standard error [14]; selection: weight function test [20].

[c] 1 'positive' finding is in the opposite direction compared with the summary effect.

[d] 2 'positive' finding is in the opposite direction compared with the summary effect.

C-section: caesarean section; TCA: tricyclic antidepressants; SSRI: selective serotonin reuptake inhibitors; OR: odds ratio; CI: confidence interval; *O:* observed number of 'positive' studies; *E:* expected number of positive studies (at the $\alpha = 0.05$ level, based on 10000 simulations); nc: no conversion of the model.

if the true OR was 1.76 rather than 1.79 for nicotine replacement, 0.32 rather than 0.30 for antibiotic prophylaxis for cesarean section, 3.64 rather than 3.75 for single-dose aspirin for acute pain. In all these cases, the required summary effect for claiming bias was very close to the observed and well within the 95% CIs of the observed summary effect. Thus, bias is probable also in these meta-analyses. Conversely, in the tamoxifen meta-analysis of individual-level data, θ_{lim} was 0.79, a value well outside the 95% CIs of the observed summary OR (0.71, 95% CI, 0.67–0.76). In the tricyclic antidepressants versus selective serotonin inhibitors meta-analysis, even if there were absolutely no effect (OR = 1.00), still four 'positive' studies would be expected by chance among 135 performed, not significantly different from the seven observed 'positive' studies. Binomial probability and χ^2 tests gave practically identical results.

Consideration of different α levels

Figure 1 shows the statistical significance of the difference between O and E as a function of α. Even assuming that the observed summary OR is not inflated, signals of bias were detectable for six of the eight meta-analyses, with lowest P-values for significance chasing seen for α values close to 0.05 (range 0.03–0.10). In the nicotine replacement meta-analysis, P-values did not drop below 0.10 for any α, but approached 0.10 for $\alpha = 0.03$. In the tamoxifen meta-analysis, there was no such signal in the whole range of α values.

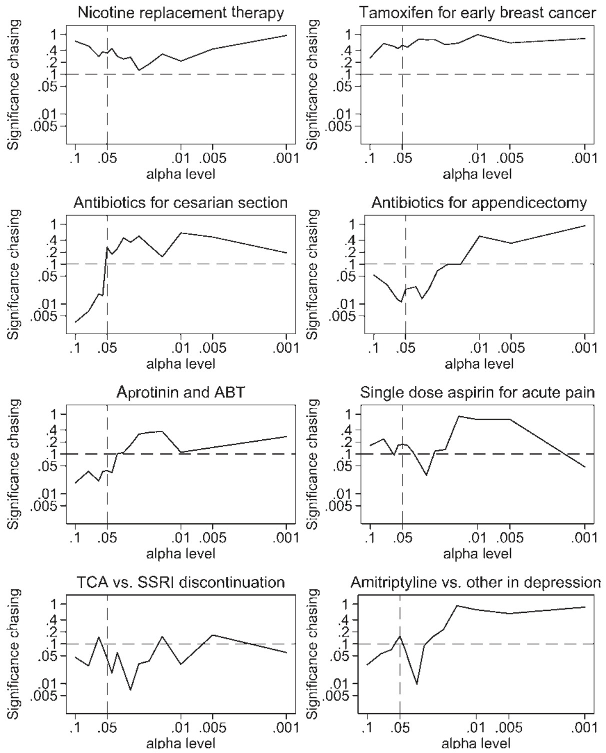

Figure 1 Plots of the *P*-value for the exploratory bias test as a function of the α level at which a study is considered 'positive'. Each plot corresponds to a meta-analysis listed in Table 1

Meta-analyses with typical numbers of studies

We used a database of meta-analyses of mental health-related interventions with binary outcomes [38]. Of those, we selected the 10 meta-analyses that addressed the efficacy of different neuroleptic medications against placebo or against each other in patients with schizophrenia using failure as outcome (Table 3). The summary OR in 5 of the 10 meta-analyses was formally significant while in five meta-analyses no significant difference was found between the compared drugs.

