RTI (h)(s)™
Health Solutions

March 28, 2019

The End of Statistical Significance Testing

The power of **knowledge.**
The value of **understanding.**

Meet Our Team

Kenneth J. Rothman, DrPH
Distinguished Fellow, Vice President of Epidemiology Research

Heather Danysh, PhD
Research Epidemiologist

Physician's Health Study

	Aspirin N= 11,037	Placebo N=11,034	RR	95% CI
Acute MI	5	18	0.25	0.11—0.56
Stroke	6	2	3.00	0.75—12.0
Ischemic Heart Disease	9	8	1.08	0.42—2.8
Sudden Death	13	9	1.49	0.65—3.4
Other Cardiovascular	10	6	1.79	0.67—4.76
Other Cerebrovascular	1	1	1.00	0.06—16.0

"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)."

NEJM 1988; 318:262-264

Physician's Health Study

	Aspirin N= 11,037	Placebo N=11,034	RR	95% CI
Acute MI	5	18	0.25	0.11—0.56
Stroke	6	2	3.00	0.75—12.0
Ischemic Heart Disease	9	8	1.08	0.42—2.8
Sudden Death	13	9	1.49	0.65—3.4
Other Cardiovascular	10	6	1.79	0.67—4.76
Other Cerebrovascular	1	1	1.00	0.06—16.0
Total Cardiovascular	44	44	0.99	0.65—1.5

"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)."

NEJM 1988; 318:262-264

Physician's Health Study

	Aspirin N= 11,037	Placebo N=11,034	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

"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)."

NEJM 1988; 318:262-264

Serotonergic Antidepressant Use During Pregnancy and Autism

RTI (h)(s) Health Solutions

Research

JAMA | Original Investigation

Association Between Serotonergic Antidepressant Use During Pregnancy and Autism Spectrum Disorder in Children

Hilary K. Brown, PhD; Joel G. Ray, MD, MSc, FRCPC; Andrew S. Wilton, MSc; Yona Lunsky, PhD, CPsych; Tara Gomes, MHSc; Simone N. Vigod, MD, MSc, FRCPC

IMPORTANCE Previous observations of a higher risk of child autism spectrum disorder with serotonergic antidepressant exposure during pregnancy may have been confounded.

OBJECTIVE To evaluate the association between serotonergic antidepressant exposure during pregnancy and child autism spectrum disorder.

Editorial page 1533

Related article page 1553

Supplemental content

JAMA 2017;317(15):1544-1552. doi:10.1001/jama.2017.3415

Serotonergic Antidepressant Use During Pregnancy and Autism

RTI (h)(s) Health Solutions

RESULTS There were 35 906 singleton births at a mean gestational age of 38.7 weeks (50.4% were male, mean maternal age was 26.7 years, and mean duration of follow-up was 4.95 years). In the 2837 pregnancies (7.9%) exposed to antidepressants, 2.0% (95% CI, 1.6%-2.6%) of children were diagnosed with autism spectrum disorder. The incidence of autism spectrum disorder was 4.51 per 1000 person-years among children exposed to antidepressants vs 2.03 per 1000 person-years among unexposed children (between-group difference, 2.48 [95% CI, 2.33-2.62] per 1000 person-years; hazard ratio [HR], 2.16 [95% CI, 1.64-2.86]; adjusted HR, 1.59 [95% CI, 1.17-2.17]). After inverse probability of treatment weighting based on the high-dimensional propensity score, the association was not significant (HR, 1.61 [95% CI, 0.997-2.59]). The association was also not significant when exposed children were compared with unexposed siblings (incidence of autism spectrum disorder was 3.40 per 1000 person-years vs 2.05 per 1000 person-years, respectively; adjusted HR, 1.60 [95% CI, 0.69-3.74]).

CONCLUSION: "…antidepressant exposure compared with no exposure was not associated with autism spectrum disorder…."

JAMA. 2017;317(15):1544-1552. doi:10.1001/jama.2017.3415

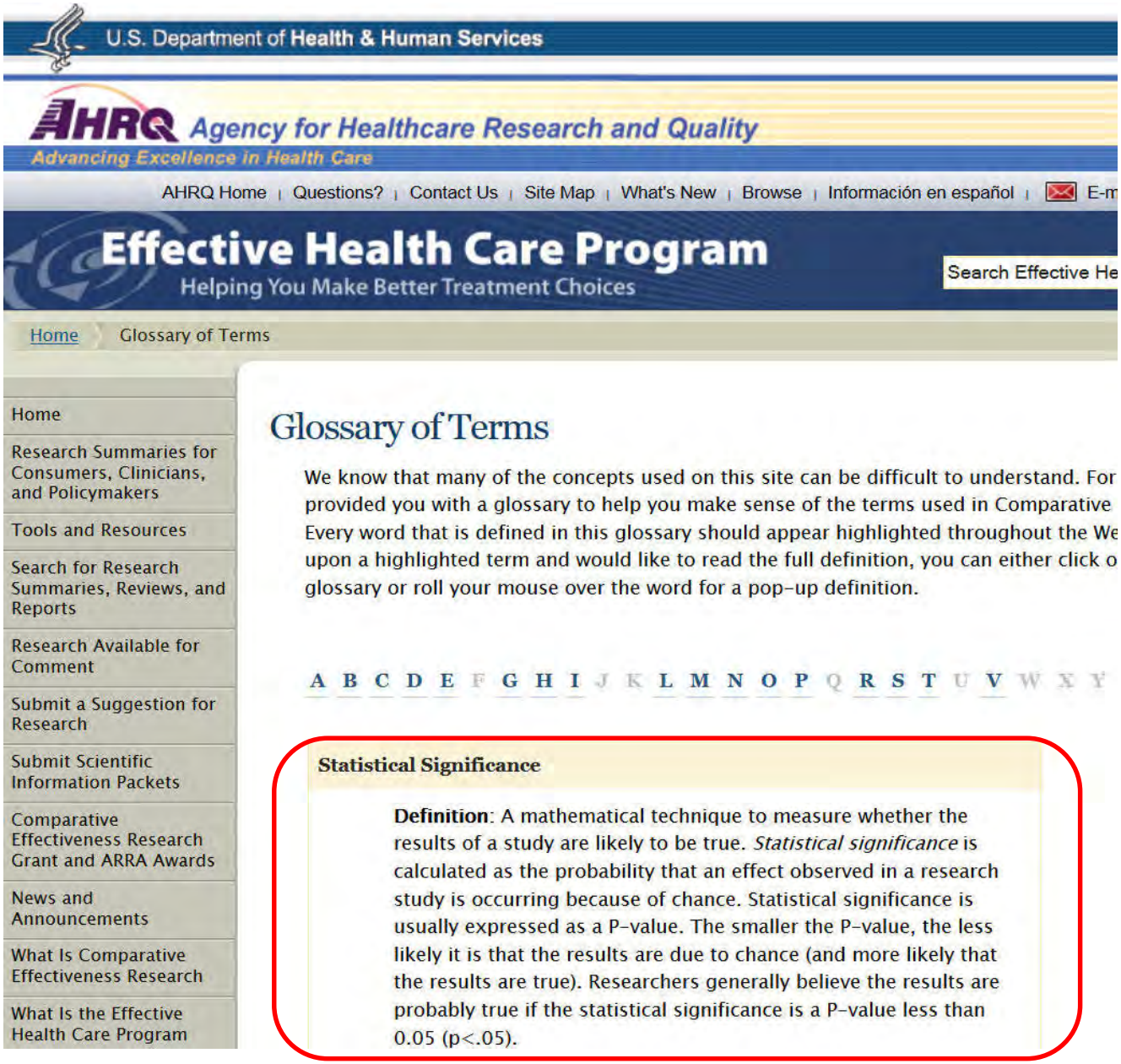
Fundamental Problems of Statistical Significance Testing

RTI (h)(s) Health Solutions

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

2. It is problematic to measure two things with one number

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



AHRQ Definition of Statistical Significance

RTI (h)(s) Health Solutions

Statistical Significance

Definition: A mathematical technique to measure whether the results of a study are likely to be true. *Statistical significance* is calculated as the probability that an effect observed in a research study is occurring because of chance. Statistical significance is usually expressed as a P-value. The smaller the P-value, the less likely it is that the results are due to chance (and more likely that the results are true). Researchers generally believe the results are probably true if the statistical significance is a P-value less than 0.05 ($p<.05$).

AHRQ Definition of Statistical Significance

Statistical Significance

Definition: A mathematical technique to measure whether the results of a study are likely to be true. *Statistical significance* is calculated as the probability that an effect observed in a research study is occurring because of chance. Statistical significance is usually expressed as a P-value. The smaller the P-value, the less likely it is that the results are due to chance (and more likely that the results are true). Researchers generally believe the results are probably true if the statistical significance is a P-value less than 0.05 ($p<.05$).

