Sifting the evidence—what's wrong with significance tests?

Jonathan A C Sterne, George Davey Smith

Department of Social Medicine, University of Bristol, Bristol BS8 2PR
Jonathan A C Sterne
senior lecturer in medical statistics
George Davey Smith
professor of clinical epidemiology
Correspondence to: J Sterne
jonathan.sterne@bristol.ac.uk

BMJ 2001;322:226–31

The findings of medical research are often met with considerable scepticism, even when they have apparently come from studies with sound methodologies that have been subjected to appropriate statistical analysis. This is perhaps particularly the case with respect to epidemiological findings that suggest that some aspect of everyday life is bad for people. Indeed, one recent popular history, the medical journalist James Le Fanu's *The Rise and Fall of Modern Medicine*, went so far as to suggest that the solution to medicine's ills would be the closure of all departments of epidemiology.[1]

One contributory factor is that the medical literature shows a strong tendency to accentuate the positive; positive outcomes are more likely to be reported than null results.[2–4] By this means alone a host of purely chance findings will be published, as by conventional reasoning examining 20 associations will produce one result that is "significant at $P = 0.05$" by chance alone. If only positive findings are published then they may be mistakenly considered to be of importance rather than being the necessary chance results produced by the application of criteria for meaningfulness based on statistical significance. As many studies contain long questionnaires collecting information on hundreds of variables, and measure a wide range of potential outcomes, several false positive findings are virtually guaranteed. The high volume and often contradictory nature[5] of medical research findings, however, is not only because of publication bias. A more fundamental problem is the widespread misunderstanding of the nature of statistical significance.

In this paper we consider how the practice of significance testing emerged; an arbitrary division of results as "significant" or "non-significant" (according to the commonly used threshold of $P = 0.05$) was not the intention of the founders of statistical inference. P values need to be much smaller than 0.05 before they can be considered to provide strong evidence against the null hypothesis; this implies that more powerful studies are needed. Reporting of medical research should continue to move from the idea that results are significant or non-significant to the interpretation of findings in the context of the type of study and other available evidence. Editors of medical journals are in an excellent position to encourage such changes, and we conclude with proposed guidelines for reporting and interpretation.

Summary points

P values, or significance levels, measure the strength of the evidence against the null hypothesis; the smaller the P value, the stronger the evidence against the null hypothesis

An arbitrary division of results, into "significant" or "non-significant" according to the P value, was not the intention of the founders of statistical inference

A P value of 0.05 need not provide strong evidence against the null hypothesis, but it is reasonable to say that $P < 0.001$ does. In the results sections of papers the precise P value should be presented, without reference to arbitrary thresholds

Results of medical research should not be reported as "significant" or "non-significant" but should be interpreted in the context of the type of study and other available evidence. Bias or confounding should always be considered for findings with low P values

To stop the discrediting of medical research by chance findings we need more powerful studies

P values and significance testing—a brief history

The confusion that exists in today's practice of hypothesis testing dates back to a controversy that raged between the founders of statistical inference more than 60 years ago.[6–8] The idea of significance testing was introduced by R A Fisher. Suppose we want to evaluate whether a new drug improves survival after myocardial infarction. We study a group of patients treated with the new drug and a comparable group treated with placebo and find that mortality in the group treated with the new drug is half that in the group treated with placebo. This is encouraging but could it be a chance finding? We examine the question by calculating a P value: the probability of getting at least a twofold difference in survival rates if the drug really has no effect on survival.

Fisher saw the P value as an index measuring the strength of evidence against the null hypothesis (in our example, the hypothesis that the drug does not affect survival rates). He advocated $P < 0.05$ (5% significance) as a standard level for concluding that there is evidence against the hypothesis tested, though not as an absolute rule. "If P is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05…."[9] Importantly, Fisher argued strongly that interpretation of the P value was ultimately for the researcher. For example, a P value of around 0.05 might lead to neither belief nor disbelief in the null hypothesis but to a decision to perform another experiment.