Only one of the 10 meta-analyses had $P < 0.10$ for bias for $\alpha = 0.05$ (there was a significant excess of observed versus expected [10 versus 6.3] 'positive' studies in the meta-analysis of clozapine versus typical neuroleptics, $P = 0.07$). The correlation test showed no significant signal in any of the 10 meta-analyses. The regression test showed a significant signal for the haloperidol versus placebo meta-analysis ($P = 0.006$) and for the thioridazine versus typical neuroleptic meta-analyses ($P = 0.04$). The latter finding is spurious and improbable to represent publication bias, since the meta-analysis shows absolutely no effect (summary OR 0.97) and there is only one 'positive' study. The selection model also found no significant signals and failed to converge (small numbers) in seven cases.

Three meta-analyses had no 'positive' study and evaluation of an excess of significant studies is not meaningful. The other two meta-analyses with non-significant summary OR had only one 'positive' study. Even if the summary OR were 1.00, E would be 0.27 and 0.28, respectively, which is not significantly different from the observed $O = 1$.

The five meta-analyses where one intervention was found to be effective (or more effective than a comparator) had a larger O than E, although the difference was typically small. Besides the clozapine versus typical neuroleptics meta-analysis that had evidence for significance-chasing bias even in the main analysis, θ_{lim} was within the 95% CIs of the observed summary effect in the other four meta-analyses, suggesting that significance-chasing was possible (Table 3).

Analysis of a domain

Across all 10 meta-analyses of neuroleptics, O was 30 and E was 21.5, even when we assumed the

Table 3 Meta-analyses of neuroleptics in schizophrenia

Compared interventions	Studies (pts)	OR (95% CI)[a]	*O*	*E*	Bias[b]
Chlorpromazine versus placebo	24 (1711)	0.31 (0.22–0.45)	10	7.0	Possible
Haloperidol versus placebo	8 (313)	0.19 (0.09–0.40)	3	2.2	Possible
Thioridazine versus placebo	5 (165)	0.19 (0.07–0.51)	2	1.9	Possible
Clozapine versus typical neuroleptics	17 (1850)	0.36 (0.23–0.55)	10	6.3	Yes
Loxapine versus typical neuroleptics	8 (359)	0.82 (0.49–1.38)	0	0.2	No (NM)
New atypical neuroleptics versus clozapine	5 (351)	0.86 (0.56–1.33)	0	0.2	No (NM)
Pimozide versus typical neuroleptics	6 (206)	1.01 (0.46–2.21)	0	0.1	No (NM)
Risperidone versus typical neuroleptics	10 (2981)	0.63 (0.50–0.79)	3	2.9	Possible
Sulpiride versus typical neuroleptics	9 (514)	0.77 (0.49–1.19)	1	0.4	None detectable
Thioridazine versus typical neuroleptics	13 (853)	0.97 (0.61–1.55)	1	0.3	None detectable

The outcome for all topics is treatment failure. OR: odds ratio; CI: confidence interval; *O*: observed number of 'positive' studies; *E*: expected number of positive studies (at the $\alpha = 0.05$ level, based on 10 000 simulations).

[a] Summary odds ratio is derived by random effects calculations; for all presented examples fixed effects estimates are very similar or even identical; two meta-analyses (chlorpromazine versus placebo and clozapine versus typical neuroleptics) have significant between-study heterogeneity ($P < 0.10$ for the Q statistic), but the amount of heterogeneity is not very excessive (I^2 of 33 and 60%, respectively), while the other four have no significant heterogeneity ($I^2 = 0$–34%).

[b] Based on binomial probability test, Yes: $P < 0.10$ in main analysis, assuming the observed summary OR is the plausible effect; Possible: θ_{limit} within 95% CI of the observed summary OR; No: regardless of the assumed plausible OR value, no significance chasing bias is documented; NM: not meaningful to test (no observed 'positive' studies).

summary OR observed in each meta-analysis as the plausible effect size. The observed excess was indicative of bias ($P = 0.041$ by χ^2, $P = 0.052$ by binomial probability test) across the domain.

Figure 2 shows that at the domain level the lowest *P*-values for bias were seen for α values in the vicinity of 0.05 (trough at $\alpha = 0.03$).

