



The ongoing tyranny of statistical significance testing in biomedical research

Andreas Stang, Charles Poole, Oliver Kuss

► To cite this version:

Andreas Stang, Charles Poole, Oliver Kuss. The ongoing tyranny of statistical significance testing in biomedical research. *European Journal of Epidemiology*, 2010, 25 (4), pp.225-230. 10.1007/s10654-010-9440-x . hal-00580111

HAL Id: hal-00580111

<https://hal.science/hal-00580111>

Submitted on 26 Mar 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The ongoing tyranny of statistical significance testing in biomedical research

Andreas Stang · Charles Poole · Oliver Kuss

Received: 16 November 2009 / Accepted: 4 March 2010 / Published online: 26 March 2010
© Springer Science+Business Media B.V. 2010

Abstract Since its introduction into the biomedical literature, statistical significance testing (abbreviated as SST) caused much debate. The aim of this perspective article is to review frequent fallacies and misuses of SST in the biomedical field and to review a potential way out of the fallacies and misuses associated with SSTs. Two frequentist schools of statistical inference merged to form SST as it is practised nowadays: the Fisher and the Neyman-Pearson school. The *P*-value is both reported quantitatively and checked against the α -level to produce a qualitative dichotomous measure (significant/nonsignificant). However, a *P*-value mixes the estimated effect size with its estimated precision. Obviously, it is not possible to measure these two things with one single number. For the valid interpretation of SSTs, a variety of presumptions and requirements have to be met. We point here to four of them: study size, correct statistical model, correct causal model, and absence of bias and confounding. It has been stated that the *P*-value is perhaps the most misunderstood statistical concept in clinical research. As in the social sciences, the tyranny of SST is

still highly prevalent in the biomedical literature even after decades of warnings against SST. The ubiquitous misuse and tyranny of SST threatens scientific discoveries and may even impede scientific progress. In the worst case, misuse of significance testing may even harm patients who eventually are incorrectly treated because of improper handling of *P*-values. For a proper interpretation of study results, both estimated effect size and estimated precision are necessary ingredients.

Keywords Statistics · *P*-value · Confidence intervals

Since its introduction into the biomedical literature, null hypothesis significance testing (abbreviated as SST) has caused much debate. As early as 1919, Boring criticized use of SST [1]. In 1957, Hogben described the logical and practical errors in theory and teaching of SST [2]. In 1970, Morrison and Henkel [3] published the compendium entitled “The Significance Test Controversy”. More recently, Cohen wrote a very influential paper on the SST controversy [4]. Despite many warnings and critical remarks about the use of SST in biomedical projects, SST is still one of the most prevalent statistical procedures in biomedical publications. The aim of this perspective article is to review frequent misconceptions and misuses of SST in the biomedical field and to review a potential way out of the fallacies associated with SSTs.

What is a *P*-value?

A *P*-value may be viewed as the probability of obtaining an estimate at least as far from a specified value (most often the null value, i.e., the value of no effect, the so called “nil

A. Stang (✉)
Institut für Klinische Epidemiologie, Medizinische Fakultät,
Martin-Luther-Universität Halle-Wittenberg, Magdeburger
Str. 8, 06097 Halle (Saale), Germany
e-mail: andreas.stang@medizin.uni-halle.de

C. Poole
Department of Epidemiology, UNC Gillings School of Global
Public Health, University of North Carolina, Chapel Hill,
NC, USA

O. Kuss
Institut für Medizinische Epidemiologie, Biometrie und
Informatik, Universitätsklinikum Halle und Medizinische
Fakultät, Martin-Luther-Universität Halle-Wittenberg,
Halle (Saale), Germany

hypothesis”) as the estimate we have obtained, if the specified (null or test) value were (note the subjunctive) the true value [5], p. 220. In other words, the P -value is a tail area probability based on the observed effect estimate; it is calculated as the probability of an effect estimate *as large as or larger than the observed estimate* (more extreme in the tails of the distribution), assuming the null hypothesis were true [6]. As Miettinen stated “The P -value is a function of the data computed under the statistical model that underlies the analysis.” [7]. It should be noted that a P -value is a confounded (mixed) measure: it mixes the estimated effect size with its estimated precision, both crucial aspects of the data [8]. Obviously, it is not possible to present two quantities (estimated effect size and precision) by one single number. We can see this most clearly when the test statistic is Z : the parameter estimate divided by its estimated standard error. The higher the parameter estimate or the lower the estimated standard error, the greater the (absolute) value of Z and the lower the (two-sided) P -value.