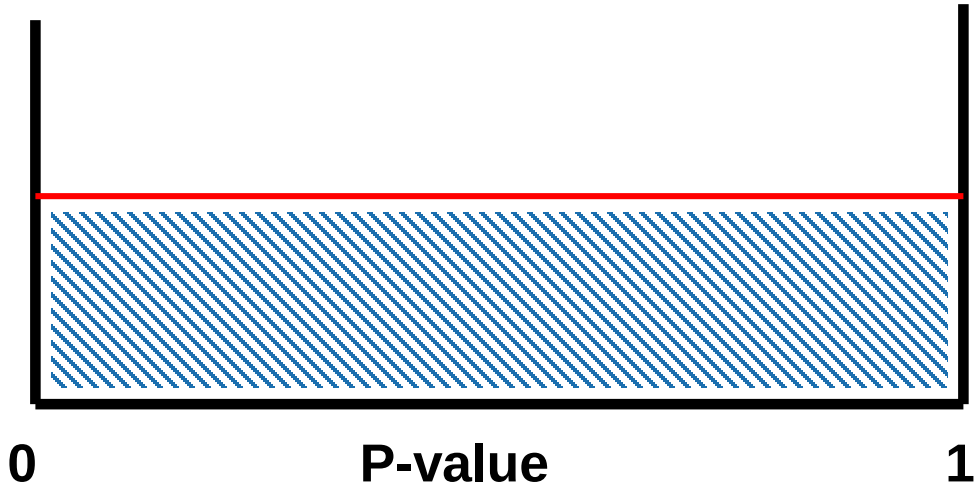
AHRQ Definition of Statistical Significance

RTI (h)(s) Health Solutions

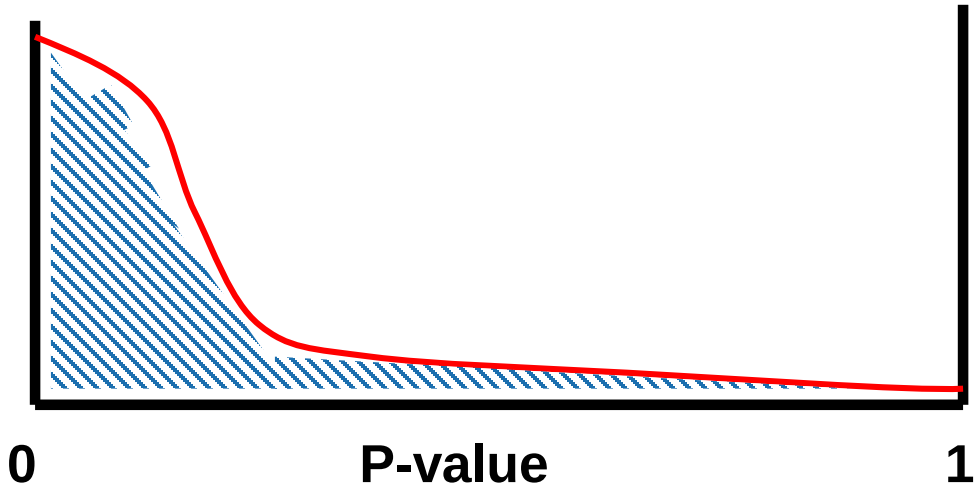
Statistical Significance

Definition: A mathematical technique to measure whether the results of a study are likely to be true. *Statistical significance* is calculated as the probability that an effect observed in a research study is occurring because of chance. Statistical significance is usually expressed as a P-value. The smaller the P-value, the less likely it is that the results are due to chance (and more likely that the results are true). Researchers generally believe the results are probably true if the statistical significance is a P-value less than 0.05 ($p<.05$).

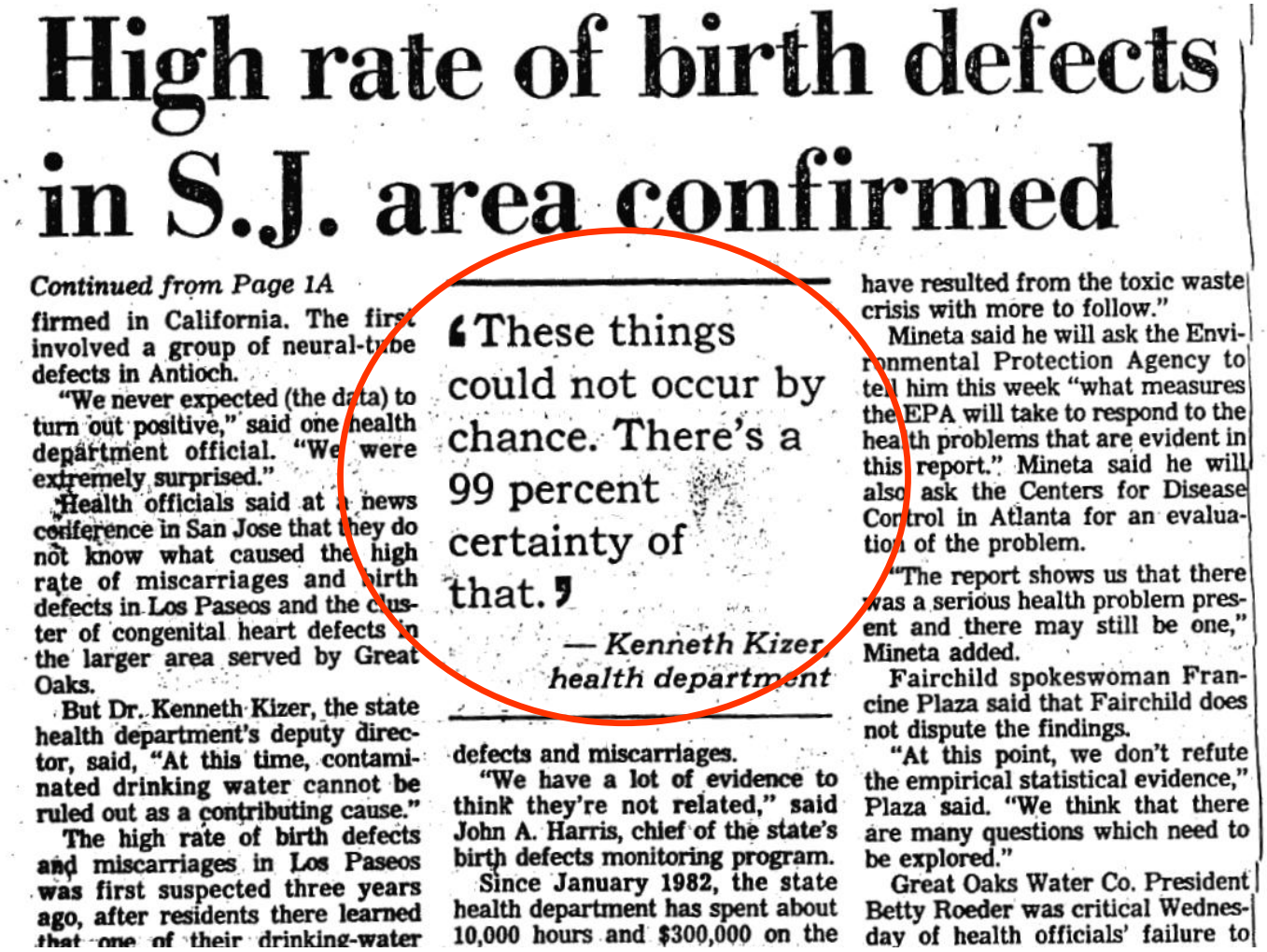
P-value



P-value



Meaning of P-value



High rate of birth defects in S.J. area confirmed

Continued from Page 1A

firmed in California. The first involved a group of neural-tube defects in Antioch.

"We never expected (the data) to turn out positive," said one health department official. "We were extremely surprised."

Health officials said at a news conference in San Jose that they do not know what caused the high rate of miscarriages and birth defects in Los Paseos and the cluster of congenital heart defects in the larger area served by Great Oaks.

But Dr. Kenneth Kizer, the state health department's deputy director, said, "At this time, contaminated drinking water cannot be ruled out as a contributing cause."

The high rate of birth defects and miscarriages in Los Paseos was first suspected three years ago, after residents there learned that one of their drinking-water

> "These things could not occur by chance. There's a 99 percent certainty of that."
>
> — *Kenneth Kizer, health department*

defects and miscarriages.

"We have a lot of evidence to think they're not related," said John A. Harris, chief of the state's birth defects monitoring program.

Since January 1982, the state health department has spent about 10,000 hours and $300,000 on the

have resulted from the toxic waste crisis with more to follow."

Mineta said he will ask the Environmental Protection Agency to tell him this week "what measures the EPA will take to respond to the health problems that are evident in this report." Mineta said he will also ask the Centers for Disease Control in Atlanta for an evaluation of the problem.

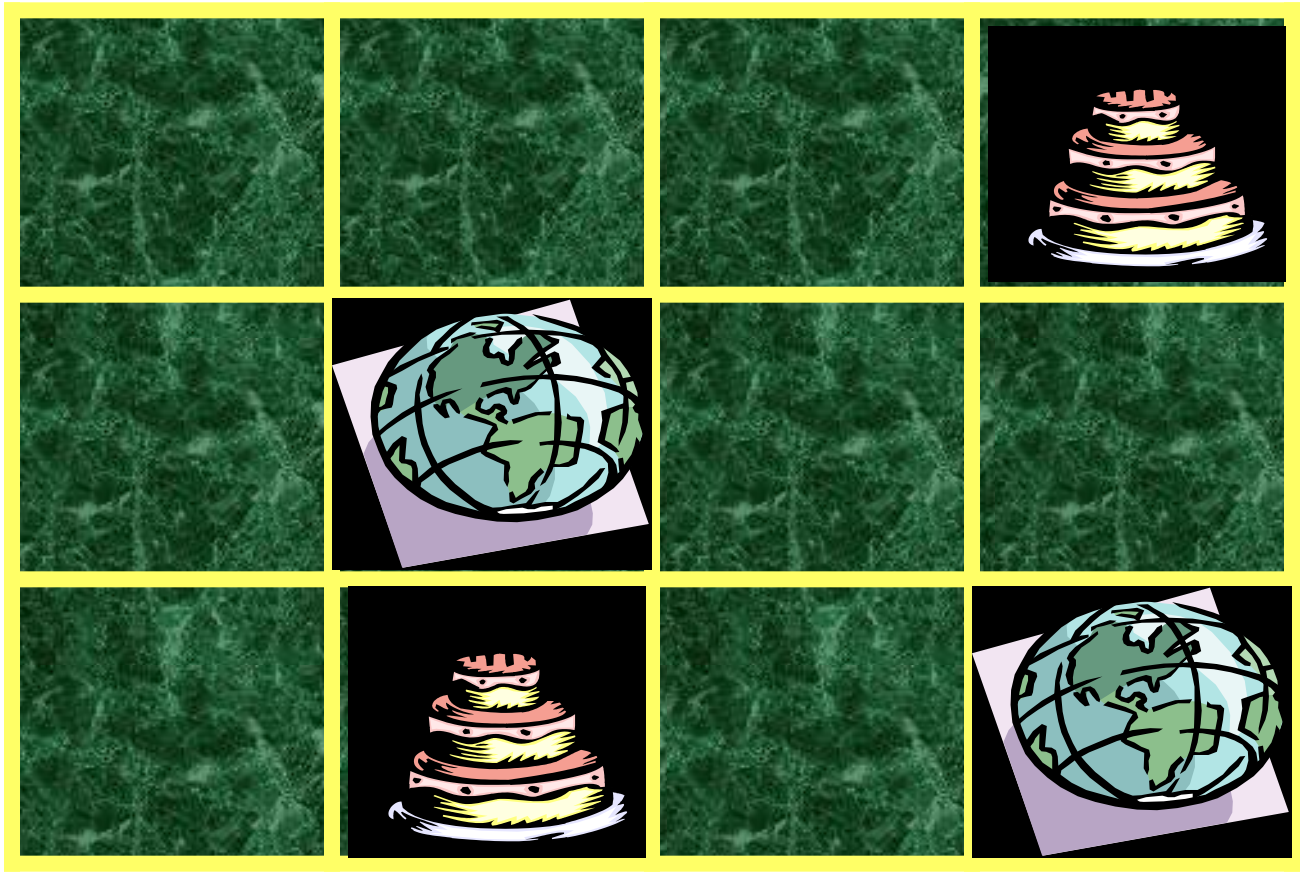
"The report shows us that there was a serious health problem present and there may still be one," Mineta added.