Consideration of a probability distribution for the effect size

The number of expected studies with 'positive' results *E* is very similar regardless of whether a distribution of effects instead of a point estimate is considered (Table 4). Results are also very similar, when we consider distributions with doubled variance. Conversely, *E* estimates are much smaller, when we consider distributions with halved mean effects, especially for meta-analyses that have statistically significant treatment effects; then there is typically a very strong excess of *O* over *E*.

Table 4 uses asymptotic power calculations. Results are very similar with simulation-based power calculations. For example, for the meta-analysis of nicotine replacement where $E = 32.6$ using only the point estimate, we get $E = 32.2$ using the distribution of effect sizes, $E = 32.6$ using a distribution with double variance and $E = 11.2$ using a distribution with halved effects (compared with 30.5, 30.5, 30.6 and 10.0, respectively, with asymptotic power calculations). However, using simulations for power calculations along with a probability density distribution for the effect is computationally (time-wise) demanding.

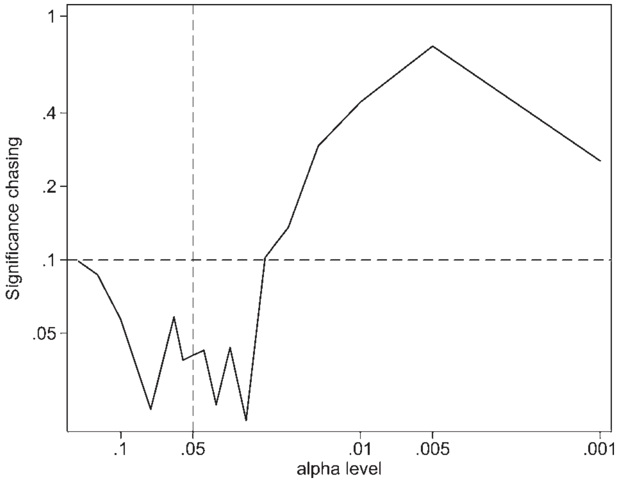

Figure 2 Plot of the *P*-value for the exploratory bias test as a function of the α level at which a study is considered 'positive'. Data have been merged across the 10 meta-analyses of neuroleptics for schizophrenia listed in Table 2

Table 4 Results using different assumptions

Topic (outcome) or comparison	Estimated E using asymptotic power calculations and different assumptions for the effect size: θ	$\sim N(\theta, s(\theta))$	$\sim N(\theta/2, s(\theta))$	$\sim N(\theta, 2 \times s(\theta))$
Meta-analyses including over 50 studies				
Nicotine replacement (not smoking at 6–12 months)	30.5	30.5	10.0	30.6
Tamoxifen for early breast cancer (recurrence)	21.6	21.5	8.4	21.3
Antibiotic prophylaxis for C-section (endometritis)	33.5	33.4	12.7	33.2
Antibiotics for appendicectomy (wound infection)	17.1	17.1	6.6	17.0
Aprotinin (allogeneic blood transfusion)	26.0	25.9	9.3	25.9
Single-dose aspirin for acute pain (relief >50%)	27.6	27.6	7.9	27.5
TCA versus SSRI discontinuation (drop-out)	6.8	7.0	6.3	7.5
Amitriptyline versus other drug in depression (drop-out)	8.5	8.2	8.3	7.9
Meta-analyses of neuroleptics				
Chlorpromazine versus placebo	6.7	6.7	2.5	6.8
Haloperidol versus placebo	2.4	2.5	0.8	2.7
Thioridazine versus placebo	1.7	1.7	0.6	1.8
Clozapine versus typical neuroleptics	5.9	5.8	2.3	5.8
Loxapine versus typical neuroleptics	0.4	0.5	0.5	0.7
New atypical neuroleptics versus clozapine	0.2	0.3	0.3	0.5
Pimozide versus typical neuroleptics	0.4	0.3	0.3	0.6
Risperidone versus typical neuroleptics	2.7	2.6	1.1	2.6
Sulpiride versus typical neuroleptics	0.3	0.5	0.4	0.8
Thioridazine versus typical neuroleptics	0.9	0.8	0.8	1.2

Discussion

We have introduced an exploratory test for examining whether there is an excess of significant findings in a body of evidence. Practical applications suggest a clear or possible excess of ‘positive’ studies in most meta-analyses of randomized controlled trials where many studies exist and similarly in a domain of many meta-analyses. No such excess is seen in a meta-analysis of individual-level data where efforts were made to collect detailed information according to standard rules.