Dislike of the subjective interpretation inherent in this approach led Neyman and Pearson to propose what they called "hypothesis tests," which were designed to replace the subjective view of the strength of evidence against the null hypothesis provided by the

P value with an objective, decision based approach to the results of experiments.[10] Neyman and Pearson argued that there were two types of error that could be made in interpreting the results of an experiment (table 1). Fisher's approach concentrates on the type I error: the probability of rejecting the null hypothesis (that the treatment has no effect) if it is in fact true. Neyman and Pearson were also concerned about the type II error: the probability of accepting the null hypothesis (and thus failing to use the new treatment) when in fact it is false (the treatment works). By fixing, in advance, the rates of type I and type II error, the number of mistakes made over many different experiments would be limited. These ideas will be familiar to anyone who has performed a power calculation to find the number of participants needed in a clinical trial; in such calculations we aim to ensure that the study is large enough to allow both type I and type II error rates to be small.

In the words of Neyman and Pearson "no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis. But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong."[10]

Thus, in the Neyman-Pearson approach we decide on a decision rule for interpreting the results of our experiment in advance, and the result of our analysis is simply the rejection or acceptance of the null hypothesis. In contrast with Fisher's more subjective view—Fisher strongly disagreed with the Neyman-Pearson approach[11]—we make no attempt to interpret the P value to assess the strength of evidence against the null hypothesis in an individual study.

To use the Neyman-Pearson approach we must specify a precise alternative hypothesis. In other words it is not enough to say that the treatment works, we have to say by how much the treatment works—for example, that our drug reduces mortality by 60%. The researcher is free to change the decision rule by specifying the alternative hypothesis and type I and type II error rates, but this must be done in advance of the experiment. Unfortunately researchers find it difficult to live up to these ideals. With the exception of the primary question in randomised trials, they rarely have in mind a precise value of the treatment effect under the alternative hypothesis before they carry out their studies or specify their analyses. Instead, only the easy part of Neyman and Pearson's approach—that the null hypothesis can be rejected if $P<0.05$ (type I error rate 5%)—has been widely adopted. This has led to the misleading impression that the Neyman-Pearson approach is similar to Fisher's.

In practice, and partly because of the requirements of regulatory bodies and medical journals,[12] the use of statistics in medicine became dominated by a division of results into significant or not significant, with little or no consideration of the type II error rate. Two common and potentially serious consequences of this are that possibly clinically important differences observed in small studies are denoted as non-significant and ignored, while all significant findings are assumed to result from real treatment effects.

Table 1 Possible errors in interpretation of experiments, according to the Neyman-Pearson approach to hypothesis testing. Error rates are proportion of times that type I and type II errors occur in the long run

	The truth	
Result of experiment	**Null hypothesis true (treatment doesn't work)**	**Null hypothesis false (treatment works)**
Reject null hypothesis	Type I error rate	Power=1–type II error rate
Accept null hypothesis		Type II error rate

These problems, noted long ago[13] and many times since,[14–17] led to the successful campaign to augment the presentation of statistical analyses by presenting confidence intervals in addition to, or in place of, P values.[18–20] By focusing on the results of the individual comparison, confidence intervals should move us away from a mechanistic accept-reject dichotomy. For small studies, they may remind us that our results are consistent with both the null hypothesis and an important beneficial, or harmful, treatment effect (and often both). For P values of around 0.05 they also emphasise the possibility of the effect being much smaller, or larger, than estimated. 95% Confidence intervals, however, implicitly use the 5% cut off, and this still leads to confusion in their interpretation if they are used simply as a means of assessing significance (according to whether the confidence interval includes the null value) rather than to look at a plausible range for the magnitude of the population difference. We suggest that medical researchers should stop thinking of 5% significance ($P<0.05$) as having any particular importance. One way to encourage this would be to adopt a different standard confidence level.