Fairchild spokeswoman Francine Plaza said that Fairchild does not dispute the findings.

"At this point, we don't refute the empirical statistical evidence," Plaza said. "We think that there are many questions which need to be explored."

Great Oaks Water Co. President Betty Roeder was critical Wednesday of health officials' failure to

Memory Game



Probability of Winning in One Play

$$\text{Prob.} = \frac{1}{11}$$

Probability of Winning in One Play

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

Probability of Winning in One Play

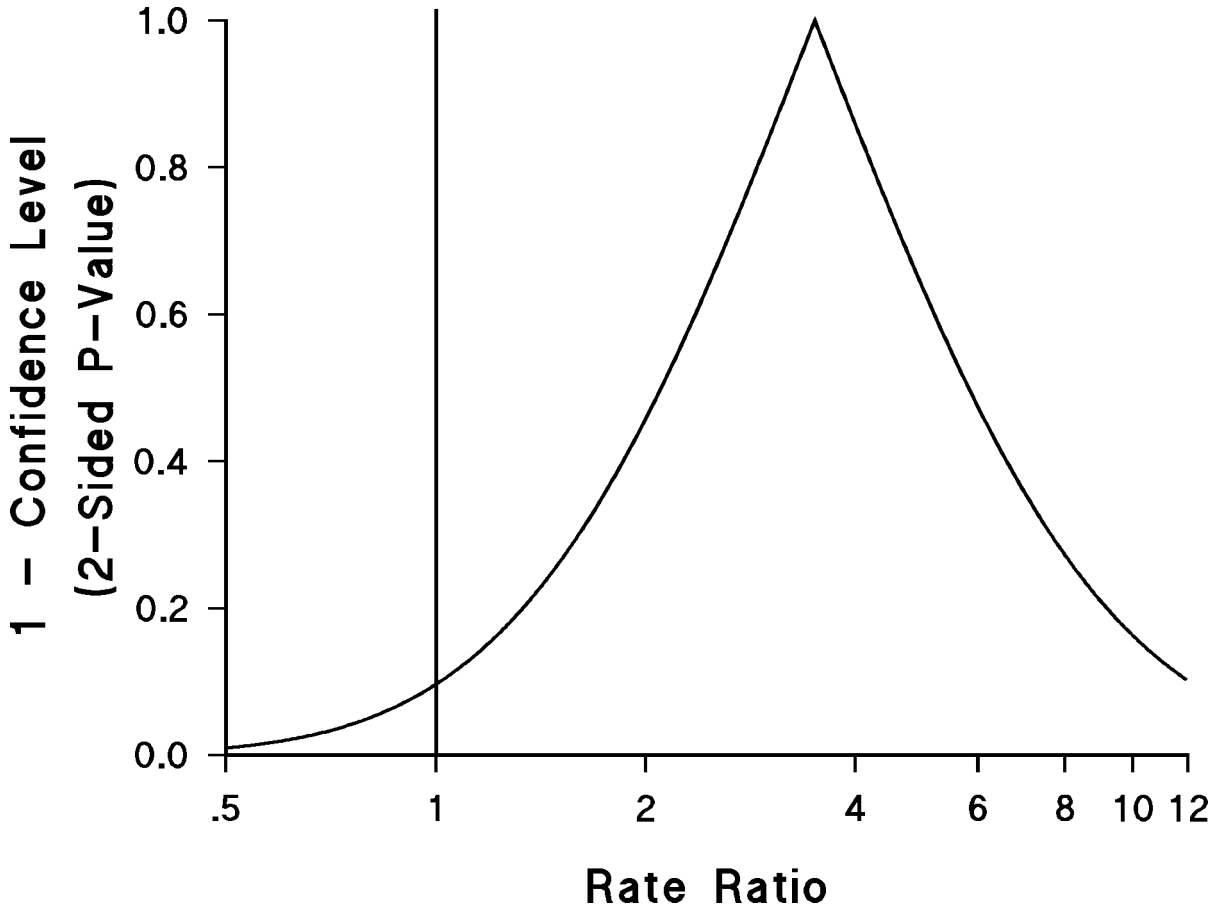
$$\text{Prob.} = \frac{1}{11} \times \frac{1}{9} \times \frac{1}{7} \times \frac{1}{5} \times \frac{1}{3}$$

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 Interval or P-value Function



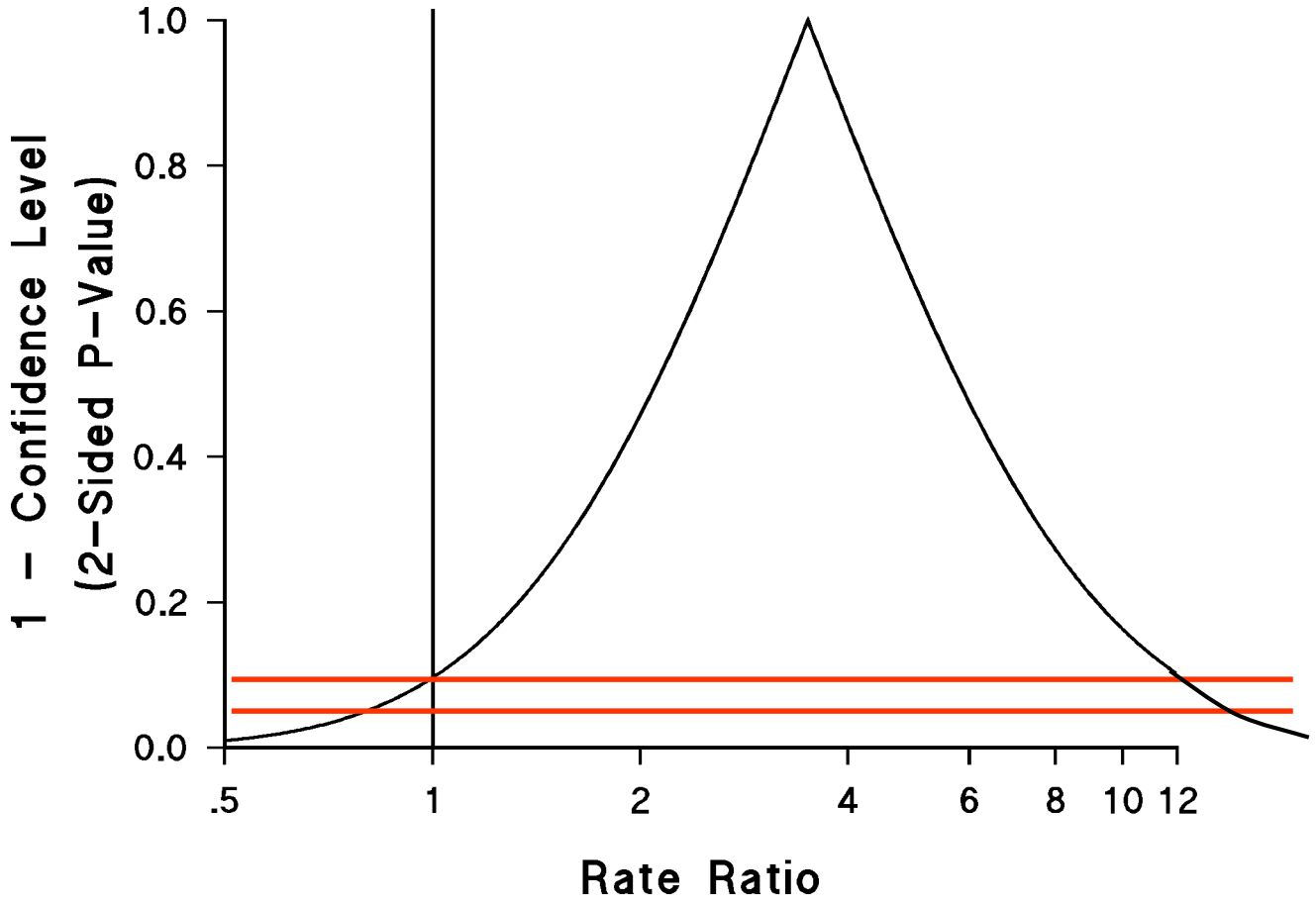
Correspondence Between P-values and Confidence Intervals