The exploratory test can be run with different methods. The simple method takes the effect θ_i for granted and examines whether the observed number of ‘positive’ studies is compatible with this assumption; different θ_i values can also be probed to derive θ_{lim}. The simplicity of the assumption is both the strength and disadvantage of this approach. Alternatively, one may specify prior probability functions for θ_i and use a more generic Bayesian approach. Results tend to be similar with the two approaches. The second approach is computationally more demanding and there is some unavoidable subjectivity in specifying prior distributions and alternatives.

Some caveats should be discussed. First, we focused on randomized trials. An excess of significant findings may be even more prominent in nonrandomized studies with greater flexibility in the design, data collection, analysis and presentation of the results [25]. In epidemiological studies a common consideration is the use of adjusted analyses. To apply the test on meta-analyses of adjusted estimates, one has to employ power calculations for multivariate models with covariates [46–48]. Applications may be limited by data availability, especially covariate correlation structure in primary studies. However, in fields where unadjusted estimates are appropriate, for example, genetic association studies, the test can be readily applied [49].

Secondly, the extent of bias seen in a meta-analysis may be more limited than what has happened in the primary studies. By definition, a meta-analysis makes efforts to streamline and standardize information and retrieve unpublished data. Meticulous meta-analysis may dissipate the bias of original study reports. Primary studies may have suggested even more significant results. For example, some investigators use asymptotic tests to claim significance for results based on small numbers where exact tests may not show formal statistical significance, or focus on selected ‘positive’ subgroups or adjusted analyses [50], while data that go into meta-analysis calculations are ‘negative’.

Thirdly, the proposed test has very low power when there are very few ‘positive’ studies and is meaningless when there are no ‘positive’ studies. This is common in meta-analyses. With few ‘positive’ studies, only θ_{lim} inferences may be meaningful.

Fourth, caution is warranted when there is genuine between-study heterogeneity. Tests of publication bias [33] generally yield spurious results in this setting. Genuine heterogeneity may be mistaken for bias. If there is strong evidence that different studies should be considered in separate subgroups, then the test may still be used using the summary subgroup estimates as the plausible effect size for each study in that subgroup.

Finally, the most challenging application of the concept may pertain to whole domains. It is unlikely that the extent of bias is the same across all research questions considered under a wider domain. Bias may have affected some questions more than others. However, evaluating the domain as a whole may provide interesting overall insights about the average bias in a scientific field at-large.

Acknowledgement

We are thankful to Dr JL Vevea for providing software to calculate weight functions in the selection models, and for useful advice on its use.

