The Death of Statistical Significance Testing

Kenneth J. Rothman

February 2, 2009

Physician's Health Study NEJM 1988; 318:262-264

"Furthermore, among the six categories of deaths from vascular causes, there was no significant excess in the aspirin group within any single category that would counterbalance the deficit in fatal myocardial infarction (5 in the aspirin group and 18 in the placebo group)."

	Aspirin	Placebo	RR	95% CI	P-value
Acute MI	5	18	0.25	0.11—0.56	0.006
Stroke	6	2	3.00	0.75—12.0	0.16
Ischemic Heart Disease	9	8	1.08	0.42—2.8	0.81
Sudden Death	13	9	1.49	0.65—3.4	0.40
Other Cardiovascular	10	6	1.79	0.67—4.76	0.31
Other Cerebrovascular	1	1	1.00	0.06—16.0	1.00
Total Cardiovascular	44	44	0.99	0.65—1.5	0.99

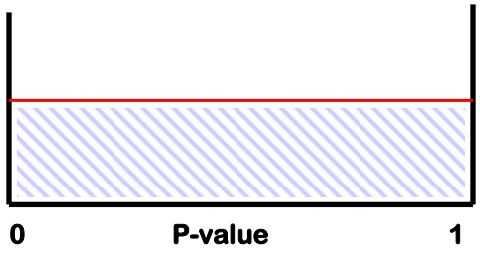
Fundamental Problems of Statistical Significance Testing

1. Significance testing is based on the P-value, which is a confounded measure: it mixes effect size with precision

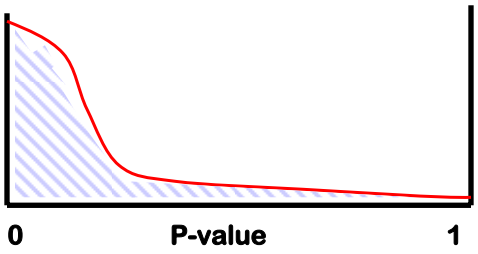
2. It is not possible to measure two things with one number

3. Significance testing reduces the quantitative P-value to a qualitative measure, yes/no

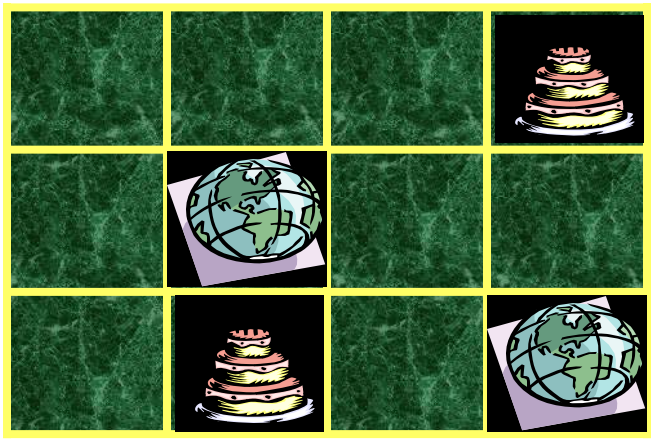
P-value



P-value



Memory Game



Probability of Winning in One Play

$$\text{Prob.} = \frac{1}{11} \times \frac{1}{9} \times \frac{1}{7} \times \frac{1}{5} \times \frac{1}{3}$$