Perhaps more important than knowing what a P -value actually *is*, is understanding what a P -value *is not*. Goodman gave as many as 12 different misconceptions of the P -value [9]. The most pervasive of these misconceptions is interpreting the P -value as the probability of the null hypothesis. The P -value is derived from the study data, assuming that the null hypothesis is true. It cannot, therefore, make a concurrent statement of the probability of that hypothesis. In addition, the P -value generally exaggerates the evidence against the null hypothesis, thus calling “into question the validity of much published work based on comparatively small, including 0.05, P values” [10]. This is because true posterior probabilities (as they could be derived, for example, in a Bayesian framework) are generally larger than the corresponding P -values.

What is a significance test?

Two frequentist schools of statistical inference merged to form SST as it is practised nowadays: the Fisher and the Neyman-Pearson school. The logic of Fisher’s null hypothesis testing consisted of two steps: (1) One sets up a null hypothesis. (2) One reports the P -value, called the level of significance. This approach “lacks a specific alternative hypothesis and therefore the concepts of statistical power, Type II error and theoretical effect size have no place in Fisher’s framework” [11]. Fisher saw the P -value as an index measuring the strength of evidence against the null hypothesis: the lower the P -value, the stronger the evidence. He argued that “no scientific worker has a fixed level of significance at which from year to year, and in all circumstances he rejects hypotheses; he rather gives mind

to each particular case in the light of his evidence and his ideas” [12, 13].

In contrast, Neyman and Pearson’s decision theoretic approach included the specification of two statistical hypotheses the null hypothesis, H_0 , and an alternative hypothesis, H_1 , the Type I and II error probabilities (α and β), and sample size before the trial or experiment is started. Based on these specifications, rejection regions for both hypotheses are defined. If the data falls into the rejection region of H_0 , H_0 is rejected in favour of H_1 , otherwise H_0 is not rejected. Thus, we use a decision rule for interpreting the results of our experiment or trial in advance, and the result of our analysis is simply the rejection or acceptance of the null hypothesis.

Today’s practice of SST is a mixture of both schools. The P -value is both reported quantitatively and checked against the α -level to decide on the destiny of the null hypothesis. In the world of SST, $P < 0.05$ warrants a call for action and the farther below 0.05 P falls, the stronger the call. $P \geq 0.05$ means that the status quo is to be maintained and the closer to 1.0 the P -value falls, the more secure the status quo seems to be. Studies with $P \geq 0.05$ are commonly considered as failed studies, that is, studies that failed to “achieve” significance [14]. Does it make sense to adopt a new therapy because the P -value of a single study was 0.048, and at the same time to reject another therapy because the P -value was 0.052? That did not make sense to Neyman and Pearson [15], but it does to some today [16].

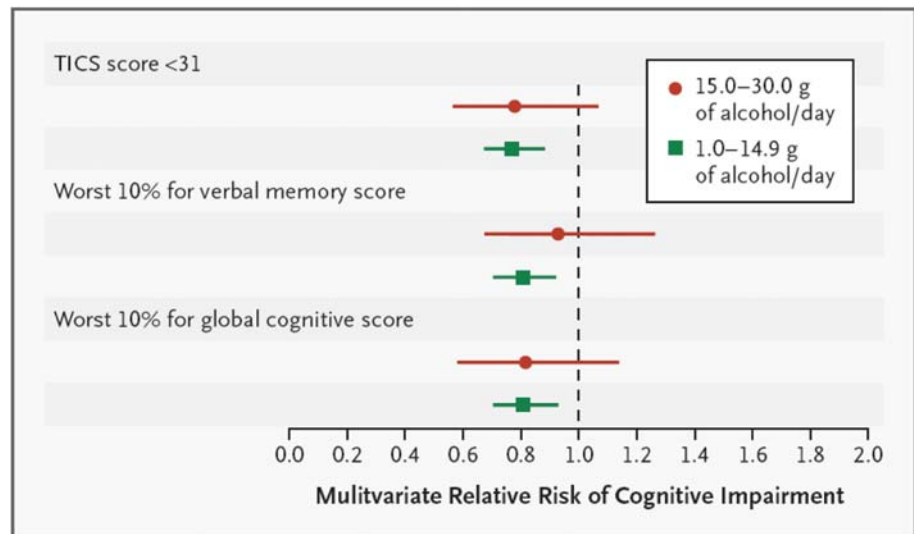
Presumptions for the interpretability of statistical significance tests

For valid interpretation of SSTs, several presumptions or requirements have to be met. We point here to four of them: study size, correct statistical model, correct causal model, and absence of bias and confounding.