References

1. **Neslon N, Rosenthal R, Rosnow RL.** Interpretation of significance levels by psychological researchers. *Am Psychol* 1986; **41**: 1299–301.
2. **Cohen J.** The earth is round (p-less-than .05). *Am Psychol* 1994; **49**: 997–1003.
3. **Rosenthal R, Gaito J.** The interpretation of levels of significance by psychological researchers. *J Psychol* 1963; **55**: 33–8.
4. **Rosenthal R, Gaito J.** Further evidence for the cliff effect in the interpretation of levels of significance. *Psychol Rep* 1964; **15**: 570.
5. **Ioannidis JP.** Why most published research findings are false. *PLoS Med* 2005; **2**: e124.
6. **Dickersin K, Min YI.** Publication bias: the problem that won't go away. *Ann N Y Acad Sci* 1993; **703**: 135–146
7. **Dickersin K, Min YI, Meinert CL.** Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *JAMA* 1992; **267**: 374–378
8. **Dickersin K.** The existence of publication bias and risk factors for its occurrence. *JAMA* 1990; **263**: 1385–1389
9. **Chalmers I.** Underreporting research is scientific misconduct. *JAMA* 1990; **263**: 1405–08.
10. **Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR.** Publication bias in clinical research. *Lancet* 1991; **337**: 867–72.
11. **Ioannidis JP.** Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA* 1998; **279**: 281–6.
12. **Light RJ, Pillemer DB.** *Summing up. The science of reviewing research*, London: Harvard University Press, 1984.
13. **Begg CB, Mazumdar M.** Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994; **50**: 1088–101.
14. **Egger M, Davey Smith G, Schneider M, Minder C.** Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; **315**: 629–34.
15. **Sterne JA, Egger M.** Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001; **54**: 1046–55.
16. **Sterne JA, Gavaghan D, Egger M.** Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol* 2000; **53**: 1119–29.
17. **Hedges LV.** Modeling publication selection effects in meta-analysis. *Stat Sci* 1992; **7**: 246–55.
18. **Copas JB.** What works?: Selectivity models and meta-analysis. *J Royal Stat Assoc* 1999; **162**: 95–109.
19. **Hedges LV, Vevea JL.** Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *J Educ Behav Stat* 1996; **21**: 299–332.
20. **Vevea JL, Hedges LV.** A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* 1995; **60**: 419–35.
21. **Iyengar S, Zhao PL.** Maximum likelihood estimation for weighted distributions. *Stat Probab Lett* 1994; **50**: 438–60.
22. **Ioannidis JP.** Differentiating biases from genuine heterogeneity: distinguishing artifactual from substantive effects. In Rothstein HR, Sutton AJ, Borenstein M eds. *Publication bias in meta-analysis – Prevention, assessment and adjustments*, Sussex: John Wiley, 2005.
23. **Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG.** Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004; **291**: 2457–65.
24. **Chan AW, Krleza-Jeric K, Schmid I, Altman DG.** Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *CMAJ* 2004; **171**: 735–40.
25. **Kyzas PA, Loizou KT, Ioannidis JP.** Selective reporting biases in cancer prognostic factor studies. *J Natl Cancer Inst* 2005; **97**: 1043–55.
26. **Michels KB, Rosner BA.** Data trawling: to fish or not to fish. *Lancet* 1996; **348**: 1152–53.
27. **Weiss RB, Gill GG, Hudis CA.** An on-site audit of the South African trial of high-dose chemotherapy for metastatic breast cancer and associated publications. *J Clin Oncol* 2001; **19**: 2771–77.
28. **Ransam J, Buyse M, George SL,** *et al*. Fraud in medical research: an international survey of biostatisticians. *Control Clin Trials* 2000; **21**: 415–27.
29. **Gardner W, Lidz CW, Hartwig KC.** Authors' reports about research integrity in clinical trials. *Contemp Clin Trials* 2005 **26**: 244–51.
30. **Al-Marzouki S, Evans S, Marshall T, Robert I.** Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *BMJ* 2005; **331**: 267–70.
31. **Lau J, Ioannidis JPA, Schmid CH.** Quantitative synthesis in systematic reviews. *Ann Intern Med* 1997; **127**: 820–26.
32. **Engels EA, Schmid CH, Terrin N, Olkin I, Lau J.** Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med* 2000; **19**: 1707–28.
33. **Terrin N, Schmid CH, Lau J, Olkin I.** Adjusting for publication bias in the presence of heterogeneity. *Stat Med* 2003; **22**: 2113–26.
34. **Lau J, Ioannidis JPA, Schmid CH.** Summing up evidence: one answer is not always enough. *Lancet* 1998; **351**: 123–7.
35. **Preston C, Ashby D, Smith R.** Adjusting for publication bias: modelling the selection process. *J Eval Clin Pract* 2004; **10**: 313–22.