Misinterpretation of P values and significance tests

Unfortunately, P values are still commonly misunderstood. The most common misinterpretation is that the P value is the probability that the null hypothesis is true, so that a significant result means that the null hypothesis is very unlikely to be true. Making two plausible assumptions, we show the misleading nature of this interpretation.

Firstly, we will assume that the proportion of null hypotheses that are in fact false is 10%—that is, 90% of hypotheses tested are incorrect. This is consistent with the epidemiological literature: by 1985 nearly 300 risk factors for coronary heart disease had been identified,

Table 2 Number of times we accept and reject null hypothesis, under plausible assumptions regarding conduct of medical research (adapted from Oakes[25])

Result of experiment	Null hypothesis true (treatment doesn't work)	Null hypothesis false (treatment works)	Total
Accept null hypothesis	855	50	905
Reject null hypothesis	45	50	95
Total	900	100	1000

and it is unlikely that more than a small fraction of these actually increase the risk of the disease.[21] Our second assumption is that because studies are often too small the average power (=1−type II error rate) of studies reported in medical literature is 50%. This is consistent with published surveys of the size of trials.[22–24]

Suppose now that we test hypotheses in 1000 studies and reject the null hypothesis if $P<0.05$. The first assumption means that in 100 studies the null hypothesis is in fact false. Because the type II error rate is 50% (second assumption) we reject the null hypothesis in 50 of these 100 studies. For the 900 studies in which the null hypothesis is true (that is, there is no treatment effect) we use 5% significance levels and so reject the null hypothesis in 45 (see table 2, adapted from Oakes[25]).

Of the 95 studies that result in a significant (that is, $P<0.05$) result, 45 (47%) are true null hypotheses and so are "false alarms"; we have rejected the null hypothesis when we shouldn't have done so. There is a direct analogy with tests used to screen populations for diseases: if the disease (the false null hypothesis) is rare then the specificity of screening tests must be high to prevent the true cases of disease identified by the test from being swamped by large numbers of false positive tests from most of the population who do not have the disease.[26] The "positive predictive value" of a significant ($P<0.05$) statistical test can actually be low—in the above case around 50%. The common mistake is to assume that the positive predictive value is 95% because the significance level is set at 0.05.

The ideas illustrated in table 2 are similar in spirit to the bayesian approach to statistical inference, in which we start with an a priori belief about the probability of different possible values for the treatment effect and modify this belief in the light of the data. Bayesian arguments have been used to show that the usual $P<0.05$ threshold need not constitute strong evidence against the null hypothesis.[27 28] Various authors over the years have proposed that more widespread use of bayesian statistics would prevent the mistaken interpretation of $P<0.05$ as showing that the null hypothesis is unlikely to be true or even act as a panacea that would dramatically improve the quality of medical research.[26 29–32] Differences between the dominant ("classic" or "frequentist") and bayesian approaches to statistical inference are summarised in box 1.

How significant is significance?

When the principles of statistical inference were established, during the early decades of the 20th century, science was a far smaller scale enterprise than it is today. In the days when perhaps only a few hundred statistical hypotheses were being tested each year, and when calculations had to be done laboriously with mechanical hand calculators (as in Fisher's photograph), it seemed reasonable that a 5% false positive rate would screen out most of the random errors. With many thousands of journals publishing a myriad hypothesis tests each year and the ease of use of statistical software it is likely that the proportion of tested hypotheses that are meaningful (in the sense that the effect is large enough to be of interest) has decreased, leading to a finding of $P<0.05$ having low predictive value for the appropriate rejection of the null hypothesis.