RTI (h)(s) Health Solutions

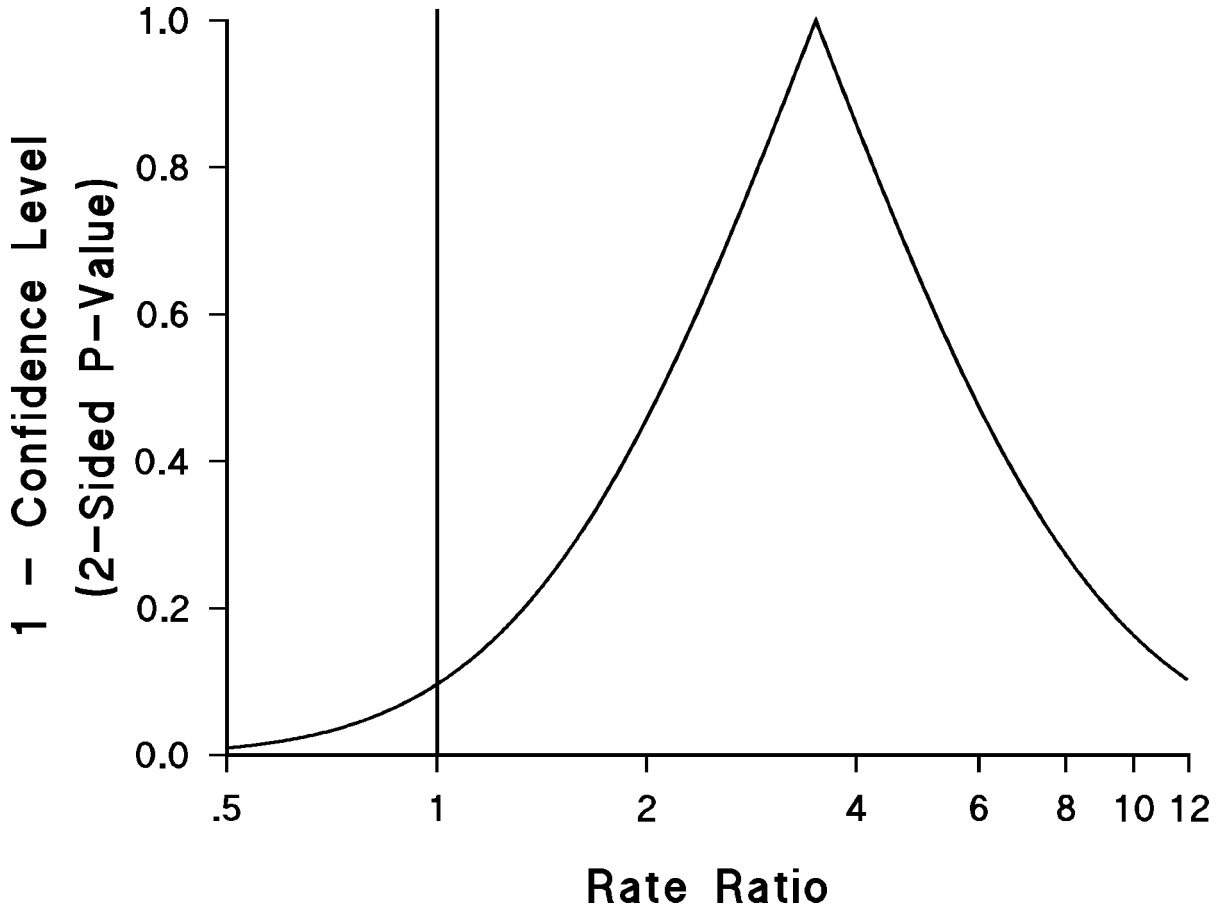
A confidence interval is a range of hypothesized parameter values for which the p-values testing those hypotheses are greater than a specified level.

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.

