ELSEVIER

Journal of Clinical Epidemiology 62 (2009) 1242−1247

Journal of Clinical Epidemiology

COMMENTARY

The Bayesian interpretation of a *P*-value depends only weakly on statistical power in realistic situations

Richard Hooper*

National Heart & Lung Institute, Imperial College London, London, UK

Accepted 3 February 2009

Abstract

Objective: It is often repeated that a low *P*-value provides more persuasive evidence for a genuine effect if the power of the test is high. However, this is based on an argument which ignores the precise *P*-value in favor of simply observing whether *P* is less than some cut-off, and which oversimplifies the possible effect sizes. In a non-Bayesian framework, there are good reasons to think that power does not affect the evidence of a given *P*-value. Here I illustrate the relationship between pre-study power and the Bayesian interpretation of a *P*-value in realistic situations.

Study Design and Setting: A Bayesian calculation, using a conventional prior distribution for the effect size and a normal approximation to the sampling distribution of the sample estimate, where the datum is the precise *P*-value.

Results: Over the range of pre-study powers typical in published research, the Bayesian interpretation of a given *P*-value varies little with power.

Conclusion: A Bayesian analysis with reasonable assumptions produces results remarkably in line with a more simple, non-Bayesian intuition—that the evidence against the null hypothesis provided by a precise *P*-value should not depend on power.

Keywords: Bayes theorem; Hypothesis testing; Inference; *P*-value; Power of a test; Significance test

1. Introduction

P-values and statements of statistical power are familiar features of reports in medical journals. But does power affect the interpretation of *P*? "In the absence of bias," wrote Wacholder et al. in 2004, "three factors determine the probability that a statistically significant finding is actually a false-positive finding"—these factors were the *P*-value, the fraction of tested hypotheses that are true, and "less appreciated" the statistical power of the test [1]. Wacholder's article was widely read—a Web of Science citation search showed that in the four years after its publication it was cited an average of 1.5 times a week. An article in *Nature* from the Wellcome Trust Case Control Consortium, for example, made the point just as strongly—"for a given significance threshold, the probability of a true association depends on the prior odds and crucially, the power" [2]. However, this "crucial" dependence of the interpretation of *P*-values on power unravels under closer scrutiny.

2. The diagnostic testing argument

Wacholder et al. were adapting an argument previously presented by Sterne and Davey Smith [3], which can be traced at least as far back as a 1976 article by Peto et al. [4]. The method, which is a simple example of a Bayesian calculation, is familiar to anyone who has learned how to evaluate the performance of a diagnostic test [5−7pp430−432]. Suppose that 90% of all null hypotheses tested are true—90% of the things we research (risk factors, treatments, and so forth) have no real effect. Suppose also that studies reported in the medical literature have an average statistical power of 50% at the 5% significance level. Then the sensitivity and specificity of the result $P \leqslant 0.05$ are, by the definitions of power and significance, 50% and 95%, respectively (see Table 1). We can then work out the probability that there is no real effect given that $P \leqslant 0.05$, and we find it is 47% (Table 1)—nearly half our "significant" results are false positives.

* Corresponding author. Respiratory Epidemiology & Public Health Group, Imperial College London, Emmanuel Kaye Building, Manresa Road, London SW3 6LR, UK. Tel.: +44-20-7352-8121 ext. 3502; fax: +44-20-7351-8322.

E-mail address: richard.hooper2@imperial.ac.uk

0895-4356/09/$ − see front matter © 2009 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2009.02.004

What is new?

- From a Bayesian viewpoint, the evidence of a given *P*-value is less convincing in studies with very low power, and also in studies with very high power, but within the range typically encountered in published research, the power affects the interpretation of *P* very little.
- From a non-Bayesian viewpoint there are also good reasons to think that power should not affect the evidence of a given *P*-value.
- Arguments made in the literature for higher-powered tests being more convincing have generally been based on an over-simplified argument. Readers can be reassured that the conventional wisdom still works well in most situations: the *P*-value weighs the evidence for an effect, and the confidence interval estimates how big that effect might be.

