Journal of Mathematical Psychology 57 (2013) 184–187

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Commentary

Clarifications on the application and interpretation of the test for excess significance and its extensions

John P.A. Ioannidis *

Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA
Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA, USA

HIGHLIGHTS

- The test for excess significance depends on several assumptions.
- Interpretation of the test should be cautious.
- Significance-related biases may follow a complex pattern.
- Likelihood ratio estimates can be used to generate the post-test probability of bias.
- Correcting effect estimates for bias is not necessarily reliable.

ARTICLE INFO

Article history:
Available online 28 April 2013

Keywords:
Bias
Publication bias
Statistical significance
Selective reporting

ABSTRACT

This commentary discusses challenges in the application of the test for excess significance (Ioannidis & Trikalinos, 2007) including the definition of the body of evidence, the plausible effect size for power calculations and the threshold of statistical significance. Interpretation should be cautious, given that it is not possible to separate different mechanisms of bias (classic publication bias, selective analysis, and fabrication) that lead to an excess of significance and in some fields significance-related biases may follow a complex pattern (e.g. Proteus phenomenon and occasional preference for "negative" results). Likelihood ratio estimates can be used to generate the post-test probability of bias, and correcting effect estimates for bias is possible in theory, but may not necessarily be reliable.

The test for excess significance (TES) was originally introduced (Ioannidis & Trikalinos, 2007) aiming to evaluate whether the observed (O) number of statistically significant results in a body of evidence is too large compared to their expected (E) number. TES has already been applied to several different fields of biomedical research, including meta-analyses of randomized trials (Ioannidis & Trikalinos, 2007), genetic association studies of diverse diseases such as Alzheimer's disease (Kavvoura et al., 2008), cutaneous melanoma (Chatzinasiou et al., 2011) and pre-eclampsia (Staines-Urias et al., 2012), brain volume abnormality studies (Ioannidis, 2011), cancer biomarkers (Tsilidis, Papatheodorou, Evangelou, & Ioannidis, 2012), and genetic associations of brain functions (Murphy et al., in press). In the psychological sciences, Francis has used the same principles to identify potential bias in a number of experimental claims in psychology (Francis, 2012a,b); now he offers an interesting overview on his approach (Francis, 2013). I take this as an opportunity to clarify some issues about the application and interpretation of these tests.

1. Nomenclature

Francis uses the terms consistency and inconsistency and defines the test as examining the consistency of a set of reported experiments (Francis, 2013). I am afraid that these terms may create some confusion in the literature. The terms "consistency" and "inconsistency" are used interchangeably with the terms "homogeneity" and "heterogeneity" in the field of meta-analysis (Higgins, Thompson, Deeks, & Altman, 2003), and TES is applied typically when many studies and meta-analyses thereof are performed. Thus, I prefer to continue using the term excess significance, which also conveys more directly what TES evaluates.

2. Definition of body of evidence

Francis has typically applied the test to probe for bias in sets of multiple experiments published by the same team in the same paper. The experiments are not necessarily the same, but may deviate

* Correspondence to: Stanford Prevention Research Center, 1265 Welch Rd, MSOB X306, Stanford University School of Medicine, Stanford, CA 94305, USA.
E-mail address: jioannid@stanford.edu.

0022-2496/$ – see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.jmp.2013.03.002

in important aspects that may or may not induce also differences in the genuine effect sizes. The number of studies included in such bodies of evidence is usually relatively small, often <10. Nevertheless TES always shows that there are too many significant results, because in the examples that Francis has probed typically all experiments have shown statistically significant results.

In biomedical sciences and related fields, I and others have focused on applications of TES in meta-analyses of studies on the same research question done by different teams and on larger bodies of evidence that include many meta-analyses of studies on the same field, typically with hundreds of studies. When excess significance is detected in a large body of evidence including many meta-analyses of many research questions, it is unknown whether all research questions are equally affected by reporting biases. However, TES is appropriate in identifying reporting biases prevalent in the field at-large.

3. Definition of plausible effect size

TES results depend on the assumptions about the plausible effect size, since these directly affect the power estimates for each study. This is a clear limitation, but, as Francis shows, the conclusions tend to be fairly robust when different assumptions are made about the plausible effect size within a sensible range. I would like to add here some additional considerations. First, it is possible to perform power calculations assuming a distribution of a plausible effect instead of a point-estimate. This distribution may be the effect size distribution of a fixed or random effects meta-analysis. In our experience (Ioannidis & Trikalinos, 2007), results are very similar to simply using the point estimate. Second, a fixed effects estimate seems a natural choice, especially when between-study heterogeneity is not prominent or seriously supported by theoretical anticipation. When many meta-analyses are examined together in a larger body of evidence, the summary effect of each meta-analysis has to be considered separately in power calculations for the studies that it contains. Third, a random effects estimate might seem reasonable when substantial between-study heterogeneity is documented or strongly suspected. However, random effects are a poor choice in the presence of selective reporting biases that affect more prominently smaller studies — this scenario can be common and it underlies the traditional small-study effect tests based on funnel plot asymmetry (Sterne et al., 2011): then, random effects estimates are substantially biased (inflated) and the result of the largest study is the best choice for the plausible effect, unless there are concerns about the quality of the largest study. In some research fields, however, there has never been a large, well-conducted study. Then, assumptions and inferences should be cautious. For example, randomized trials in animal models are almost ubiquitously of very small sample size (Sena, van der Worp, Bath, Howells, & Macleod, 2010).