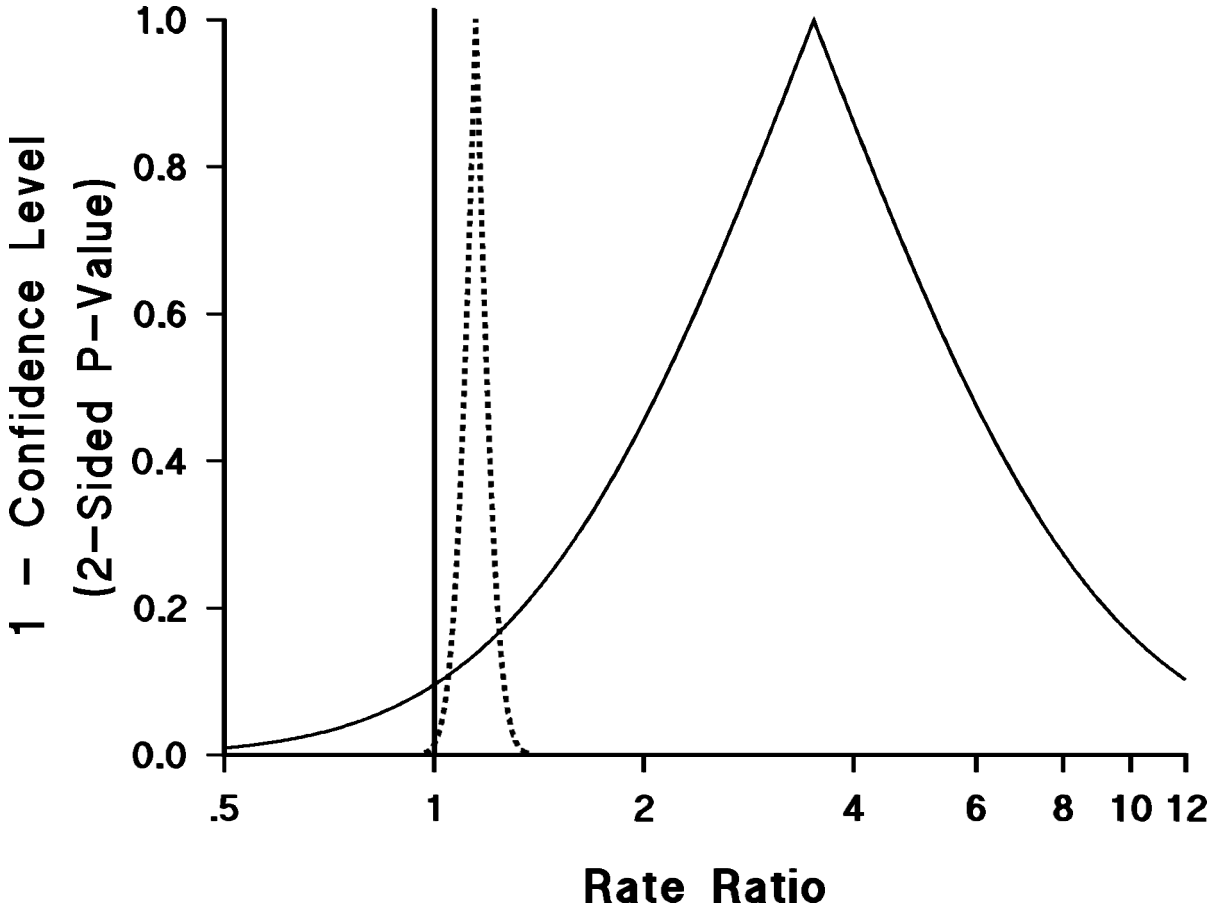
Confidence Interval or P-value Function



Confidence Interval or P-value Function



Confidence Interval or P-value Function

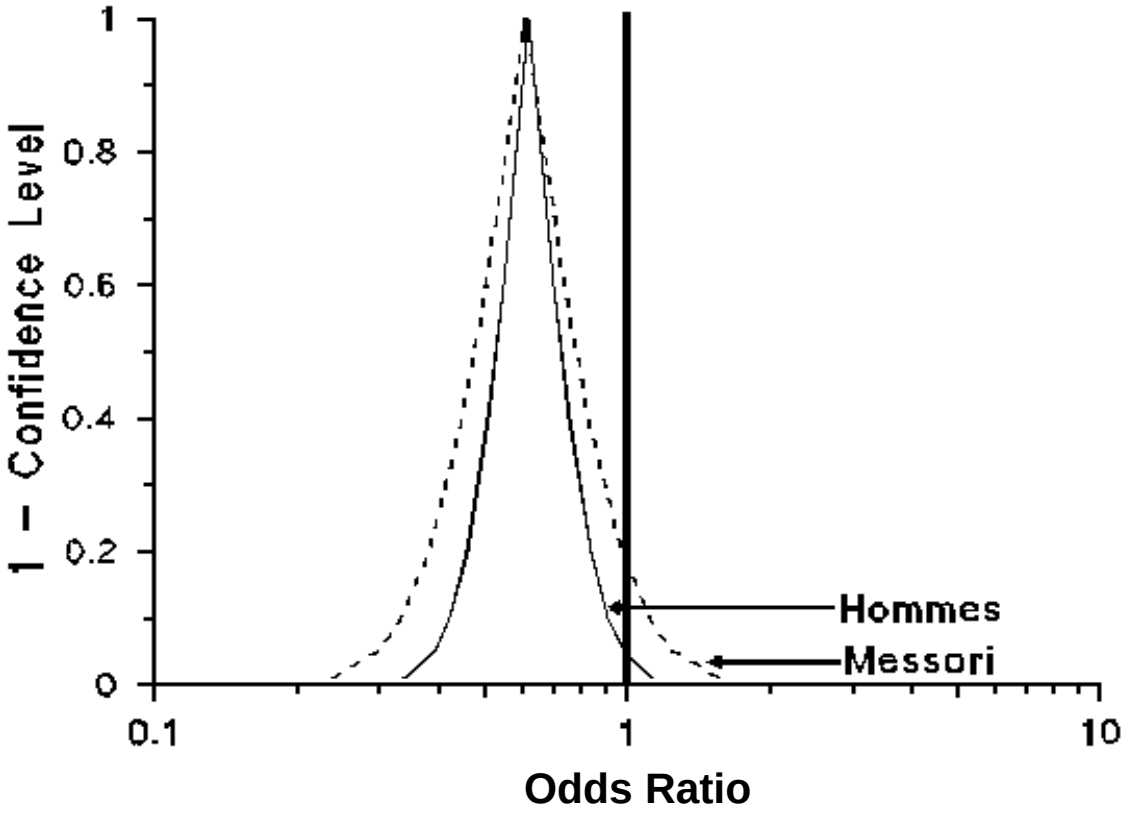


Calculation Errors in Meta-Analysis

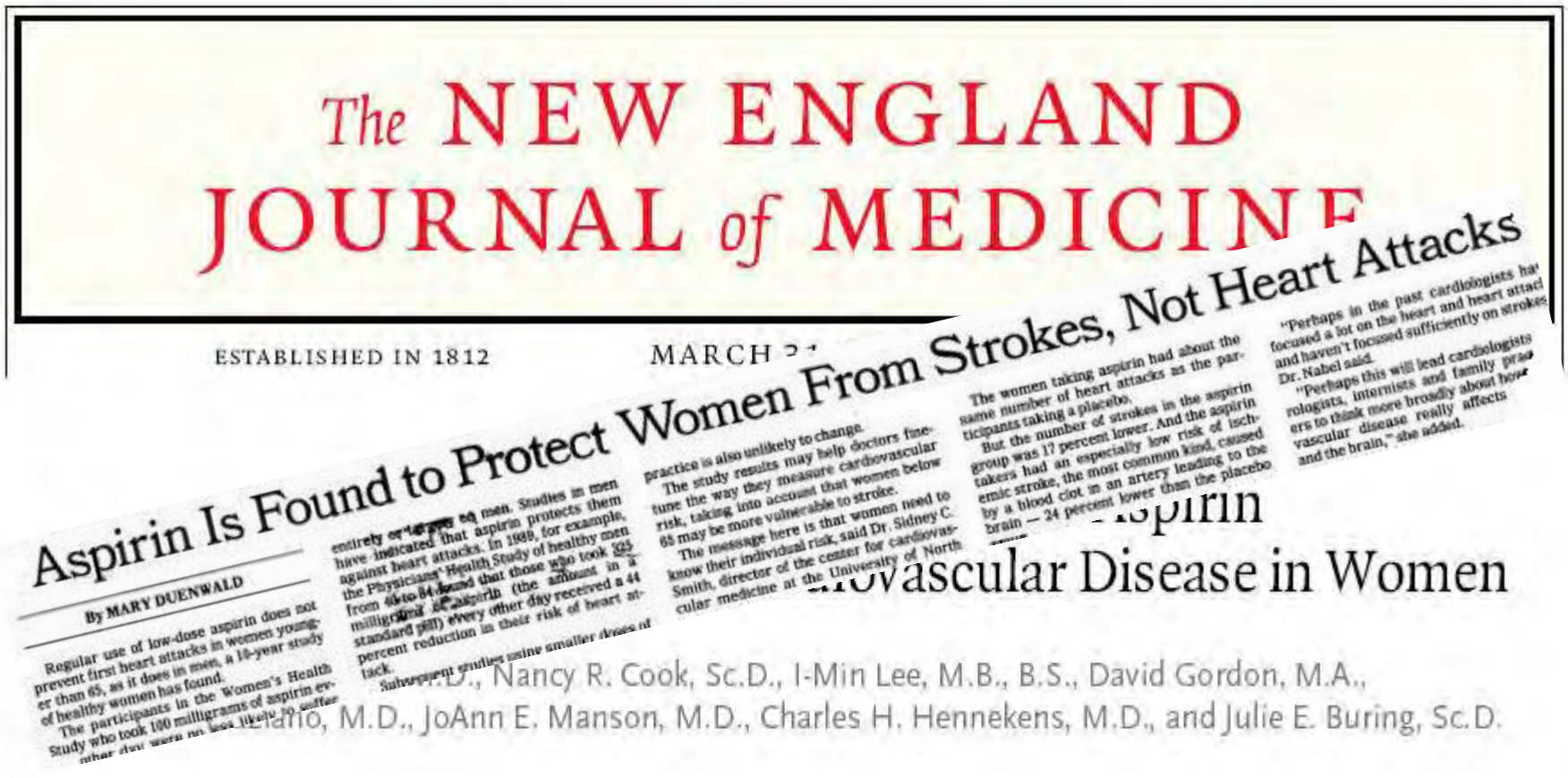
The recent paper by Hommes and colleagues reports a meta-analysis of six randomized trials comparing subcutaneous heparin with continuous intravenous heparin for the initial treatment of deep vein thrombosis....The result of our calculation was an odds ratio of 0.61 (95% CI, 0.298 to 1.251; P > 0.05); 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....

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

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



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 ev-

entirely or largely on men. Studies in men have indicated that aspirin protects them against heart attacks. In 1989, for example, the Physicians' Health Study of healthy men from 40 to 84 found that those who took 325 milligrams of aspirin (the amount in a standard pill) every other day received a 44 percent reduction in their risk of heart attack.

Subsequent studies using smaller doses of

practice is also unlikely to change.

The study results may help doctors fine-tune the way they measure cardiovascular risk, taking into account that women below 65 may be more vulnerable to stroke.

The message here is that women need to know their individual risk, said Dr. Sidney C. Smith, director of the center for cardiovascular medicine at the University of North

The women taking aspirin had about the same number of heart attacks as the participants taking a placebo.

But the number of strokes in the aspirin group was 17 percent lower. And the aspirin takers had an especially low risk of ischemic stroke, the most common kind, caused by a blood clot in an artery leading to the brain — 24 percent lower than the placebo

"Perhaps in the past cardiologists have focused a lot on the heart and heart attacks and haven't focused sufficiently on strokes," Dr. Nabel said.

"Perhaps this will lead cardiologists, rologists, internists and family practi ers to think more broadly about how vascular disease really affects and the brain," she added.

...spirin

...vascular Disease in Women

...M.D., Nancy R. Cook, Sc.D., I-Min Lee, M.B., B.S., David Gordon, M.A., ...llano, 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

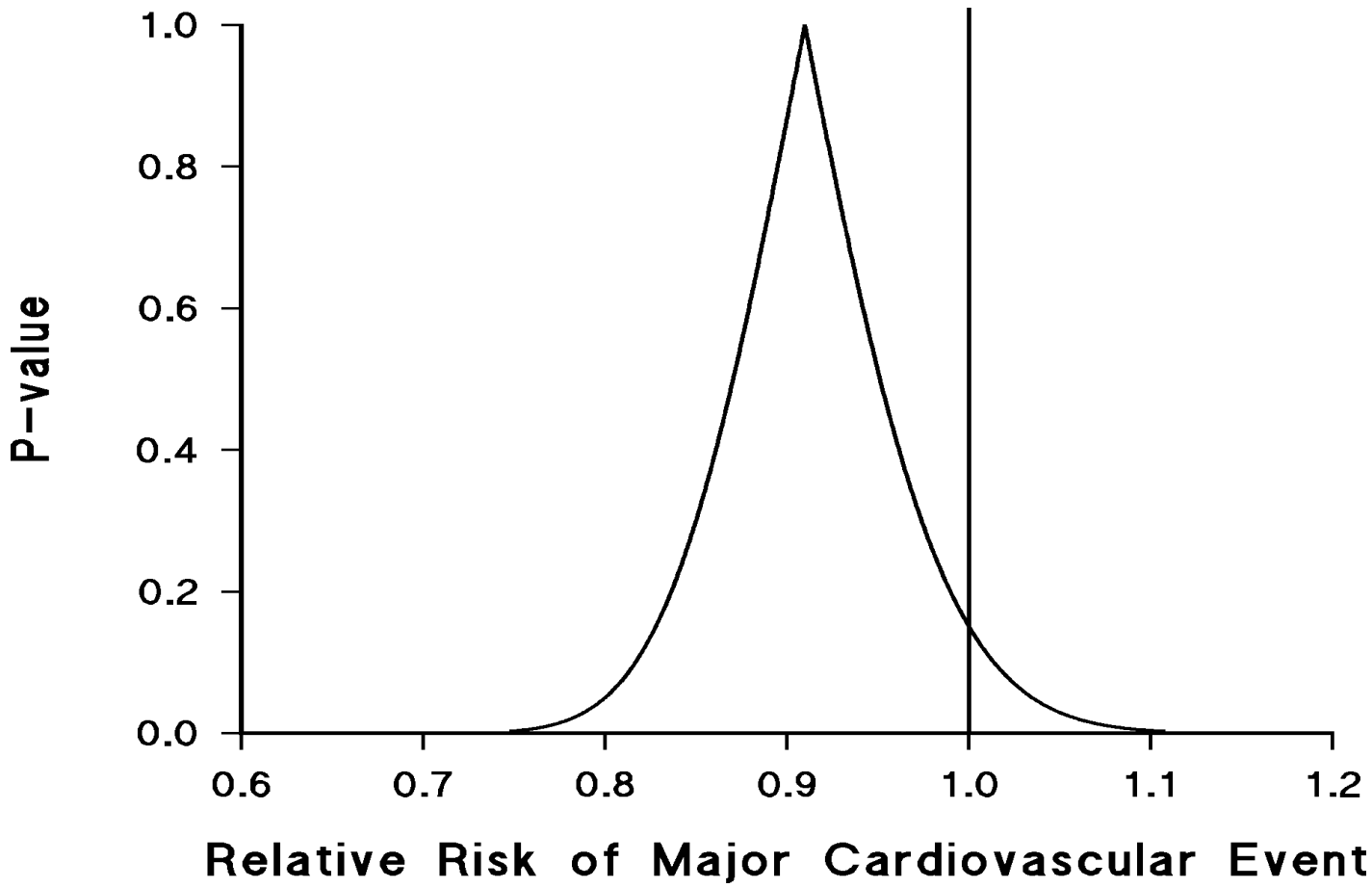
RTI (h)(s) Health Solutions

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

RTI (h)(s) Health Solutions

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.