A more general formula can be derived using the same method—if the proportion of null hypotheses which are true is π_0, and the power to detect a clinically important effect at the α significance level is $(1-\beta)$, then the probability that there is no real effect given that $P \leqslant \alpha$ is $a\pi_0/[a\pi_0 + (1-\beta)(1-\pi_0)]$. Figure 1 illustrates the relationship between power and this probability (assuming $\alpha = 0.05$ and $\pi_0 = 0.9$), showing that the probability of no real effect decreases with increasing power—$P \leqslant \alpha$ argues more persuasively for a genuine effect if the power is greater.

3. Problems with the diagnostic testing argument

One problem is that the calculation described above is only valid if we only know that P is below some fixed cut-off α. Other authors have pointed out that if we know the precise *P*-value the situation is quite different [8,9pp179–184]—in this case the null hypothesis becomes increasingly *more* likely as the power increases, rather than less, as Wacholder et al. implied—a counterintuitive result known as Lindley's paradox after a 1957 article in which it was presented [10,11]. Note that the precise *P*-value is implicit in our data (and usually easily obtained), and if we only record that $P \leqslant \alpha$ we are throwing away evidence.

Wacholder et al. recommended substituting the observed *P*-value for α in the formula in Section 2 above, but this is not the same as making use of the precise *P*-value (observing $P = p$ is not the same thing as being told that $P \leqslant p$). Wacholder et al. later clarified that their method might be considered as giving the lowest possible probability of no real effect for a criterion that allowed the observed result to be considered significant [12].

Table 1
Diagnostic testing analogy for interpreting $P \leqslant 0.05$

	No effect	Genuine effect	Total	
$P \leqslant 0.05$	45 (5%)	50 (50%)	95	Proportion of "significant" findings that are false positives = 45/95 = 47 %
$P > 0.05$	855 (95%)	50 (50 %)	905	
Total	900 (100%) Specificity (proportion of true positives correctly identified by the test) = 95%	100 (100%) Sensitivity (proportion of true negatives correctly identified by the test) = 50%	1000	

The table shows the expected results of 1000 tests, only 100 of which are tests of a genuine effect, assuming that the power to detect a genuine effect is 50% at the 5% significance level.

The diagnostic testing analogy also assumes that the effect is either null or else precisely the value at which power is calculated. To use the language of Bayesian statistics, this "prior distribution" for the effect is a very strange one, because actually if we believe in an odds ratio of 1.0 or 2.0, for example, then we surely entertain the possibility that it might be 1.5 or 2.5, or other ratios in between [13]. A Bayesian calculation, which I will reproduce below, assuming a continuum of possible effect sizes, and showing that Lindley's paradox still applies at high powers, is presented by Spiegelhalter et al. [14pp130–133]. They conclude that "the pragmatic interpretation of *P*-values strongly depends on sample size" [14p136], in other words, on power—but now (because of Lindley's paradox) in the opposite sense to Wacholder et al.

4. Non-Bayesian interpretation of *P*-values

Whatever the direction of the association, power is an odd thing for our interpretation of P to depend on, at least from a traditional, non-Bayesian statistical viewpoint. Power is about our uncertainty over what P will turn out to be—once we have our data this uncertainty evaporates. Interpreting power after the results have been obtained is a practice which, although common, has been criticized [15,16]. Power is more useful at the planning stage of a study, when it helps ensure a narrow confidence interval and a good chance of finding evidence for an effect. In fact, the post-study power (using the *observed* effect instead of the smallest clinically significant effect in the power calculation) is always the same for a given *P*-value, at least in the idealized testing situation described further below. For example, if P is exactly 0.05 then the post-study power is always 50%. This invariant post-study power suggests that the pre-study power should not affect the interpretation of the *P*-value.