Study size

Given a fixed effect size, the P -value is a function of the sample size. Studies with small sample sizes tend to miss important clinical differences in significance tests. In contrast, studies with large study sizes tend to produce significant findings that are clinically meaningless. Despite these well-known relations between study size and statistical significance, study size related misinterpretations of SSTs are ubiquitous even after decades of educational articles. For example, the authors of a recent randomized trial hailed a statistically significant ($P = 0.04$) increase in median survival from 10.1 months to 11.3 among patients with advanced non-small-cell lung cancer as “a new

Fig. 1 Association between daily alcohol consumption and risk of cognitive impairment in the nurses health study [21]. Multivariate relative risk of cognitive impairment among women who drank 1.0 to 14.9 g of alcohol per day, as compared to nondrinkers. TICS denotes the telephone interview for cognitive status. The verbal memory score reflects the results of immediate and delayed recall of both the TICS 10-word list and the East Boston Memory Test. The global cognitive score is the average of the results of all cognitive tests



Copyright © 2005 Massachusetts Medical Society. All rights reserved.

treatment option” [16]. The clinical significance of that statistically significant difference is questionable.

In 1995, Altman and Bland noted that “the non-equivalence of statistical significance and clinical importance has long been recognized” [17]. It is just as well understood, and just as frequently ignored, that findings of potentially great clinical importance may be statistically non-significant. For example, the frequently replicated association between greatest tumor dimension (GTD) and prognosis is one component of the TNM-staging of uveal melanoma [18]. The authors of one study [19] that replicated this association, with an estimated relative risk of poor outcome of 2.40 and a 95% CI of 0.98–5.88, failed to emphasize this result because, “...of the clinical characteristics, only the presence of extraocular extension was associated with poor outcome”. By “associated,” the authors meant statistically significantly associated. The exceedingly plausible, highly replicated, clinically significant association with GTD was treated as no association merely because $P \geq 0.05$. We note in passing that a post hoc power analysis would not have changed that misinterpretation and, contrary to common beliefs, does not give any additional information in such situations [20].

Stampfer et al. [21] recently studied the association between daily alcohol drinking at baseline and cognitive impairment (three primary outcomes) during follow-up in the Nurses Health Study among 11,102 women. They concluded that “women [drinking 1.0–14.9 g of alcohol per day] had significantly better mean cognitive scores than nondrinkers on all three primary outcomes”. They stressed that there were no significant associations with higher levels of drinking (15.0–30.0 g per day). A closer look at the distribution of drinking habits at baseline reveals that 51% of all women were nondrinkers, 44% were moderate

drinkers (1.0–14.9 g per day), and only 5% (648 women) were in the highest drinking category. The results (Fig. 1) showed essentially the same association for the moderate and higher drinking levels compared with zero consumption, with a less precise estimate for the higher category owing to its low prevalence in the study population.

Correct statistical model

The validity of the calculated P -value, and therefore of the SST, depends on the correct statistical model. Referring again to the study of Stampfer et al., the reported estimated relative risks are derived from a standard logistic regression model with alcohol consumption and several other confounders as covariates. Using this model explicitly assumes that (1) all confounders act linearly (and not, for example, in a quadratic relation) on the log odds of cognitive impairment, (2) there are no interactions between confounders, (3) all important confounders were included in the model, (4) the link function between the probability of cognitive impairment and the confounders is indeed the logit function, and (5) no bias-inducing adjustment variables were included. Violations of any of these presumptions can potentially result in grossly biased P -values.