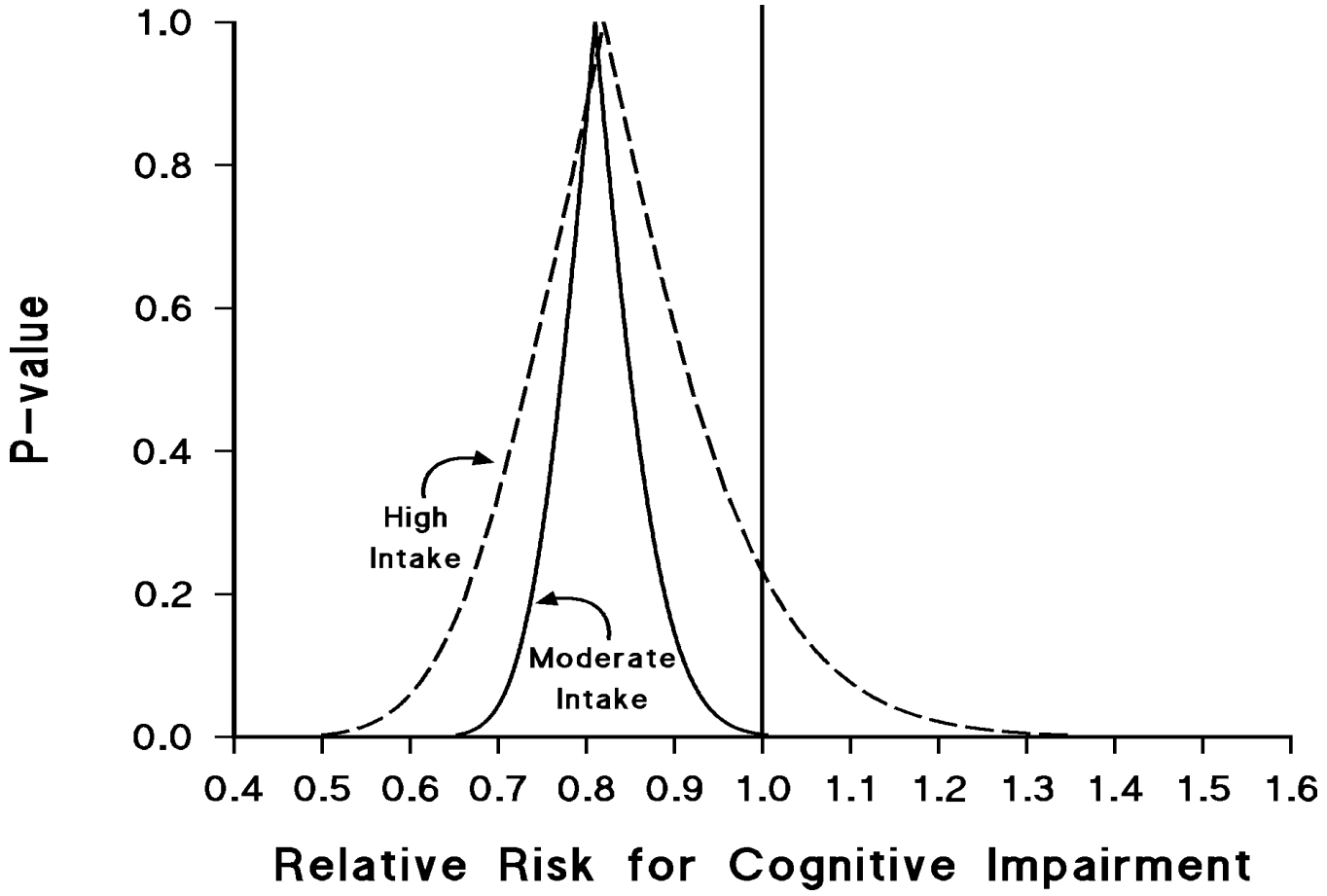
CI/P-value Function: Alcohol and Cognitive Impairment

RTI (h)(s) Health Solutions

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.

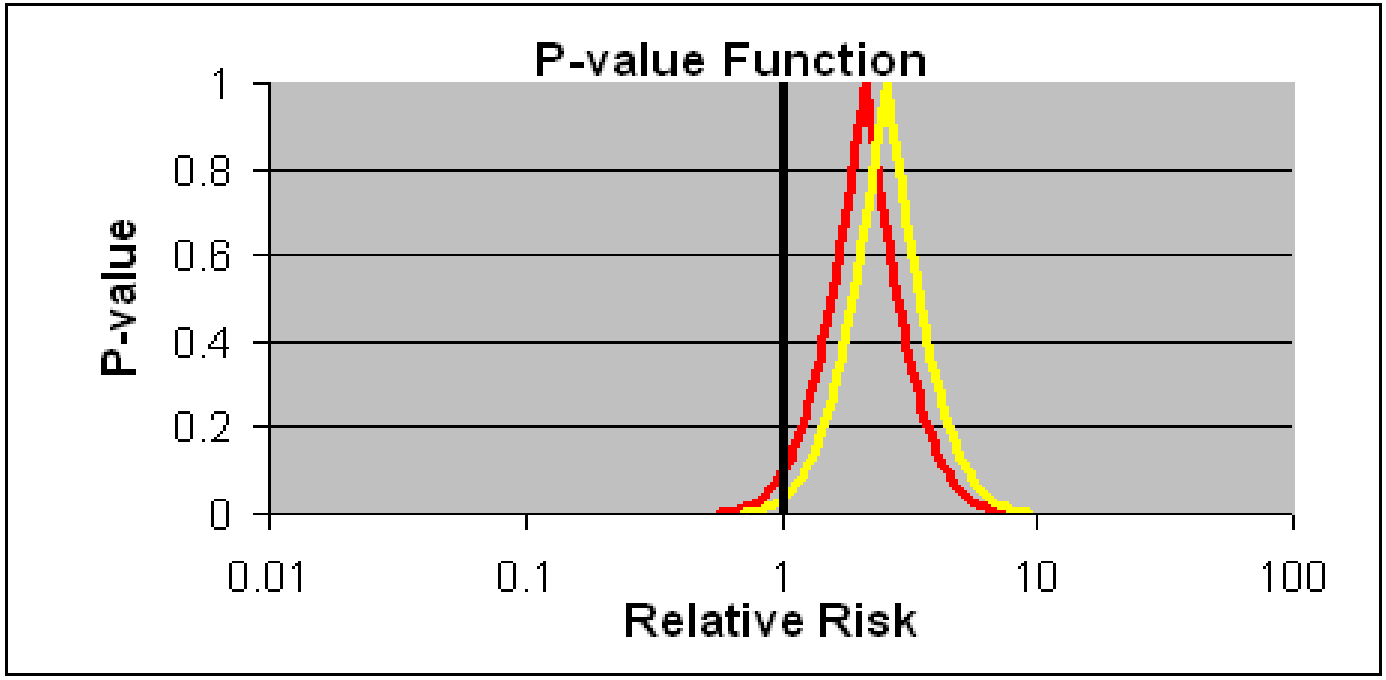
CI/P-value Function: Alcohol and Cognitive Impairment



