University of California
San Francisco

The celebrated fall of the *p*-value: Time for a new paradigm

Liz Whitlock, MD, MSc
John Severinghaus Assistant Professor In Residence
University of California, San Francisco
Department of Anesthesia & Perioperative Care

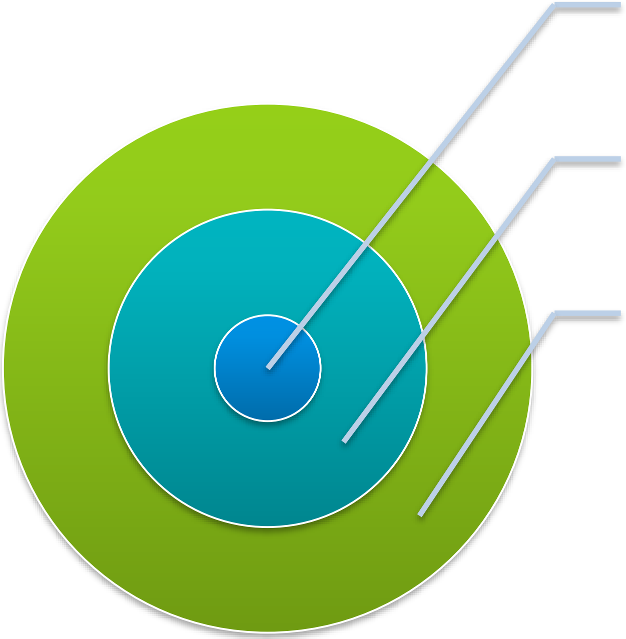
MPOG Research Retreat
October 12, 2018

No financial relationships to disclose

Funding: NIA GEMSSTAR R03AG059822, Foundation for Anesthesia Education & Research Jahnigan Career Development Award

UCSF

Who’s who in this talk’s audience



Novice researchers

(And consumers of the literature)

Experienced researchers

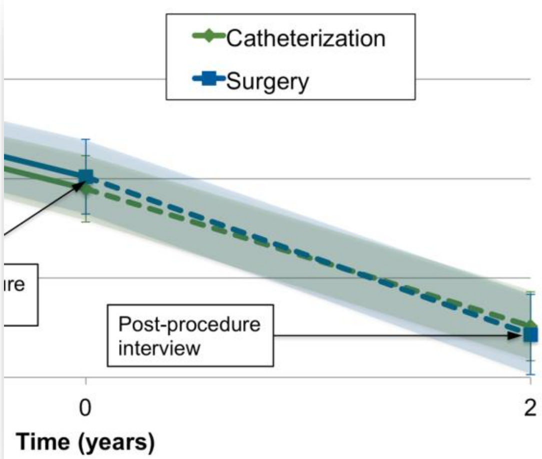
Reviewers

(As the gateway to academic currency)

It all begins with a “case”…

35yo F junior researcher submits a manuscript

- Major finding: CABG associated with equivalent of additional 4 months [95% CI -1 to 10] of cognitive aging at up to 2 years post-procedure
  - *p* = 0.12
  - *"Population-level impact of surgery, if it exists, is likely to be subtle"*