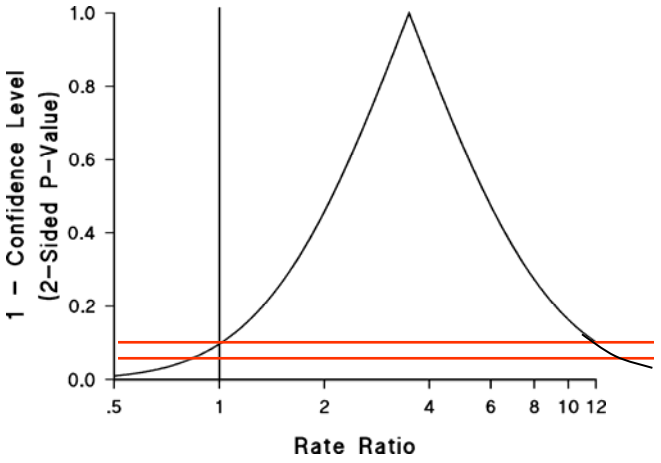
$$= 0.000096$$

Confidence Intervals

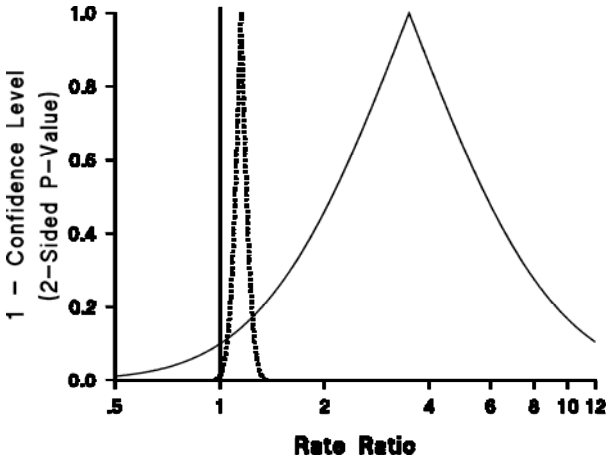
A confidence interval is a range of hypothesized parameter values for which the p-values testing those hypotheses are greater than a specified level.

If we measure RR, for example, the 90% CI for a RR is the range of RR values for which the corresponding p-values would be greater than 0.1.

Confidence Interval or P-value Function



Confidence Interval or P-value Function

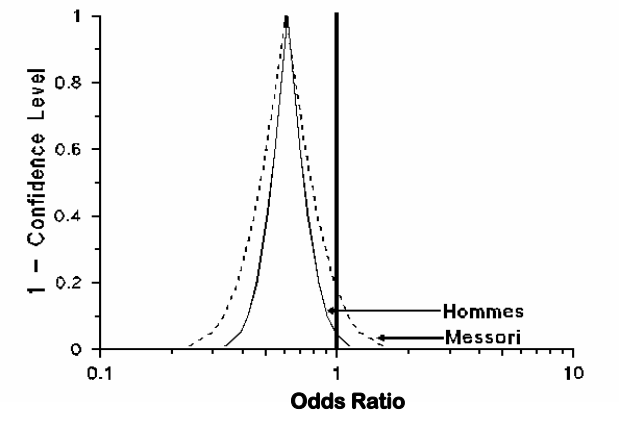


Calculation Errors in Meta-Analysis

(Messori et al., Ann Intern Med 1993;118:77-78)

The recent paper by Hommes and colleagues reports a meta-analysis of six randomized trials comparing subcutaneous heparin with continuous intravenous heparin for the initial treatment of deep vein thrombosis....The result of our calculation was an odds ratio of 0.61 (95% CI, 0.298 to 1.251; P > 0.05); this figure differs greatly from the value reported by Hommes and associates (odds ratio, 0.62; 95% CI, 0.39 to 0.98; P < 0.05)....Based on our recalculation of the overall odds ratio, we concluded that subcutaneous heparin is not more effective than intravenous heparin, exactly the opposite to that of Hommes and colleagues....

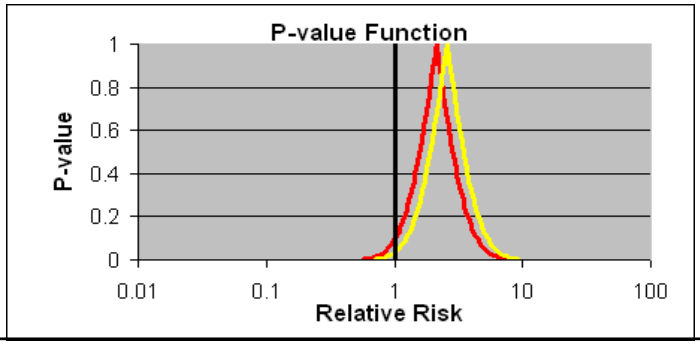
CI/P-value Functions: Hommes et al. and Messori et al.