36. **Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR.** Bayesian methods in health technology assessment: a review. *Health Technol Assess* 2000; **4**: 1–130.
37. **Sutton AJ, Abrams KR, Jones DR.** Generalized synthesis of evidence and the threat of dissemination bias: the example of electronic fetal heart rate monitoring (EFM). *J Clin Epidemiol* 2002; **55**: 1013–24.
38. **Trikalinos TA, Churchill R, Ferri M,** *et al*. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *J Clin Epidemiol* 2004 **57**: 1124–30.
39. **Ioannidis JPA, Lau J.** Evolution of treatment effects over time: empirical evidence from recursive cumulative meta-analyses. *Proc Natl Acad Sci USA* 2001; **98**: 831–6.
40. **Counsell C.** Formulating questions and locating primary studies for inclusion in systematic reviews. *Ann Intern Med* 1997; **127**: 380–7.
41. **Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N.** Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 2004; **96:** 434–42.
42. **Efron B, Tibshirani R.** Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol* 2002; **23:** 70–86.
43. **Spiegelhalter DJ, Abrams KR, Myles JP.** *Bayesian approaches to clinical trials and health-care evaluation*, Chichester: John Wiley & Sons, 2004.
44. **Lilford RJ, Braunholtz D.** The statistical basis of public policy. BMJ 1996; **313**: 603–7.
45. **Carlin BP, Louis TA.** *Bayes and empirical Bayes methods for data analysis*, Boca Raton: Chapman and Hall/CRC, 2000.
46. **Greenland S.** Power, sample size and smallest detectable effect determination for multivariate studies. *Stat Med* 1985; **4**: 117–27.
47. **Lui KJ.** Sample size determination for case-control studies: the influence of the joint distribution of exposure and confounder. *Stat Med* 1990; **9**: 1485–93.
48. **Lubin JH, Gail MH.** On power and sample size for studying features of the relative odds of disease. *Am J Epidemiol* 1990; **131**: 552–66.
49. **Pan Z, Trikalinos TA, Kavvoura FK, Lau J, Ioannidis JP.** Local literature bias in genetic epidemiology: an empirical evaluation of the Chinese literature. *PLoS Med* 2005; **2**: e334.
50. **Oxman AD, Guyatt GH.** A consumer's guide to subgroup analyses. *Ann Intern Med* 1992; **116**: 78–84.
51. **Ioannidis JPA, Rosenberg PS, Goedert JJ, O'Brien TR.** Meta-analysis of individual participants' data in genetic epidemiology. *Am J Epidemiol* 2002; **156**: 204–10.
52. **DerSimonian R, Laird N.** Meta-analysis in clinical trials. *Control Clin Trials* 1986; **7**: 177–88.
53. **Higgins JP, Thompson SG.** Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002; **21**: 1539–58.

Appendix – Power calculations

In all meta-analyses the observed power of each study to detect a difference equal to the summary OR given the sample size in each arm and the proportion of events in the controls has been calculated based on 10 000 simulations in Intercooled Stata 8.2 (Stata Corp, College Station, TX, USA). For any given study i in a meta-analysis of k studies, let $n_{1,i}$ and $n_{2,i}$ be the (observed) sample sizes in the treatment (or intervention, or exposure) and control arms, respectively. Let $r_{1,i}$ and $r_{2,i}$ be the observed events, and $\pi_{1,i}$, $\pi_{2,i}$ be the true (latent) pertinent risks in these two arms. We assume that the events follow binomial distributions: $r_{1,i}$ ~Bin($\pi_{1,i}$, $n_{1,i}$), and $r_{2,i}$ ~Bin($\pi_{2,i}$, $n_{2,i}$).

Let θ be the true OR that all studies in the meta-analysis aim to estimate. A good estimator for θ is the observed summary OR, $\hat{\theta}$. The estimator $\hat{\pi}_{2,i}$ of the true risk $\pi_{2,i}$ in the control arm of each study is given by $\hat{\pi}_{2,i} = r_{2,i}/n_{2,i}$. Then, the estimator $\hat{\pi}_{1,i}$ of the true risk $\pi_{1,i}$ in the intervention (or exposure) arm of each study is $[1+[(n_{2,i} - r_{2,i})/\hat{\theta} r_{2,i}]]^{-1}$

We simulated each study i 10 000 times, selecting the number of events in the intervention and control arms from the pertinent binomial distributions (i.e., drawing random event numbers from *Bin*($\hat{\pi}_{1,i}$) and *Bin*($\hat{\pi}_{2,i}$,$n_{2,i}$), respectively). The observed power was defined as the proportion of simulated studies among the 10 000 replicates in which the number of events across the two arms differed beyond chance, based on Fisher's exact test.

When working with distributions of effect, we split the interval covering ±3.5 standard deviations of the mean effect into 101 segments of equal length. The middle value of each segment was used to calculate the *E* estimate; these estimates were weighted by the probability density value at the middle value of their segment and numerically integrated over the whole range of assessed effect values, excluding the segment containing the null effect. The software modules for the simulation-based power calculations are available from the authors or can be downloaded from the website www.dhe.med.uoi.gr