Summary effect sizes are more likely to be over- rather than under-estimates of the true effects (Pereira & Ioannidis, 2011). Even when fixed effects assumptions seem tenable and not refuted by statistical homogeneity testing, this is not fully reassuring. By default, excess significance bias for whatever reason will tend to inflate the observed summary effect size. Excess significance bias may also generate between-study heterogeneity, but statistical testing (e.g. with Cochran's Q test) has low power to detect heterogeneity for most research questions and meta-analyses (Pereira, Patsopoulos, Salanti, & Ioannidis, 2010). Thus, in the presence of bias, inferences of TES are conservative and excess significance may be missed. Conversely, TES may yield some false signals of bias when between-study heterogeneity exists due to genuine reasons rather than bias (Ioannidis & Trikalinos, 2007; Johnson & Yuan, 2007). I revisit this issue in the section of post-test probability of bias below.

4. Definition of nominal statistical significance threshold

Francis has used the $p = 0.05$ threshold to separate "positive" from "negative" results. This threshold acts as an attractor for investigators in many fields (Bakker, van Dijk, & Wicherts, 2012; Simmons, Nelson, & Simonsohn, 2011), but it is not absolute. Some fields increasingly require more stringent thresholds and/or use multiplicity-corrections, some investigators may bias the results of their analysis too much and strike to get p-values much below 0.05, and investigators occasionally make leaps to claim significance for trends not reaching $p<0.05$. It is interesting to study the excess (or deficit) of "positive" results in different ranges of p-values to understand better p-value distributions in different scientific fields. This can be done by an extension of the excess significance concept to examine p-value bins; methods and an application are shown in Kavvoura et al. (2008). Empirical application of this approach in genetic epidemiology shows some interesting observations: the pattern of excess significance in different p-value bins may differ in situations where there is or not between-study heterogeneity; and the bin of p-values of 0.05–0.15 may also show an excess, perhaps because some investigators consider such results good enough or use a different analytical practice that makes the presented p-values better than they really are. There is also evidence from randomized clinical research that often investigators put "spin" in their interpretation and claim significance for results that are not nominally significant (Boutron, Dutton, Ravaud, & Altman, 2010). Furthermore, I expect that bin patterns may vary in different scientific fields and they are worth investigating in large scale. For example, p-hacking may tend to create a heap of p-values slightly less than 0.05 in the psychological sciences, but in genome studies that operate with genome-wide thresholds of $p < 5 \times 10^{-8}$ (Chanock et al., 2007), p-value hacking may result in heaps of p-values close to that "telescoped" threshold.

5. Separating mechanisms of reporting bias

There are many mechanisms of selective reporting. I agree with Francis that fabrication bias, i.e. clear fraud, is unlikely to be a major player in most scientific fields. However, I also doubt that classic publication bias is the main explanation for excess significance in most fields. Classic publication bias means that "negative" results entirely disappear (by authors and/or editors/reviewers). The prevalence of this bias may vary across different scientific fields, proportional to the ease of making a study disappear and the difficulty of making a "negative" study become "positive" with changes in the analysis plans and/or outcome definition. Most datasets can be data-dredged to yield eventually a nominally statistically significant result, even if the original intention and analysis plan has yielded a non-significant one. Often there is not even a pre-specified intention and analysis plan, let alone protocol. Registration for some types of designs (e.g. randomized trials) has been a major step forward for diminishing classic publication bias, but a priori registration of protocols is not yet common and detailed enough to abort questionable research practices in data analyses. A particularly prevalent form of bias in biomedicine may be the reporting of only some among many outcomes in a study or changing outcomes and analyses plans after the study has been completed and analyzed (Dwan et al., 2011). Analytical flexibility is apparently also very common in the psychological sciences (Fanelli, 2010; John, Loewenstein, & Prelec, 2012). Thus selective analysis and outcome reporting may be more prevalent than classic publication bias (Ferguson & Brannick, 2012). Emphasis on classic publication bias can sometimes even be absurd, e.g. when investigators apply tests to detect such bias in situations where all studies are pre-registered or in prospective meta-analysis of individual level data, situations where classic publication bias (but not selective analysis and outcome reporting bias) can be excluded by default.