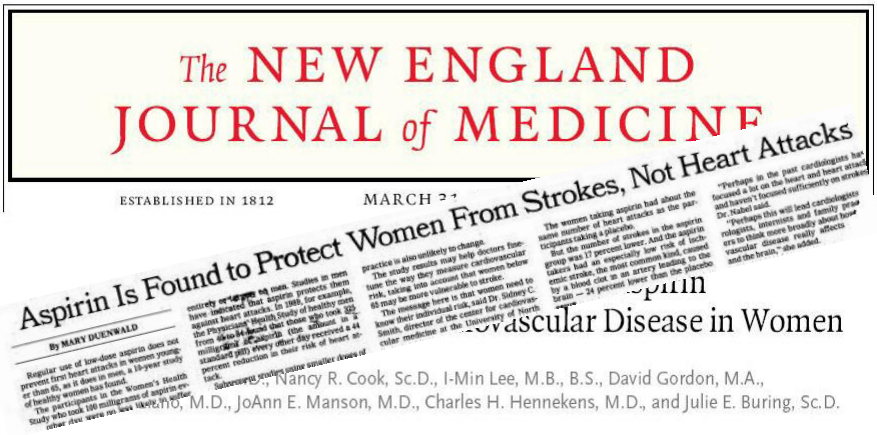
Inference by Statistical Significance

Effect in Men: **RR = 2.6 95% CI: 1.1 – 6.0**

Effect in Women: **RR = 2.1 95% CI: 0.9 – 5.0**



CI/P-value Function: Women's Health Study


The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 MARCH 31

Aspirin Is Found to Protect Women From Strokes, Not Heart Attacks

By MARY DUENWALD

Regular use of low-dose aspirin does not prevent first heart attacks in women younger than 65, as it does in men, a 10-year study of healthy women has found.

The participants in the Women's Health Study who took 100 milligrams of aspirin every other day were not less likely to ...

...spirin ...vascular Disease in Women

...Nancy R. Cook, Sc.D., I-Min Lee, M.B., B.S., David Gordon, M.A., ...M.D., JoAnn E. Manson, M.D., Charles H. Hennekens, M.D., and Julie E. Buring, Sc.D.

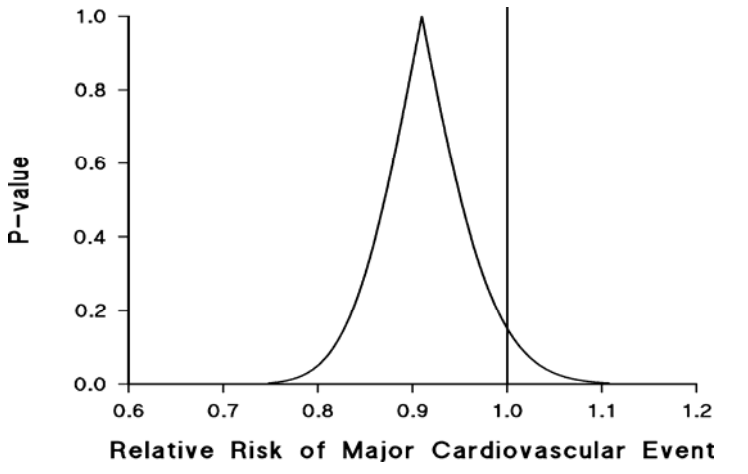
CI/P-value Function: Women's Health Study

CONCLUSIONS

In this large, primary-prevention trial among women, aspirin lowered the risk of stroke without affecting the risk of myocardial infarction or death from cardiovascular causes, leading to a nonsignificant finding with respect to the primary end point.