Correct causal model

To validly estimate the association between a treatment and outcome, presumptions have to be met regarding the timeline of treatment and treatment effect (see Fig. 2). Let us assume a beneficial treatment has to last for at least 6 months to reduce the risk of an adverse outcome and it takes an additional 6 months for that risk reduction to occur. A clinical trial of this beneficial treatment would be

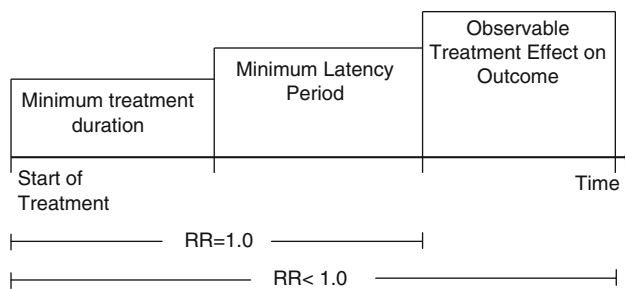


Fig. 2 Minimum treatment duration, minimum latency period, and observable treatment effect on outcome

meaningless if the treatment duration, the follow-up duration or both are too short. In the field of outcomes research, rough a priori ideas about these time intervals are often available. However, a priori ideas about these intervals may be more vague in other fields of biomedical research, including the study of etiologic factors. The Women's Health Initiative (WHI) study assessed the association between hormone replacement therapy (HRT) and risk of breast cancer [22]. The risk of breast cancer among HRT users was not increased during the first 3–4 years of follow-up. Thereafter, the cumulative risks diverged, showing a higher risk among women with HRT. If the follow-up time had been only 3 years, the increased breast cancer risk could not have been observed in that study.

No bias and no confounding

Although it sounds like a truism, the interpretability of P -values and SSTs also depends on the absence of bias and confounding, because bias and confounding can influence the test size, power, or both—for better or worse. For Ronald Fisher, randomization was necessary to ensure that the estimates of error and tests of significance “should be fully valid” [23]. In contrast to randomized trials, in which randomization provides a solid theoretical basis for the probability models from which P -values are derived, such a mechanism is missing in observational studies. Hence, precise and hairsplitting interpretations of P -values and confidence intervals are even less wise in observational research than they are in randomized trials [24].

Interpretation of nullhypothesis significance tests in presence of prior studies

How can we ignore available prior studies that dealt with the same study question as our own study when interpreting our own results [14]? For example, a nonsignificant association between smoking (ever vs. never) and lung cancer (e.g., $RR = 2.0$, 95% CI: 0.9–4.2) in a cohort study

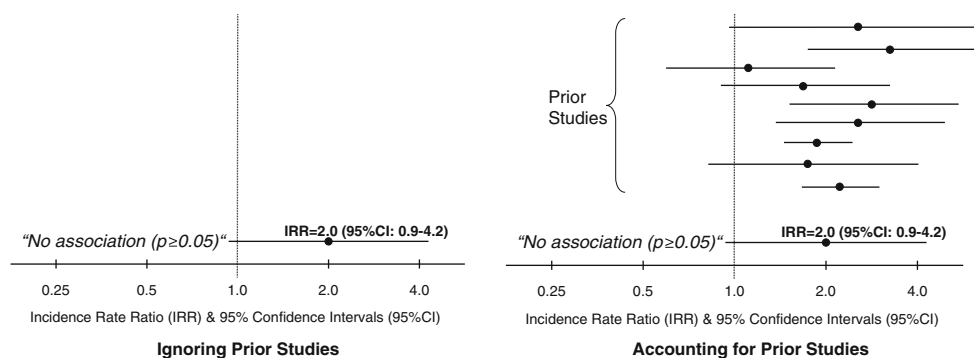
should not prompt investigators to claim that there is no association between smoking and risk of lung cancer and therefore contradicting evidence. Their non-significant finding must be seen in light of the compelling evidence that smoking is a risk factor of lung cancer. Yet, many authors discuss their own study findings as if their study would be the first and only study addressing the specific study question, as in the study of GTD and uveal melanoma prognosis discussed above [19]. In medical research, decisions are rarely made on the basis of a single trial or experiment [25] and many studies that are not statistically significant contribute to the overall weight of the evidence upon which associations come to be accepted as causal. Figure 3a, b illustrate the two ways of thinking: ignoring any prior evidence and accounting for it. Conducting systematic reviews has the immense side benefit that one comes to view the contributions of individual studies in a greatly improved light [14].