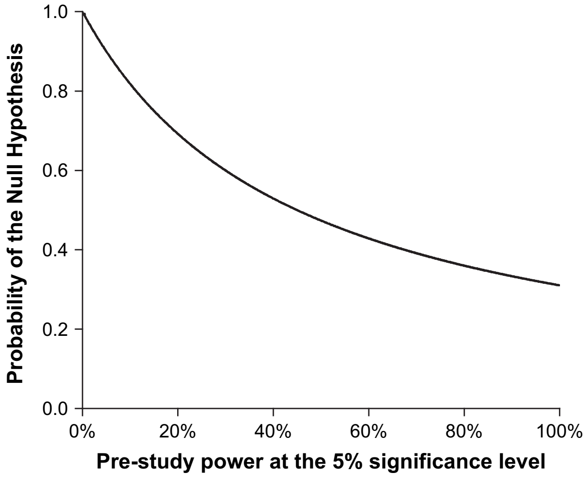

Fig. 1. Relationship between power and the probability that the null hypothesis is true given that $P \leqslant 0.05$, according to the diagnostic testing analogy, and assuming that 90% of all null hypotheses tested are true.

The idea of simply recording whether $P \leqslant \alpha$ was central to the frequentist approach to statistics of Neyman and Pearson, who argued that we should follow rules for making inferences which mean making few mistakes in the long run (α is the proportion of times in the long run where we conclude there is an effect where none is present) [17]. However, Neyman and Pearson's argument in terms of long-run error probabilities has been shown to be circular [18pp103–106], and contemporary medical statistics textbooks teach us that the precise P-value is a measure of the strength of evidence against the null hypothesis [7p72]. (Note, although, that P does not tell us about the possible size of the effect, for which we need a confidence interval [19].)

The tests we use to obtain P are often equivalent to tests based on an intuitive measure of the strength of evidence called the maximum likelihood ratio [7pp309–314], and Wilks showed in 1938 that the maximum likelihood ratio has the same approximate distribution under the null hypothesis, that is, a given ratio leads to the same P-value, whatever the power [20]. This is another strong hint that power is not relevant to the interpretation of P. Of course, a Bayesian researcher is more likely to be persuaded by a Bayesian calculation, which I now present.

5. Bayesian calculation

The Bayesian calculation can be thought of as a more elaborate version of the diagnostic testing argument, but now based on the evidence of the exact P-value, rather than the observation that $P \leqslant \alpha$, and now assuming a continuum of possible effect sizes.

5.1. Population effect

I assume there is just one parameter we are interested in—the "effect" in the population, which might be a mean difference between exposed and unexposed groups, or a log odds ratio, for example. This is measured on a numerical scale where zero represents no effect (the null hypothesis). I will assume that under the alternative hypothesis the effect can be either positive or negative, that is, the alternative is two-sided (in an epidemiological study this means looking for both harmful and protective effects of the exposure). The units of the numerical scale are arbitrary—I will re-scale them so that an effect of 1 (or -1) is the smallest effect I would describe as clinically significant, in other words, I measure any effect as a number of clinically significant units.

5.2. Data

In a quantitative research study, the population effect is estimated from sample data. I will make the simplifying assumption that the sample estimate has a normal sampling distribution, centered on the population effect, with known standard error σ. This means that the 95% confidence interval for the population effect is the sample estimate $\pm 1.96\sigma$. The sample estimate divided by its standard error gives a z-statistic from which a precise P-value can be obtained [7pp61–63].