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



Serotonergic Antidepressant Use During Pregnancy and Autism

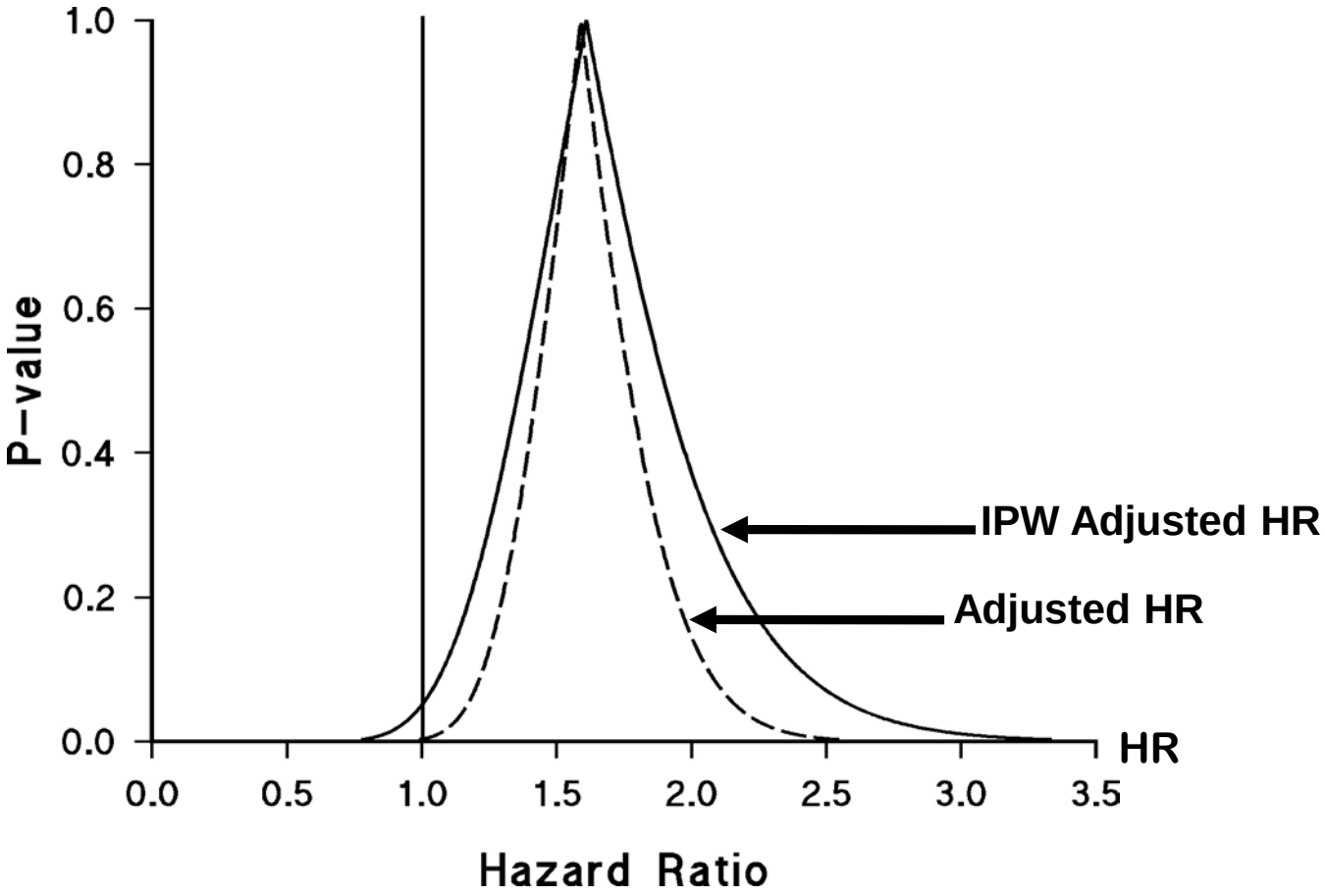
RTI (h)(s) Health Solutions

RESULTS There were 35 906 singleton births at a mean gestational age of 38.7 weeks (50.4% were male, mean maternal age was 26.7 years, and mean duration of follow-up was 4.95 years). In the 2837 pregnancies (7.9%) exposed to antidepressants, 2.0% (95% CI, 1.6%-2.6%) of children were diagnosed with autism spectrum disorder. The incidence of autism spectrum disorder was 4.51 per 1000 person-years among children exposed to antidepressants vs 2.03 per 1000 person-years among unexposed children (between-group difference, 2.48 [95% CI, 2.33-2.62] per 1000 person-years; hazard ratio [HR], 2.16 [95% CI, 1.64-2.86]; adjusted HR, 1.59 [95% CI, 1.17-2.17]). After inverse probability of treatment weighting based on the high-dimensional propensity score, the association was not significant (HR, 1.61 [95% CI, 0.997-2.59]). The association was also not significant when exposed children were compared with unexposed siblings (incidence of autism spectrum disorder was 3.40 per 1000 person-years vs 2.05 per 1000 person-years, respectively; adjusted HR, 1.60 [95% CI, 0.69-3.74]).

CONCLUSION: "…antidepressant exposure compared with no exposure was not associated with autism spectrum disorder…."

(JAMA. 2017;317(15):1544-1552. doi:10.1001/jama.2017.3415)

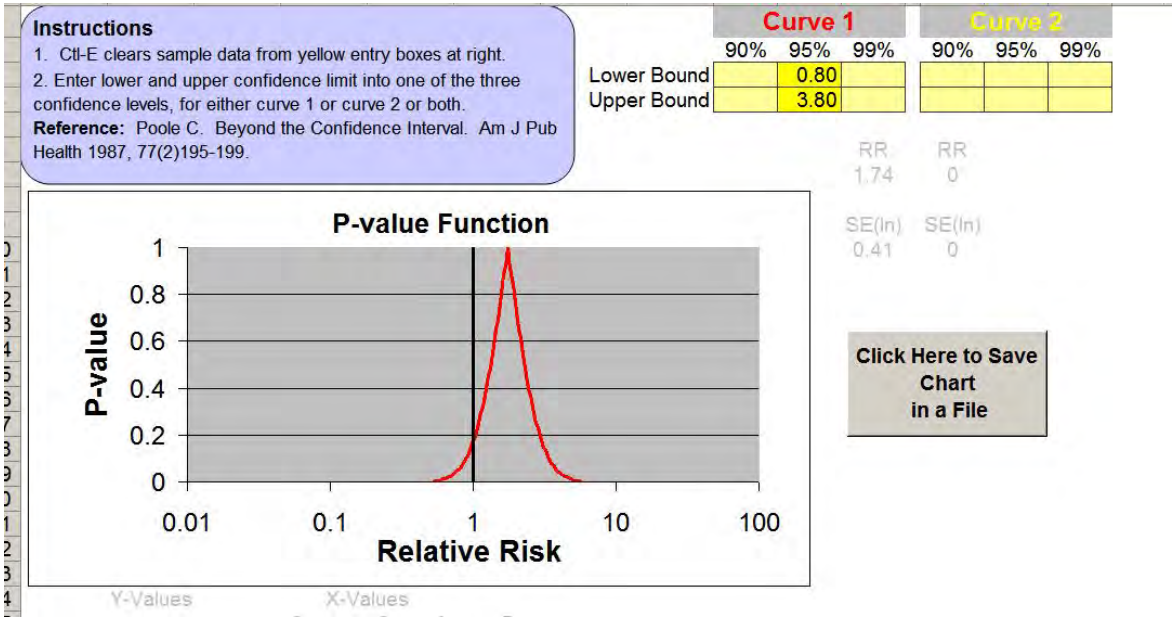
CI/P-value Functions: Antidepressants and Autism



How to Generate a P-value Function

1. You can use episheet:
   http://krothman.org/episheet.xls

2. The only input required is the lower bound and the upper bound of a confidence interval.



Criticism of Significance Testing is Not New

RTI (h)(s) Health Solutions

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*"

1997: William W. Rozeboom:

"*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.*"

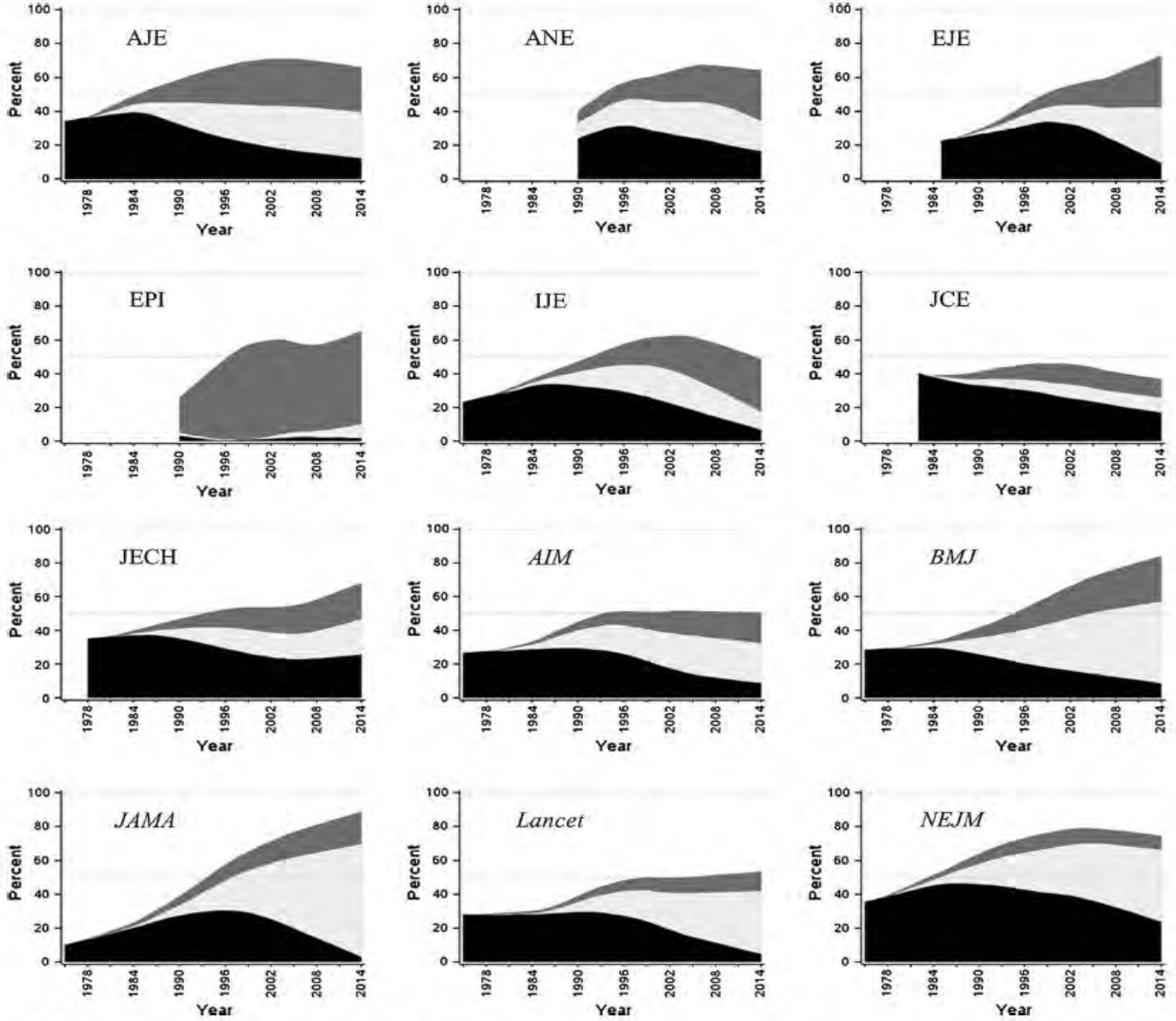


Fig. 2 Flexibly estimate time trends 1975–2004 in the prevalence of null hypothesis significance testing only, null hypothesis significance testing in combination with confidence intervals, and confidence intervals only in the abstracts of seven major epidemiology and five major medical journals. Flexibly (LOESS) fitted trend of the prevalence of statistical inference in abstracts; *black area* NHST-only; *light gray area* NHST combined with CIs; *dark gray area* CI-only; *white top area* percentage of abstracts that do not contain statistical inference; *AIM* Annals of Internal Medicine; *AJE* American Journal of Epidemiology; *ANE* Annals of Epidemiology; *EJE* European Journal of Epidemiology; *EPI* Epidemiology; *IJE* International Journal of Epidemiology; *JCE* Journal of Clinical Epidemiology; *JECH* Journal of Epidemiology and Community Health