N ENGL J MED 352;13 WWW.NEJM.ORG MARCH 31, 2005

CI/P-value Function: Women's Health Study



CI/P-value Function: Alcohol and Cognitive Impairment

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Effects of Moderate Alcohol Consumption on Cognitive Function in Women

Meir J. Stampfer, M.D., Jae Hee Kang, Sc.D., Jennifer Chen, M.P.H., Rebecca Cherry, M.D., and Francine Grodstein, Sc.D.

ABSTRACT

BACKGROUND

The adverse effects of excess alcohol intake on cognitive function are well established, but the effect of moderate consumption is uncertain.

METHODS

Between 1995 and 2001, we evaluated cognitive function in 12,480 participants in the Nurses' Health Study who were 70 to 81 years old, with follow-up assessments in 11,102 two years later. The level of alcohol consumption was ascertained regularly beginning in 1980. We calculated multivariate-adjusted mean cognitive scores and multivariate-adjusted risks of cognitive impairment (defined as the lowest 10 percent of the

From the Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School (M.J.S., J.H.K., J.C., F.G.); and the Departments of Epidemiology (M.J.S., F.G.) and Nutrition (M.J.S.), Harvard School of Public Health — all in Boston; and Vanderbilt Children's Hospital, Nashville (R.C.).

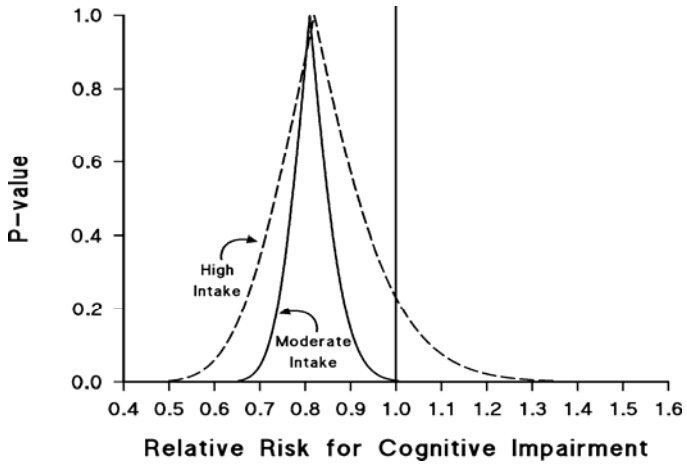
N Engl J Med 2005;352:245-53.

CI/P-value Function: Alcohol and Cognitive Impairment

RESULTS

After multivariate adjustment, moderate drinkers (those who consumed less than 15.0 g of alcohol per day [about one drink]) had better mean cognitive scores than nondrinkers. Among moderate drinkers, as compared with nondrinkers, the relative risk of impairment was 0.77 on our test of general cognition (95 percent confidence interval, 0.67 to 0.88) and 0.81 on the basis of a global cognitive score combining the results of all tests (95 percent confidence interval, 0.70 to 0.93). The results for cognitive decline were similar; for example, on our test of general cognition, the relative risk of a substantial decline in performance over a two-year period was 0.85 (95 percent confidence interval, 0.74 to 0.98) among moderate drinkers, as compared with nondrinkers. There were no significant associations between higher levels of drinking (15.0 to 30.0 g per day) and the risk of cognitive impairment or decline. There were no significant differences in risks according to the beverage (e.g., wine or beer) and no interaction with the apolipoprotein E genotype.

CI/P-value Function: Alcohol and Cognitive Impairment



Statistical Significance in the News

Boston Scientific Stent Study Flawed

By KEITH J. WINSTEIN
August 14, 2008; Page B1

A heart stent manufactured by Boston Scientific Corp. and expecting approval for U.S. sales is backed by flawed research despite the company's claims of success in a clinical trial, according to a Wall Street Journal review of the data.