5.3. Prior distribution

In a Bayesian framework, the prior distribution describes my beliefs about population effects before collecting data [14pp139–140]. Figure 2 shows the form of the prior distribution that I will consider. The graph shows how densely distributed my prior belief is over different effects, and the distinctive spike at zero represents that fact that the null hypothesis—an effect of precisely zero—is considered to be a distinct possibility *a priori* [21,22]. This favoring in

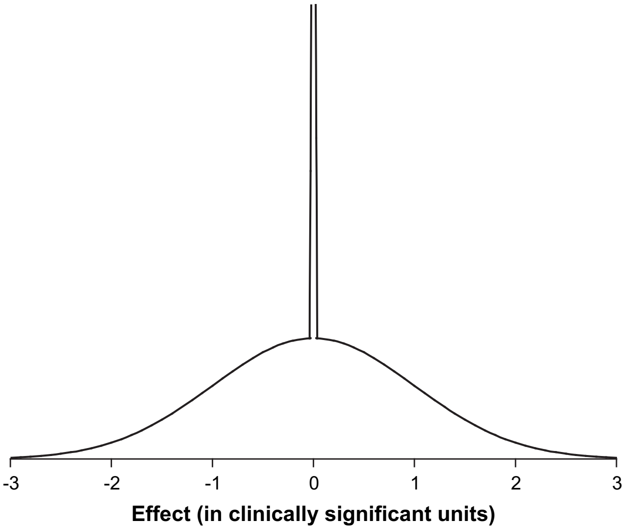

Fig. 2. Form of the prior distribution for the population effect. The graph shows prior probability density (the area under the curve between any two points on the horizontal axis is the prior probability that the effect lies within that interval).

particular of the hypothesis of no effect—this inclination to believe in the simplest theory—is an application of Occam's razor [23], and has been described as a form of "rational cynicism" [3]. Diamond and Forrester may have been among the first to use this kind of prior distribution in the clinical literature [6], although they do not provide explicit details of the form and parameters of their prior, which follows Jeffreys [24]. The approach differs from that of Brophy and Joseph, who assumed nothing special about the null hypothesis [25]. In epidemiological research it has been suggested that the prior probability of the null hypothesis, which I will denote π_0, is as large as 0.9—a 90% chance of no effect at all [3]—but here I also give results for a more optimistic $\pi_0 = 0.5$.

Over non-zero effects (the alternative hypothesis) many people would consider larger and larger effects to be increasingly less plausible *a priori*, which suggests that we use a prior distribution tailing off to zero in either direction (indeed, it turns out to be a mathematical impossibility to have the prior distribution continue at a uniform level up to infinitely large positive and negative effects). Like others before me [14p130,22], I have chosen a normal distribution symmetric around zero. I will use the letter τ to denote the standard deviation of this normal distribution. As an indication of how sensitive my results are to the choice of τ, I report results for two different values: $\tau = 1$ and $\tau = 2$. To understand what τ means, consider that if the alternative hypothesis is true, roughly 95% of our prior belief is concentrated on effects in the range -2τ to 2τ, on a scale where an effect greater than one is clinically significant. Wakefield warns against setting τ too high, to avoid the possibility that we end up believing in the null hypothesis just because the observed effect is smaller than anticipated by the prior. However, he gives an example where τ is around 0.5 (i.e., half the clinically significant value) [22], which seems unreasonably prejudiced against the possibility of a clinically significant effect in the population.

5.4. Pre-study power

I assume that pre-study power is calculated at the 5% significance level, using the pre-study definition of clinical