Stang et al., EJE 2016

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA

ASA Statement: Six Principles

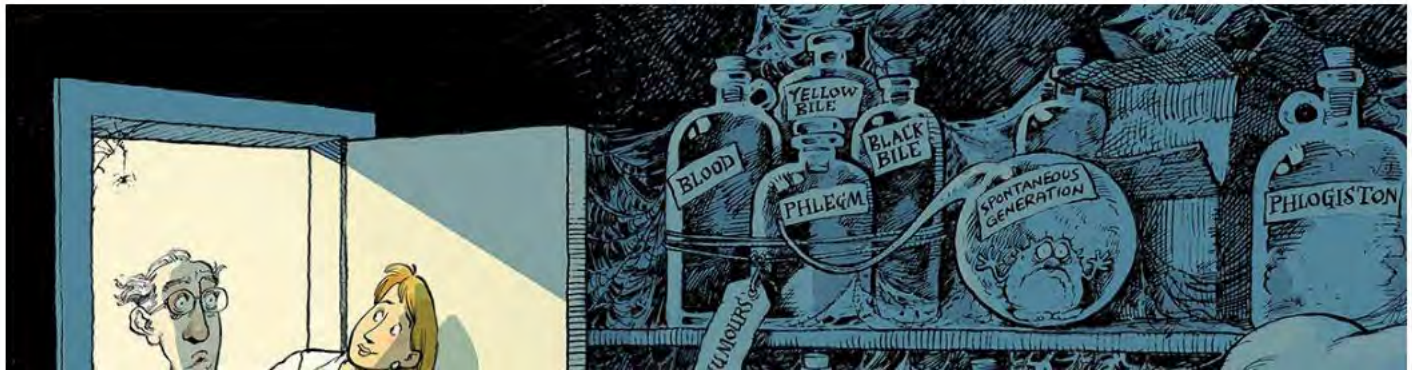
1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

COMMENT · 20 MARCH 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

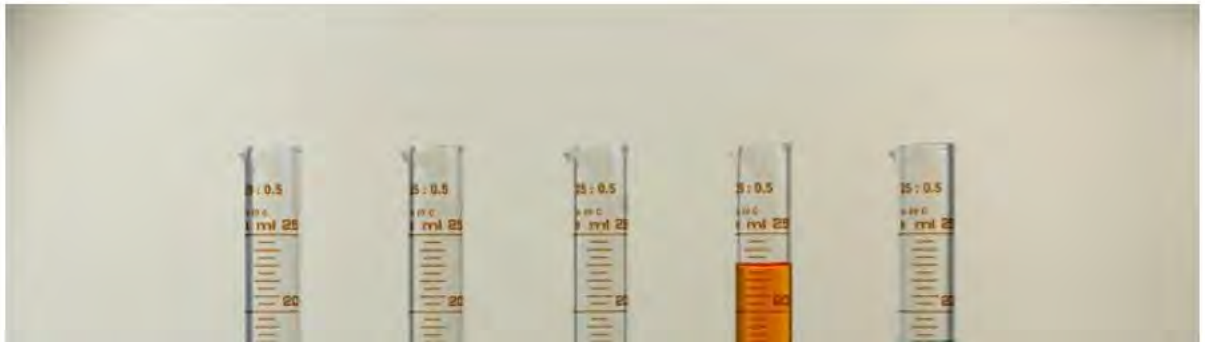
Valentin Amrhein, Sander Greenland & Blake McShane



EDITORIAL · 20 MARCH 2019

It's time to talk about ditching statistical significance

Looking beyond a much used and abused measure would make science harder, but better.



PDF version

RELATED ARTICLES

Scientists rise up against statistical significance

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 1–19: Editorial
https://doi.org/10.1080/00031305.2019.1583913

Taylor & Francis
Taylor & Francis Group

EDITORIAL

OPEN ACCESS

Moving to a World Beyond "$p < 0.05$"

Some of you exploring this special issue of *The American Statistician* might be wondering if it's a scolding from pedantic statisticians lecturing you about what *not* to do with *p*-values, without offering any real ideas of what *to do* about the very hard problem of separating signal from noise in data and making decisions under uncertainty. Fear not. In this issue, thanks to 43 innovative and thought-provoking papers from forward-looking statisticians, help is on the way.

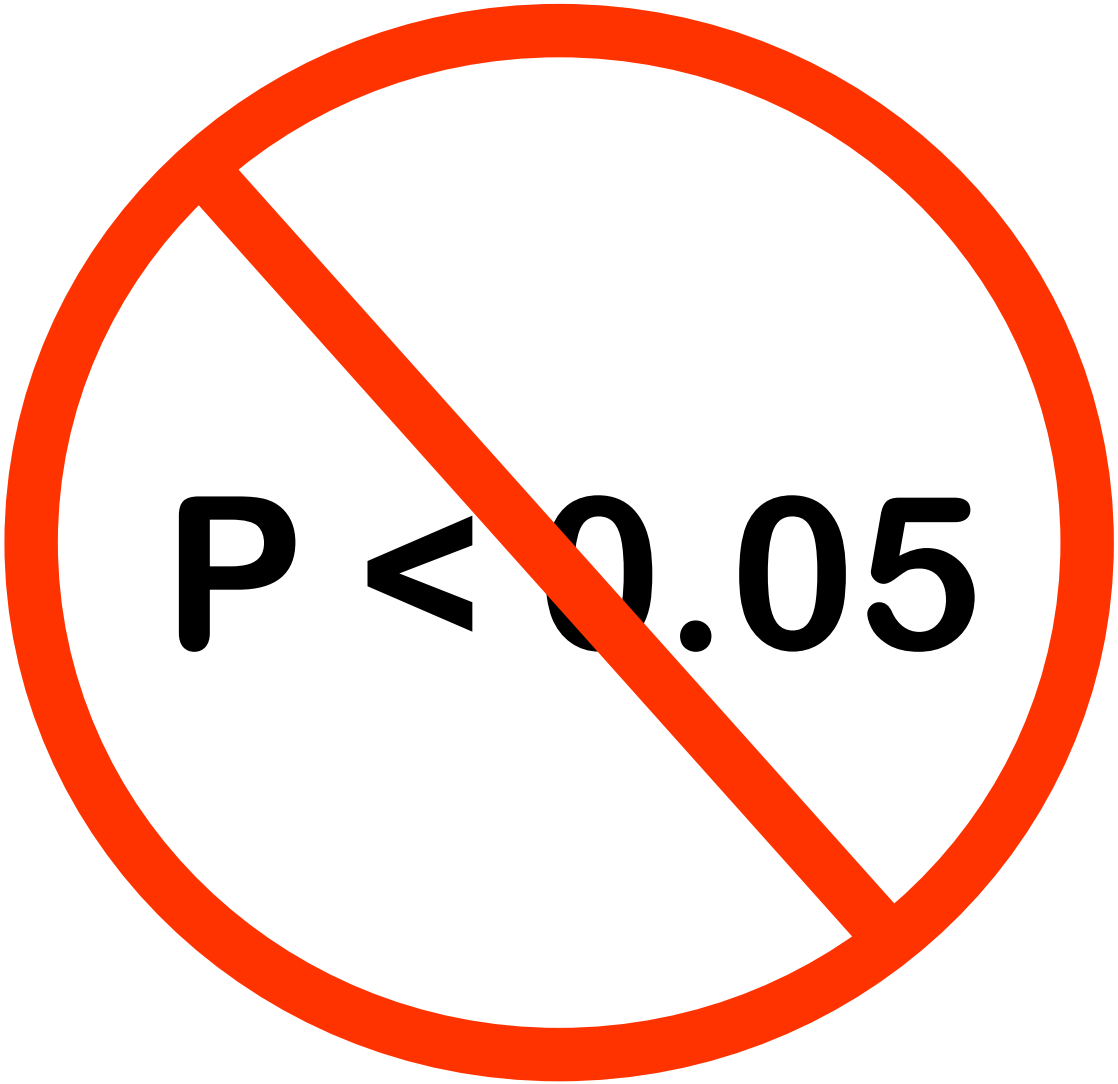
1. "Don't" Is Not Enough

There's not much we can say here about the perils of *p*-values and significance testing that hasn't been said already for decades (Ziliak and McCloskey 2008; Hubbard 2016). If you're just arriving to the debate, here's a sampling of what not to do:

special issue of *The American Statistician*. Authors were explicitly instructed to develop papers for the variety of audiences interested in these topics. If you use statistics in research, business, or policymaking but are not a statistician, these articles were indeed written with YOU in mind. And if you are a statistician, there is still much here for you as well.

The papers in this issue propose many new ideas, ideas that in our determination as editors merited publication to enable broader consideration and debate. The ideas in this editorial are likewise open to debate. They are our own attempt to distill the wisdom of the many voices in this issue into an essence of good statistical practice as we currently see it: some do's for teaching, doing research, and informing decisions.

Yet the voices in the 43 papers in this issue do not sing as one. At times in this editorial and the papers you'll hear deep dissonance, the echoes of "statistics wars" still simmering today



Thank You

Questions?

Generating knowledge and providing greater understanding so that you—and those who regulate, pay for, prescribe, and use your products—can make better decisions.

rtihs.org

Contact Us

Kenneth J. Rothman
krothman@rti.org
@ken_rothman

Heather Danysh
hdanysh@rti.org