The mixture of SST schools (Fisher vs. Neyman-Pearson) tempts investigators with strong prior beliefs of their hypotheses to switch between SST schools depending on the results of their SSTs. For example, consider an investigator who strongly believes that a factor would be an important predictor of a poor outcome of a disease observes a non-significant association at $\alpha = 0.05$ (the predefined Type I error according to Neyman and Pearson) between this factor and the outcome. This investigator might write, “Our study result was close to significance ($P = 0.08$),” and might even seriously discuss the “almost positive” study finding or “trend” [26]. Frequently, those investigators claim that their study finding would have become significant if the study size would have been larger. However, if the investigator does not believe in a real association between a factor and an outcome and the association becomes significant, this investigator might write, “Our study result ($P = 0.03$) may be due to chance,” and might argue why this finding most likely is an artefact due to unexplained errors including bias and confounding—again circumventing the Neyman and Pearson decision-theoretic approach.

Null hypothesis significance testing where chance plays the only role or no role

P -values are frequently completely senseless. One example is the typical table 1 of the publication of a randomized controlled trial (two groups) presenting the distributions of study characteristics just after randomization by group. These tables are frequently accompanied by SSTs comparing the treatment groups with each other in order to assess comparability (exchangeability) between the groups. The mechanism that produced these distributions is a chance mechanism by definition. That is, we know in

Fig. 3 Interpretation of study results ignoring and accounting for prior studies



advance that the SST of equal means or proportions in treatment groups is true and so it is completely useless to test this hypothesis. If randomization worked, we expect that on average one out of 20 variables will show statistically significant differences ($P < 0.05$) between the groups. Therefore, the assessment of chance as an explanation of an imbalance of confounders between the groups by SSTs is unreasonable as the answer is yes, regardless what the result of the SST is. It is the magnitude of imbalances between the groups that is important to consider [27]. The meaning of these imbalances is based not on statistics but on subject matter judgment. For example, in a large antihypertensive treatment trial, at randomization one group may have a mean systolic blood pressure of 148 mmHg (standard deviation, $SD = 5.6$) and the other group a mean systolic blood pressure of 150 mmHg ($SD = 5.7$). Is the 2 mmHg difference of the mean blood pressure values a relevant difference from a clinical perspective? This question cannot be answered by SSTs. It can be answered by clinical judgment or by a quantitative evaluation of confounding.

Similar unreasonable uses of SSTs can be observed in observational studies. For example, it is popular to study the distribution of matching factors (for example age) between cases and controls (see for example [28]). However, the similarity of the distributions of the matching factors among cases and controls is dependent on the quality of the matching, which is under control of the investigator. Comparisons of those distributions are therefore not scientific findings.

What is a confidence interval?

For a given estimate, a 95% confidence interval (CI) is the set of all parameter values (i.e., hypotheses) for which $P \geq 0.05$ [24]. If the underlying statistical model is correct and there is no bias, the proportion of CIs derived over unlimited repetitions of the study containing the true parameter value is no smaller than 95% and is usually close to that value. This means that a confidence interval produces a move from a single value, or point estimate, to a

range of possible effects in the population about which we want to draw conclusions.

A confidence interval conveys information about both magnitude of effect and precision and therefore is preferable to a single P -value [29]. Although CIs can be used for a mechanistic accept–reject dichotomy by rejecting any null hypothesis if the value indicating the null effect is outside the interval, they can and should move us away from that dichotomy. The width of a CI indicates how much the point estimate is influenced by chance. For instance, when a relative risk estimate of 4.1 (95%CI: 1.2–14.0), $P = 0.02$, is compared with an estimate of 1.4 (95%CI: 0.8–2.4), $P = 0.20$, the first is much more influenced by chance (as reflected by its wider CI) and is therefore much less trustworthy, even though the first estimate is statistically significant and the second is not [24]. Since 1999, the Board of Scientific Affairs (BSA) of the American Psychological Association (APA) has recommended the presentation of confidence intervals for all effect size estimates [30].

Conclusions

It has been stated that the P -value is perhaps the most misunderstood statistical concept in clinical research [6]. As in the social sciences [31], the tyranny of significance testing is still highly prevalent in the biomedical literature, even after decades of warnings against SST [25]. The ubiquitous misuse and even tyranny of significance testing, given the authoritarian way in which it tends to be practiced, threatens scientific discoveries and may even impede scientific progress [9]. In the worst case, it may even harm patients who eventually are incorrectly treated because of improper handling of P -values. As Sterne and Davey Smith recently reminded us, “In many cases published medical literature requires no firm decision: it contributes incrementally to an existing body of knowledge” [13]. Biomedical research is an endeavour in measurement. Its objective is to obtain a valid and precise estimate of a measure of effect, the true value of which may be no effect,

a small and unimportant effect, or a large and meaningful effect. We should strive for good data description and careful interpretation of estimated effect measures and their accompanying precision rather than mechanical significance testing with yes–no answers which lead inherently to biased interpretations. An important way out of significance fallacies such as those we have described is to interpret statistical findings based on confidence intervals that convey both the size and precision of estimated effect measures.