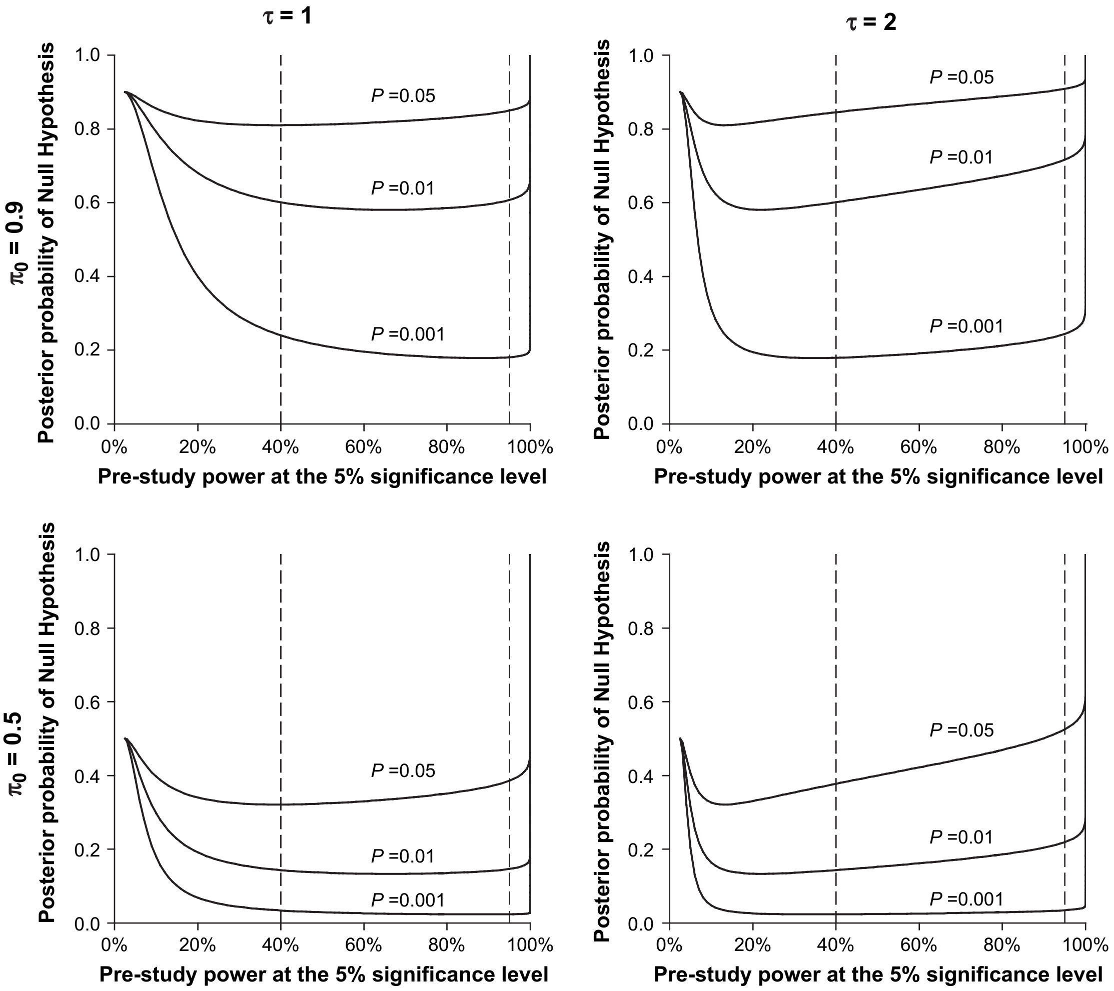

Fig. 3. Posterior probability of the null hypothesis plotted against pre-study power, for $P = 0.05$, 0.01, and 0.001. π_0 is the prior probability of the null hypothesis; τ is the standard deviation of the prior distribution over the alternative hypothesis. If the alternative hypothesis is true, 95% of prior belief is concentrated on effects in the range -1.96τ to 1.96τ, where an effect greater than 1 is clinically significant.

significance. By "power" I mean the probability that the two-sided P-value is less than 0.05, with a positive estimate of the effect, given that the population effect is just clinically significant. This can potentially range from 2.5% to 100%. Most studies are expected to achieve at least 80% power, but literature reviews have suggested that much published research fails this standard [26]. It is certainly unusual for power to exceed 95%, perhaps not only because there are ethical objections to recruiting more subjects than necessary [27], but also because the required sample size is prohibitive. We typically see results of studies with a pre-study power between 40% and 95% (for a given study design, the sample size varies more than four-fold over this range of powers).