It is often perfectly possible to increase the power of studies by increasing either the sample size or the precision of the measurements. Table 3 shows the pre-

Box 1: Comparison of frequentist and bayesian approaches to statistical inference

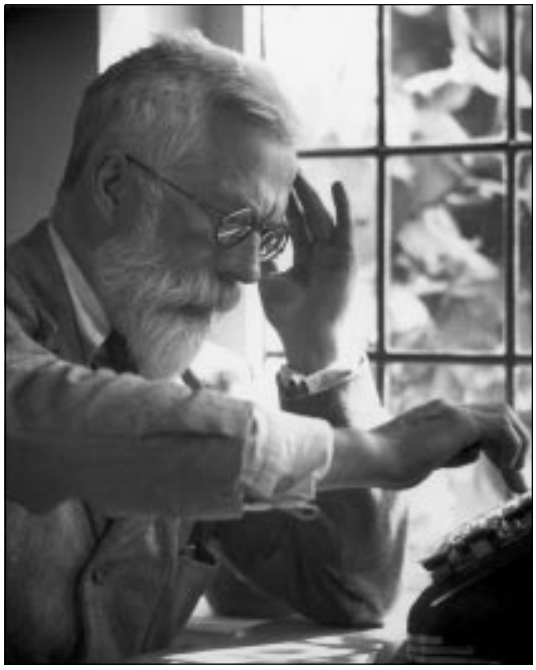
Let us assume that we want to evaluate whether a new drug improves one year survival after myocardial infarction by using data from a placebo controlled trial. We do this by estimating the risk ratio—the risk of death in patients treated with the new drug divided by the risk of death in the control group. If the risk ratio is 0.5 then the new drug reduces the risk of death by 50%. If the risk ratio is 1 then the drug has no effect.

Frequentist statistics	Bayesian statistics
Like Mulder and Scully in *The X-Files*, frequentist statisticians believe that "the truth is out there." We use the data to make inferences about the true (but unknown) population value of the risk ratio	Bayesians take a subjective approach. We start with our prior opinion about the risk ratio, expressed as a probability distribution. We use the data to modify that opinion (we derive the posterior probability distribution for the risk ratio based on both the data and the prior distribution)
The 95% confidence interval gives us a plausible range of values for the population risk ratio; 95% of the times we derive such a range it will contain the true (but unknown) population value	A 95% credible interval is one that has a 95% chance of containing the population risk ratio
The P value is the probability of getting a risk ratio at least as far from the null value of 1 as the one found in our study	The posterior distribution can be used to derive direct probability statements about the risk ratio—for example, the probability that the drug increases the risk of death

If our prior opinion about the risk ratio is vague (we consider a wide range of values to be equally likely) then the results of a frequentist analysis are similar to the results of a bayesian analysis; both are based on what statisticians call the likelihood for the data:

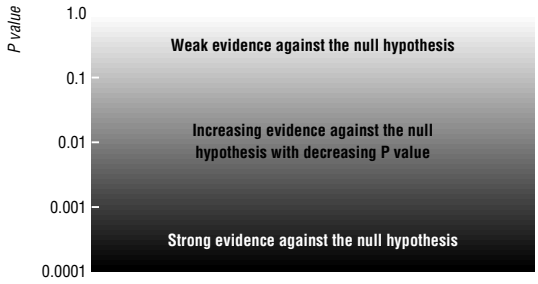
- The 95% confidence interval is the same as the 95% credible interval, except that the latter has the meaning often incorrectly ascribed to a confidence interval;
- The (one sided) P value is the same as the bayesian posterior probability that the drug increases the risk of death (assuming that we found a protective effect of the drug).

The two approaches, however, will give different results if our prior opinion is not vague, relative to the amount of information contained in the data.



A BARRINGTON-BROWN

R A Fisher, the founder of statistical inference, working on a mechanical calculator



Suggested interpretation of P values from published medical research

dictive value of different P value thresholds under different assumptions about both the power of studies and the proportion of meaningful hypotheses. For any choice of P value, the proportion of "significant" results that are false positives is greatly reduced as power increases. Table 3 suggests that unless we are very pessimistic about the proportion of meaningful hypotheses, it is reasonable to regard P values less than 0.001 as providing strong evidence against the null hypothesis.