Boston Scientific submitted the results of the 2006 trial to the Food and Drug Administration to gain U.S. approval for the Taxus Liberte, which already is one of the top-selling stents abroad. Coronary stents -- tiny scaffolds that prop open arteries clogged by heart disease -- are one of the most popular methods for treating heart patients, and have been implanted in more than 15 million people world-wide.

Statistical Significance in the News

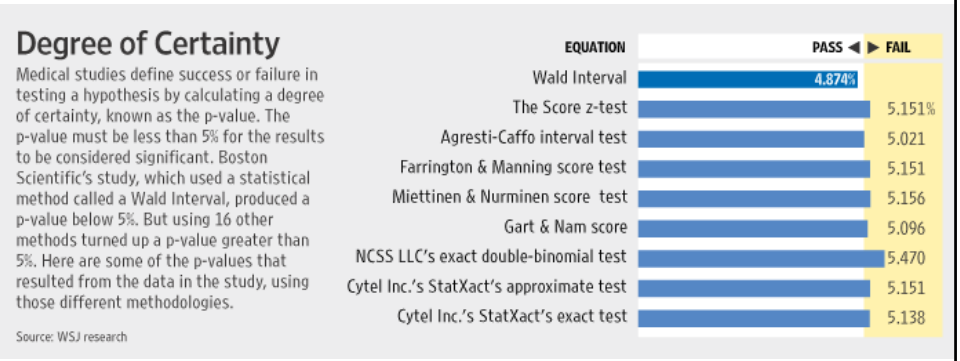
"Boston Scientific's claim was based on a flawed statistical equation that favored the Liberte stent, a Journal analysis has found. Using a number of other methods of calculation -- including 14 available in off-the-shelf software programs -- the Liberte study would have been a failure by the common standards of statistical significance in research."

Statistical Significance in the News

"Scientists generally regard studies with p-values above 5% to be failures, and medical journals typically won't publish them....

"[T]he Journal's calculations found that the Liberte study's p-value was about 5.1%. Although the difference seems small -- 0.2 of a percentage point -- it is the difference between success and failure for a product on which Boston Scientific has spent some tens of millions of dollars."

Statistical Significance in the News



Criticism of Sigificance Testing is Not New

1919: Edwin Boring criticizes early use of statistical significance testing

1957: Lancelot Hogben describes logical and practical errors in theory and teaching of statistical significance testing

1970: Morrison Henkel publish compendium entitled "The Significance Test Controversy"

Criticism of Sigificance Testing

The statistical significance test "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!"

-- *Cohen (1994)*

Criticism of Sigificance Testing

Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution.