5.5. *Posterior probability of the null hypothesis*

Given the above it is possible to calculate the degree to which we believe in a null effect given the data. In a Bayesian framework this is called the posterior probability of the null hypothesis (PPNH), and it is calculated using Bayes' theorem [14p57] (details of all calculations are in the online supplement, available on the journal's website at www.elsevier.com). By varying the standard error of the sample estimate, σ, one can also show how the PPNH varies with pre-study power.

Figure 3 shows PPNH plotted against power, for different values of P and different choices of prior distribution. Each PPNH curve has a characteristic flattened U-shape, with a wide "operating range" of power over which the curve is reasonably level, the exact shape and orientation being sensitive to the choice of prior distribution. (A Microsoft Excel application for drawing the graph for any values of τ, π_0, and P is provided as an online extra, available on the journal's website at www.jclinepi.com.)

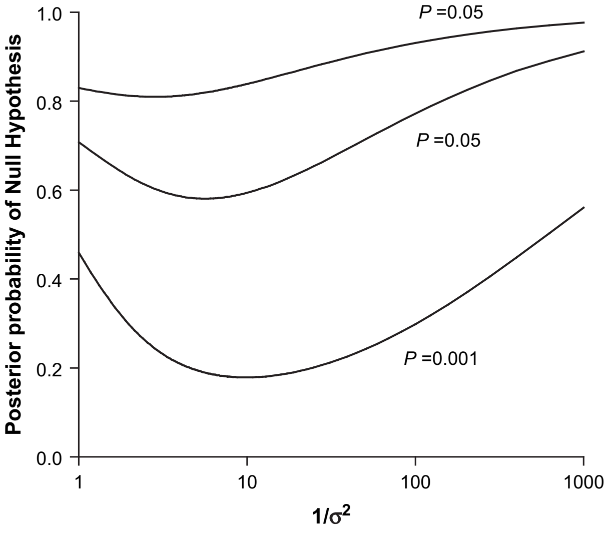

Fig. 4. Posterior probability of the null hypothesis plotted against $1/\sigma^2$, for $P = 0.05$, 0.01, 0.001, assuming a prior probability of the null hypothesis of 90%, and a standard deviation of the prior distribution over the alternative hypothesis of 1. $1/\sigma^2$ is proportional to the sample size.

6. *Discussion of the Bayesian solution*

The flattened shape of the curves shows that under reasonable assumptions, in a Bayesian framework, a low P-value such as $P = 0.001$ supports the hypothesis of a non-zero effect to virtually the same degree whether the pre-study power is 40% or 90%, or anywhere in between.

A non-zero effect should not be confused with a clinically significant effect. For a given P-value a clinically significant effect becomes increasingly *less* likely with increasing power, because much smaller effects can be detected. This is particularly important to remember if the P-value is large but the power is low. You should always look at the confidence interval to evaluate the possible magnitude of the effect.

Spiegelhalter et al. plot a similar graph to my Fig. 3 [14p133] but with an apparently different conclusion. The reason lies in their choice of horizontal axis, which shows τ^2/σ^2 on a log scale, ranging from 1 to 1000 (for given τ, τ^2/σ^2 will be proportional to the sample size). In Fig. 4, as an example, I have re-drawn Fig. 3 for $\tau = 1$, $\pi_0 = 0.9$, with $P = 0.05$, 0.01, and 0.001, using this same axis. Spiegelhalter et al. conclude that the interpretation of P "strongly depends" on sample size [14p136]. However, their axis gives undue emphasis to very high powers: by the mid-point of their graph, where $\tau^2/\sigma^2 = 30$, the power is already 99.98% assuming $\tau = 1$. The relationship with power, as illustrated in Fig. 3, presents quite a different picture.

Even outside the "operating range" of powers in Fig. 3, it is arguable how much we should adjust our interpretation of a given P-value. The increase in the PPNH at very high powers has been called paradoxical [10]. At low power the PPNH is raised because the estimated effect is larger than the range anticipated in the prior, but this prior range is a judgment call—a sensitivity analysis can always find a value of τ with which the estimate is consistent, causing the PPNH to bottom out.