One argument against changing the strength of evidence regarded as conclusively showing that the null hypothesis is false is that studies would have to be far bigger. Surprisingly, this is not true. For illustrative purposes it can be shown, by using standard power calculations, that the maximum amount by which a study size would have to be increased is by a factor of only 1.75 for a move from $P<0.05$ to $P<0.01$ and 2.82 from $P<0.05$ to $P<0.001$. It is also possible, and generally preferable, to increase power by decreasing measurement error rather than by increasing sample size.[33] Thus by doing fewer but more powerful studies it is perfectly possible to stop the discrediting of medical research. The need for large, statistically precise studies has been emphasised for many years by Richard Peto and colleagues.[34] The practice of medical research will not be improved, however, if we simply substitute one arbitrary P value threshold (0.05) with another one (0.001).

Interpreting P values: opinions, decisions, and the role of external evidence

In many cases published medical research requires no firm decision: it contributes incrementally to an existing body of knowledge. In the results sections of papers the precise P value should be presented, without reference to some arbitrary threshold. In communicating the individual contribution of a single study we suggest the P value should be interpreted as illustrated in the figure. P values in the "grey area" provide some, but not conclusive, evidence against the null hypothesis.

It is rare that studies examine issues about which nothing is already known. Increasing recognition of this is reflected in the growth of formal methods of research synthesis,[35] including the presentation of updated meta-analyses in the discussion section of original research papers.[36] Here the prior evidence is simply the results of previous studies of the same issue. Other forms of evidence are, of course, admissible: findings from domains as different as animal studies and tissue cultures on the one hand and secular trends and ecological differences in human disease rates on the other will all influence a final decision as to how to act in the light of study findings.[37]

In many ways the general public is ahead of medical researchers in its interpretation of new "evidence." The reaction to "lifestyle scares" is usually cynicism, which, for many reasons, may well be rational.[38] Popular reactions can be seen to reflect a subconscious bayesianism in which the prior belief is that what medical researchers, and particularly epidemiologists, produce is gobbledegook. In medical research the periodic calls for a wholesale switch to the use of bayesian statistical inference have been largely ignored. A major reason is that prior belief can be difficult to quantify. How much weight should be given to a particular constellation of biological evidence as against the concordance of a study finding with international differences in disease rates, for example?

Table 3 Proportion of false positive significant results with three different criteria for significance

Power of study (proportion (%) of time we reject null hypothesis if it is false)	Percentage of "significant" results that are false positives		
	P=0.05	P=0.01	P=0.001
80% of ideas correct (null hypothesis false)			
20	5.9	1.2	0.10
50	2.4	0.5	0.05
80	1.5	0.3	0.03
50% of ideas correct (null hypothesis false)			
20	20.0	4.8	0.50
50	9.1	2.0	0.20
80	5.9	1.2	0.10
10% of ideas correct (null hypothesis false)			
20	69.2	31.0	4.30
50	47.4*	15.3	1.80
80	36.0	10.1	1.10
1% of ideas correct (null hypothesis false)			
20	96.1	83.2	33.10
50	90.8	66.4	16.50
80	86.1	55.3	11.00

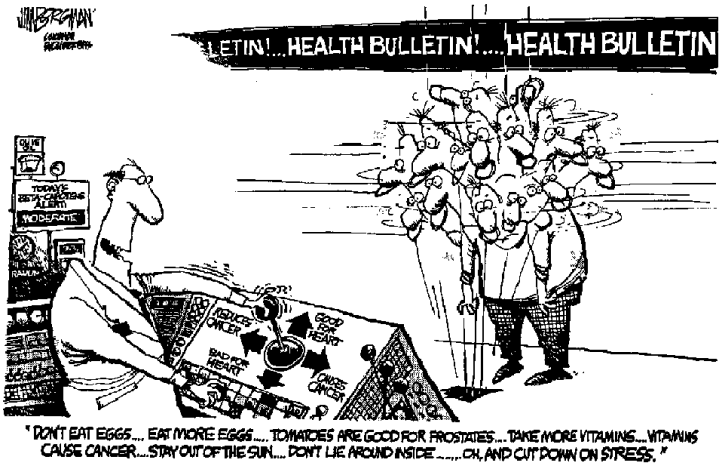
*Corresponds to assumptions in table 2.