-- *Schmidt and Hunter (1997)*

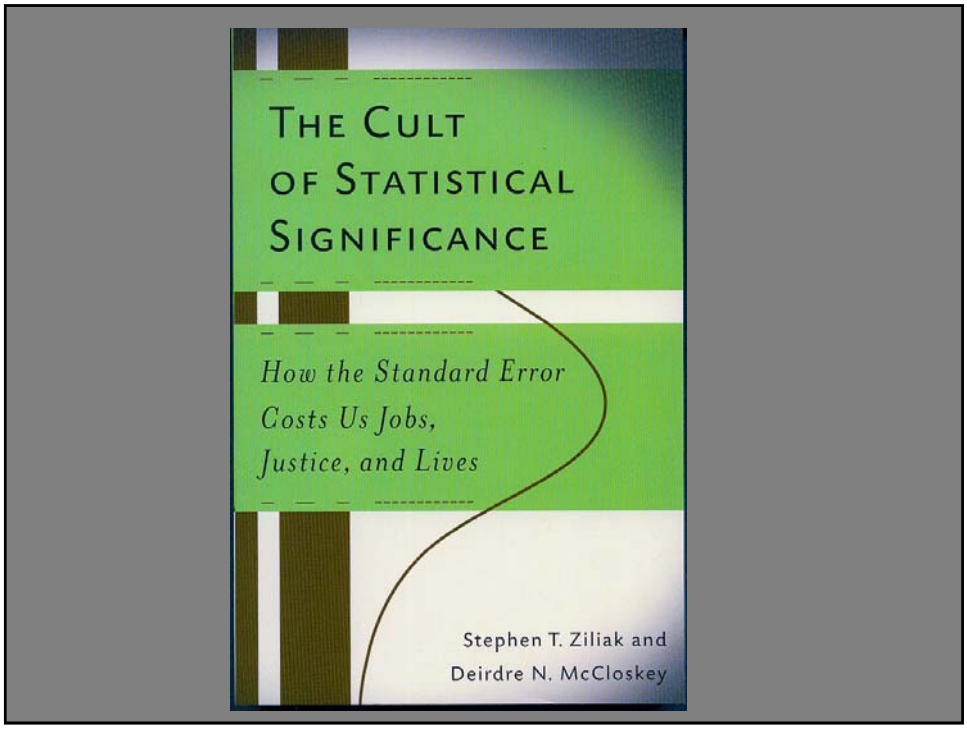
Criticism of Sigificance Testing

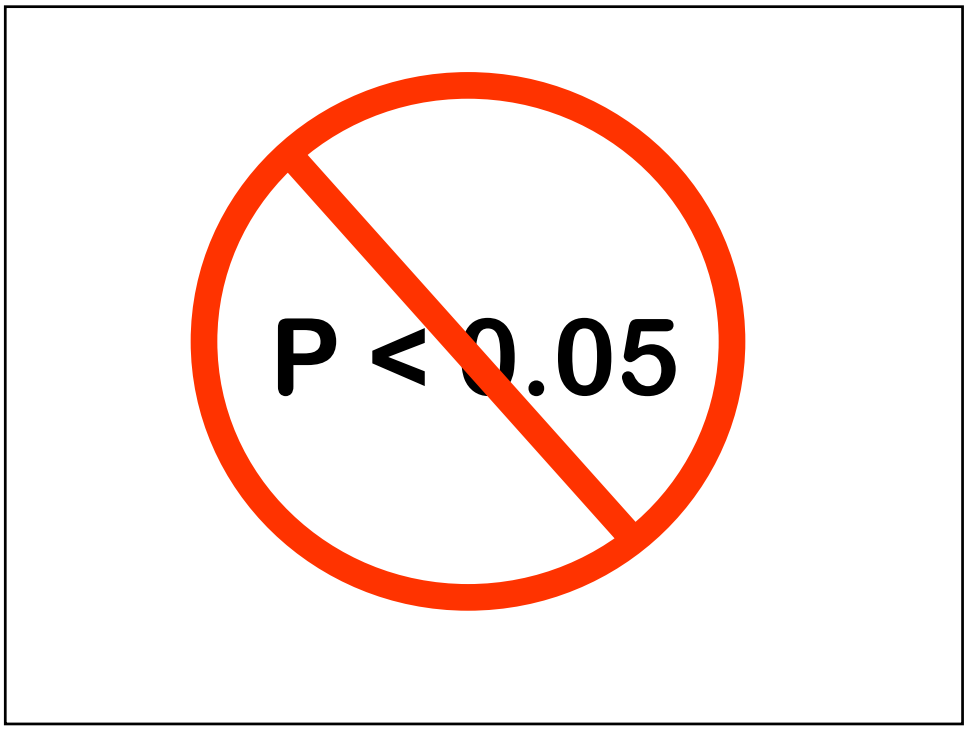
Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students... It is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism.

-- *Rozeboom (1997)*

Arguments Offered in Favor of Statistical Significance Testing

- **Tests and confidence intervals are mathematically equivalent.**
- **In the real world, decisions need to be made. Tests provide a basis for decision-making.**
- **Tradition.**
- **Need to say something about data.**


THE CULT
OF STATISTICAL
SIGNIFICANCE
How the Standard Error
Costs Us Jobs,
Justice, and Lives
Stephen T. Ziliak and
Deirdre N. McCloskey


P < 0.05