Note finally, and importantly, how high the PPNH can still be given a low P-value—if $P = 0.05$ and $\pi_0 = 0.9$, the probability of the null hypothesis is still at least 81%—a lot greater than the figure of 5% which some mistakenly assume [28]. This lack of persuasiveness of $P = 0.05$ (whatever the power) has been pointed out before [3,6,29].

7. Conclusion

A careful Bayesian analysis with reasonable assumptions produces results remarkably in line with a more simple, non-Bayesian intuition: that the evidence against the null hypothesis provided by a precise P-value does not depend on power. If you are interested in the potential size of the effect then you should also consult the confidence interval.

Acknowledgments

The author's work is part of a research program funded by the Department of Health's Policy Research Program. This article arose out of conversations with Prof. John Thompson at the University of Leicester.

Appendix

Supplementary material

Supplementary material can be found, in the online version, at 10.1016/j.jclinepi.2009.02.004.

References

[1] 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.

[2] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78.

[3] Sterne JAC, Davey Smith G. Sifting the evidence—what's wrong with significance tests? BMJ 2001;322:226–31.

[4] Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 1976;34:585–612.

[5] Browner WS, Newman TB. Are all significant P-values created equal—the analogy between diagnostic-tests and clinical research. JAMA 1987;257:2459–63.

[6] Diamond GA, Forrester JS. Clinical trials and statistical verdicts: probable grounds for appeal. Ann Intern Med 1983;98:385–94.

[7] Kirkwood BR, Sterne JAC. Essential medical statistics. Oxford, UK: Blackwell; 2003.

[8] Royall RM. The effect of sample size on the meaning of significance tests. Am Stat 1986;40:313–5.

[9] Senn SJ. Statistical issues in drug development. Chichester, UK: Wiley; 1997.

[10] Lindley DV. A statistical paradox. Biometrika 1957;44:187–92.

[11] Bartlett MS. A comment on D.V. Lindley's statistical paradox. Biometrika 1957;44:533–4.

[12] Wacholder S, Chanock S, Garcia-Closas M, Katki HA, El Ghormli L, Rothman N. Re: Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Nat Cancer Inst 2004;96:1722.

[13] Thomas DC, Clayton DG. Betting odds and genetic associations. J Natl Cancer Inst 2004;96:421–3.

[14] Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Chichester, UK: Wiley; 2004.

[15] 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;121:200–6.

[16] Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 2001;55:19–24.

[17] Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc A 1933;231:289–337.

[18] Hacking I. Logic of statistical inference. Cambridge, UK: Cambridge University Press; 1965.

[19] Gardner MJ, Altman DG. Confidence intervals rather than P-values—estimation rather than hypothesis testing. BMJ 1986;292:746–50.

[20] Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann Math Stat 1938;9:60–2.

[21] Hughes MD. Reporting Bayesian analyses of clinical-trials. Stat Med 1993;12:1651–63.

[22] Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Gen 2007;81:208–27.

[23] Wears RL, Lewis RJ. Statistical models and Occam's razor. Acad Emerg Med 1999;6:93–4.

[24] Jeffreys H. Theory of probability. Oxford, UK: Oxford University Press; 1961.

[25] Brophy JM, Joseph L. Placing trials in context using Bayesian analysis: GUSTO revisited by Reverend Bayes. JAMA 1995;273:871–5.

[26] Moher D, Dulberg CS, Wells GA. Statistical power, sample-size, and their reporting in randomized controlled trials. JAMA 1994;272:122–4.

[27] Altman DG. Statistics and ethics in medical research: III How large a sample? BMJ 1980;281:1336–8.

[28] Sellke T, Bayarri MJ, Berger JO. Calibration of P-values for testing precise null hypotheses. Am Stat 2001;55:62–71.

[29] Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2:696–701.