Similarly, the predictive value of $P<0.05$ for a meaningful hypothesis is easy to calculate on the basis of an assumed proportion of "meaningful" hypotheses in the study domain, but in reality it will be impossible to know what this proportion is. Tables 2 and 3 are, unfortunately, for illustration only. If we try to avoid the problem of quantification of prior evidence by making our prior opinion extremely uncertain then the results of a bayesian analysis become similar to those in a standard analysis. On the other hand, it would be reasonable to interpret $P=0.008$ for the main effect in a clinical trial differently to the same P value for one of many findings from an observational study on the basis that the proportion of meaningful hypotheses tested is probably higher in the former case and that bias and confounding are less likely.

What is to be done?

There are three ways of reducing the degree to which we are being misled by the current practice of significance testing. Firstly, table 3 shows that $P<0.05$ cannot be regarded as providing conclusive, or even strong, evidence against the null hypothesis. Secondly, it is clear that increasing the proportion of tested hypotheses that are meaningful would also reduce the degree to which we are being misled. Unfortunately this is difficult to implement; the notion that the formulation of prior hypotheses is a guarantor against being misled is itself misleading. If we do 100 randomised trials of useless treatments, each testing only one hypothesis and performing only one statistical hypothesis test, all "significant" results will be spurious. Furthermore, it is impossible to police claims that reported associations were examined because of existing hypotheses. This has been satirised by Philip Cole, who has announced that he has, via a computer algorithm, generated every possible hypothesis in epidemiology so that all statistical tests are now of a priori hypotheses.[39] Thirdly, the most important need is not to change statistical paradigms but to improve the quality of studies by increasing sample size and precision of measurement.

While there is no simple or single solution, it is possible to reduce the risk of being misled by the results of hypothesis tests. This lies partly in the hands of journal editors. Important changes in the presentation of statistical analyses were achieved after guidelines insisting on presentation of confidence intervals were introduced during the 1980s. A similar shift in the presentation of hypothesis tests is now required. We suggest that journal editors require that authors of research reports follow the guidelines outlined in box 2.



Box 2: Suggested guidelines for the reporting of results of statistical analyses in medical journals

1. The description of differences as statistically significant is not acceptable
2. Confidence intervals for the main results should always be included, but 90% rather than 95% levels should be used. Confidence intervals should not be used as a surrogate means of examining significance at the conventional 5% level. Interpretation of confidence intervals should focus on the implications (clinical importance) of the range of values in the interval
3. When there is a meaningful null hypothesis, the strength of evidence against it should be indexed by the P value. The smaller the P value, the stronger is the evidence
4. While it is impossible to reduce substantially the amount of data dredging that is carried out, authors should take a very sceptical view of subgroup analyses in clinical trials and observational studies. The strength of the evidence for interaction—that effects really differ between subgroups—should always be presented. Claims made on the basis of subgroup findings should be even more tempered than claims made about main effects
5. In observational studies it should be remembered that considerations of confounding and bias are at least as important as the issues discussed in this paper[40]

We are grateful to Professor S Goodman, Dr M Hills, and Dr K Abrams for helpful comments on previous versions of the manuscript; this does not imply their endorsement of our views. Bristol is the lead centre of the MRC Health Sevices Research Collaboration.

Funding: None.

Competing interests: Both authors have misused the word significance in the past and may have overestimated the strength of the evidence for their hypotheses.