Disclosure None of the authors reports any conflict of interest.

References

1. Boring EG. Mathematical vs. scientific significance. *Psychol Bull.* 1919;15(10):335–8.
2. Hogben LT. Statistical theory: an examination of the contemporary crisis in statistical theory from a behaviourist viewpoint. London: George Allen & Unwin; 1957.
3. Morrison DE, Henkel RE. The significance test controversy: a reader. Chicago: Aldine Pub; 1970.
4. Cohen J. The earth is round ($p < .05$). *Am Psychol.* 1994; 49(12):997–1003.
5. Greenland S, Rothman KJ. Fundamentals of epidemiologic data analysis. In: Rothman KJ, Greenland S, Lash TL, editors. *Modern epidemiology*. 3rd ed. Philadelphia: Wolters Kluwer, Lippincott Williams & Wilkins; 2008. p. 213–37.
6. Blume J, Peipert JF. What your statistician never told you about P-values. *J Am Assoc Gynecol Laparosc.* 2003;10(4):439–44.
7. Miettinen OS. *Theoretical epidemiology*. Albany: Delmar Publishers Inc.; 1985.
8. Lang JM, Rothman KJ, Cann CI. That confounded P-value. *Epidemiology.* 1998;9(1):7–8.
9. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol.* 2008;45(3):135–40.
10. Hubbard R, Lindsay RM. Why p-values are not a useful measure of evidence in statistical significance testing. *Theory Psychol.* 2008;18(1):69–88.
11. Gigerenzer G. Mindless statistics. *J Socio-Econ.* 2004;33: 587–606.
12. Fisher RA. *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd; 1956.
13. Sterne JA, Davey SG. Sifting the evidence—what’s wrong with significance tests? *BMJ.* 2001;322(7280):226–31.
14. Poole C, Peters U, Il'yasova D, Arab L. Commentary: this study failed? *Int J Epidemiol.* 2003;32(4):534–5.
15. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika.* 1928;20A:175–240.
16. Rabe KF. Treating COPD—the TORCH trial, P values, and the Dodo. *N Engl J Med.* 2007;356(8):851–4.
17. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ.* 1995;311(7003):485.
18. Sobin LH, Wittekind Ch. *TNM classification of malignant tumours*. 6th ed. New York: Wiley-Liss, Inc.; 2002.
19. White VA, Chambers JD, Courtright PD, Chang WY, Horsman DE. Correlation of cytogenetic abnormalities with the outcome of patients with uveal melanoma. *Cancer.* 1998;83(2):354–9.
20. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med.* 1994;121(3):200–6.
21. Stampfer MJ, Kang JH, Chen J, Cherry R, Grodstein F. Effects of moderate alcohol consumption on cognitive function in women. *N Engl J Med.* 2005;352(3):245–53.
22. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women’s health initiative randomized controlled trial. *JAMA.* 2002;288(3):321–33.
23. Fisher RA. *The design of experiments*. Edinburgh: Oliver & Boyd; 1935.
24. Poole C. Low P-values or narrow confidence intervals: which are more durable? *Epidemiology.* 2001;12(3):291–4.
25. Rothman KJ. A show of confidence. *N Engl J Med.* 1978; 299(24):1362–3.
26. Pocock SJ, Ware JH. Translating statistical findings into plain English. *Lancet.* 2009;373(9679):1926–8.
27. Altman DG. A fair trial? *Br Med J (Clin Res Ed).* 1984; 289(6441):336–7.
28. Main KM, Kiviranta H, Virtanen HE, Sundqvist E, Tuomisto JT, Tuomisto J, et al. Flame retardants in placenta and breast milk and cryptorchidism in newborn boys. *Environ Health Perspect.* 2007;115(10):1519–26.
29. Rothman KJ. Significance questing. *Ann Intern Med.* 1986; 105(3):445–7.
30. Wilkinson L. Task force on statistical inference. Statistical methods in psychology journals: guidelines and explanations. *Am Psychol.* 1999;54(8):594–604.
31. Loftus GR. On the tyranny of hypothesis testing in the social sciences. *Contemp Psychol.* 1991;36(2):102–5.