1 Le Fanu J. *The rise and fall of modern medicine*. New York: Little, Brown, 1999.
2 Berlin JA, Begg CB, Louis TA. An assessment of publication bias using a sample of published clinical trials. *J Am Stat Assoc* 1989;84:381-92.
3 Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;337:867-72.
4 Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *JAMA* 1992;263:374-8.
5 Mayes LC, Horwitz RI, Feinstein AR. A collection of 56 topics with contradictory results in case-control research. *Int J Epidemiol* 1988;17:680-5.
6 Goodman SN. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 1993;137:485-96.
7 Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *J Am Stat Assoc* 1993;88:1242-9.
8 Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med* 1999;130:995-1004.
9 Fisher RA. *Statistical methods for research workers*. London: Oliver and Boyd, 1950:80.
10 Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. *Philos Trans Roy Soc A* 1933;231:289-337.
11 Fisher RA. *Statistical methods and scientific inference*. London: Collins Macmillan, 1973.
12 Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. *J Clin Epidemiol* 1998;51:355-60.
13 Berkson J. Tests of significance considered as evidence. *J Am Stat Assoc* 1942;37:325-35.
14 Rozeboom WW. The fallacy of the null-hypothesis significance test. *Psychol Bull* 1960;57:416-28.
15 Freiman JA, Chalmers TC, Smith HJ, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of

the randomized control trial. Survey of 71 "negative" trials. *N Engl J Med* 1978;299:690-4.
16 Cox DR. Statistical significance tests. *Br J Clin Pharmacol* 1982;14:325-31.
17 Rothman KJ. Significance questing. *Ann Intern Med* 1986;105:445-7.
18 Altman DG, Gore SM, Gardner MJ, Pocock, SJ. Statistical guidelines for contributors to medical journals. *BMJ* 1983;286:1489-93.
19 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ* 1986;292:746-50.
20 Gardner MJ, Altman DG. *Statistics with confidence. Confidence intervals and statistical guidelines*. London: BMJ Publishing, 1989.
21 Hopkins PN, Williams RR. Identification and relative weight of cardiovascular risk factors. *Cardiol Clin* 1986;4:3-31.
22 Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial. In: Bailar JC, Mosteller F, eds. *Medical uses of statistics*. Boston, Ma: NEJM Books, 1992:357-73.
23 Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994;272:122-4.
24 Mulward S, Gøtzsche PC. Sample size of randomized double-blind trials 1976-1991. *Dan Med Bull* 1996;43:96-8.
25 Oakes M. *Statistical inference*. Chichester: Wiley, 1986.
26 Browner WS, Newman TB. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 1987;257:2459-63.
27 Edwards W, Lindman H, Savage LJ. Bayesian statistical inference for psychological research. *Psychol Rev* 1963;70:193-242.
28 Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of *P* values and evidence. *J Am Stat Assoc* 1987;82:112-22.
29 Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. *BMJ* 1996;313:603-7.
30 Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. *JAMA* 1995;273:871-5.
31 Burton PR, Gurrin LC, Campbell MJ. Clinical significance not statistical significance: a simple Bayesian alternative to p values. *J Epidemiol Community Health* 1998;52:318-23.
32 Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* 1999;130:1005-13.
33 Phillips AN, Davey Smith G. The design of prospective epidemiological studies: more subjects or better measurements? *J Clin Epidemiol* 1993;46:1203-11.
34 Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;3:409-22.
35 Egger M, Davey Smith G. Meta-analysis. Potentials and promise. *BMJ* 1997;315:1371-4.
36 Danesh J, Whincup P, Walker M, Lennon L, Thomson A, Appleby P, et al. Chlamydia pneumoniae IgG titres and coronary heart disease: prospective study and meta-analysis. *BMJ* 2000;321:208-13.
37 Morris JN. *The uses of epidemiology*. Edinburgh: Churchill-Livingstone, 1975.
38 Davey Smith G. Reflections on the limits to epidemiology. *J Clin Epidemiol* (in press).
39 Cole P. The hypothesis generating machine. *Epidemiology* 1993;4:271-3.
40 Davey Smith G, Phillips AN. Confounding in epidemiological studies: why "independent" effects may not be all they seem. *BMJ* 1992;305:757-9.

(Accepted 9 November 2000)

Another comment on the role of statistical methods

D R Cox

Nuffield College, Oxford OX1 1NF
D R Cox
professor
david.cox@nuf.ox.ac.uk

The cartoons in Sterne and Davey Smith's paper describe implicitly a double threat to progress. Firstly, there is the bombardment of an apparently nervous and litigious public with ill based stories. This leads on to the undermining of meticulous studies that may indeed point towards improved health. Statistical methods, sensibly and modestly used, are both some protection against false alarms and, more importantly, an aid, via principles of study design and analysis, to well founded investigations and ultimately to enhanced health.

To comment here on detailed statistical issues would be out of place. While guidelines too rigidly enforced are potentially dangerous, the thoughtful recommendations of box 2 are consistent with mainstream statistical thinking. That is, design to minimise bias is crucial and the estimation of magnitudes of effects, relative risks, or whatever, is central and best done by limits of error, confidence or posterior limits, or estimates and standard errors. Statistical significance testing has a limited role, usually as a supplement to estimates. Quantitative notions of personalistic probability may have some place, especially perhaps in the planning stage of an investigation, but seem out of place in the general reporting of conclusions.

The authors' castigation of the search for subgroup effects in largely null studies is indeed thoroughly justified. All reports of large effects confined, however, to Aston Villa supporters over the age of 75 and living south of Birmingham should go into the wastepaper basket, however great the interest in that particular subgroup, or, in less extreme cases, put into the pile of topics for future independent investigation. More might be made of a limited and preplanned search for effect modifiers, what in statistical jargon rather misleadingly tends to be called interaction. Even the most carefully planned and implemented randomised controlled trial with full compliance estimates only an average effect across the population of patients giving informed consent. The basis for extending the conclusions to different populations and to individual patients often lies primarily in scientific understanding of the mode of action of the treatments concerned but is reinforced by some check of the stability of any effect found, even if such checks are relatively insensitive.

All these issues are essentially ones of public education about the nature of scientific inquiry and the uncertainties involved. As the authors note, modern statistical thinking owes much to the statistician and geneticist R A Fisher, in particular for two books.[1 2] In the second, the same year that Karl Popper introduced the hypotheticodeductive method, Fisher wrote "Every experiment may be said to exist only to give the facts the chance of disproving the null hypothesis." On the 25th anniversary of the publication of the first book, Fisher's friend F Yates wrote an assessment of its impact, in particular criticising Fisher for his emphasis on significance testing.[3] In one form or another this criticism has been repeated many times since. To distinguish several types of hypothesis that might be tested it helps to understand the issues.[4] In the research laboratory it may be possible to set up an experiment for which outcome can be predicted if the understanding of an underlying process is correct. The key issue is then consistency with that prediction. On the other hand, in many epidemiological studies and randomised controlled trials, with rare exceptions (mobile phones and brain tumours, for instance), there may be no reason for expecting the effect to be null. The issue tends more to be whether the direction of an effect has been reasonably firmly established and whether the magnitude of any effect is such as to make it of public health or clinical importance.

1 Fisher RA. *Statistical methods for research workers*. 1st ed. Edinburgh: Oliver and Boyd, 1925. Reprinted by Oxford University Press.
2 Fisher RA. *Design of experiments*. 1st ed. Edinburgh: Oliver and Boyd, 1935. Reprinted by Oxford University Press.
3 Yates F. The 25th anniversary of statistical methods for research workers. *J Am Stat Assoc* 1950;46:19-34.
4 Cox DR. Statistical significance tests. *Br J Clin Pharmacol* 1982;14:325-31.