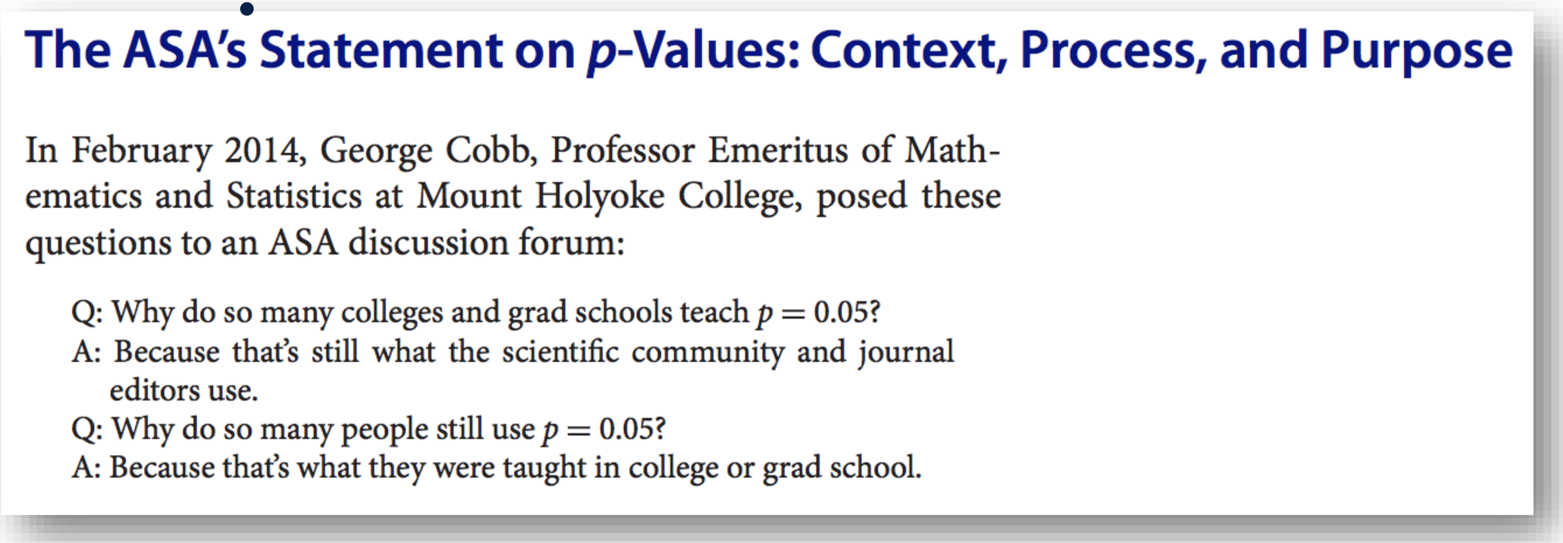


"Your borderline finding (P=0.12) and 95% CI cannot support this and so we remain uncertain."

Whitlock et al, Ann Thorac Surg, *in press*

The problem

Sadly, not
our ASA



The ASA’s Statement on *p*-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach $p = 0.05$?
> A: Because that’s still what the scientific community and journal editors use.
> Q: Why do so many people still use $p = 0.05$?
> A: Because that’s what they were taught in college or grad school.

Wasserstein & Lazar, Am Stat 2016 70(2):129-133, 2016

$p<0.05$

Null hypothesis significance testing
versus
the estimation approach

“Beer-significance”

A new(ish?) way of thinking about *significance*

Null hypothesis significance testing

- Dichotomous: difference or not
- A
  -
  - failing to reject h_0
- Frequently becomes "probability h_0 is true" – inverse probability fallacy

"A p value 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!"

The estimation approach

- Also deals with…*significance*
- What *do* you want to know?
  - How much?
  - Does it really matter?
  - Should it change what you do?
- Study design (and bias)

Cumming G & Calin-Jageman R. Introduction to the New Statistics. Routledge (New York, NY), 2017. Pp 142-150.
Jacob Cohen, *quoted in* Introduction to the New Statistics

Where did this tension come from? Historical context

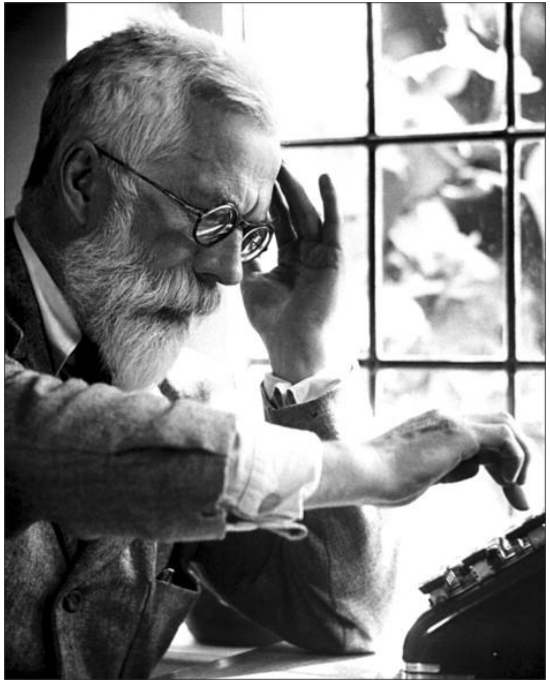
Fisherian vs Gossetian statistics

- Randomized design
- Validity
- Statistical significance
- “Student’s” t-test

“Theoretically plausible symmetric error distribution…around the mean result”

William Sealy Gosset (1876-1937) had a different definition of “validity”

“Beer-significance” Efficacy, value, strength, robustness

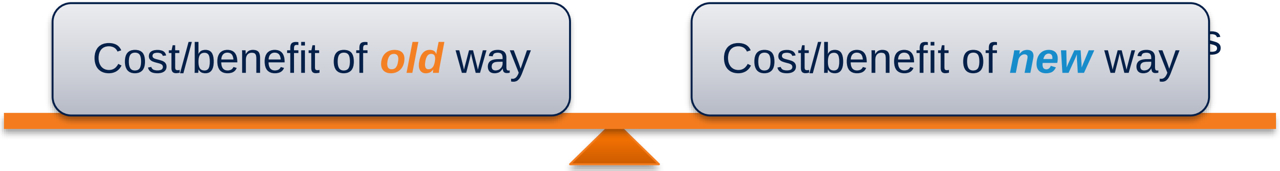


Ronald A. Fisher (1890-1962), deeply contemplating statistical significance

Ziliak ST. The *Validus Medicus* and a new gold standard. Lancet 2010 Jul 31; 376(9738): 324-5
Ziliak ST & McCloskey DN. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. Univ of Michigan Press, 2008
Fisher RA. Statistical methods for research workers. Oliver & Boyd (Edinburgh), 1925

Gosset, 1905, in a letter to Karl Pearson

“In such work as ours, the degree of certainty…must depend on the advantage to be gained by following the result of the experiment, compared with…the cost of the new method, and the cost of each experiment.”



Cost/benefit of ***old*** way

Cost/benefit of ***new*** way

So, why don’t we do it Gosset’s way?

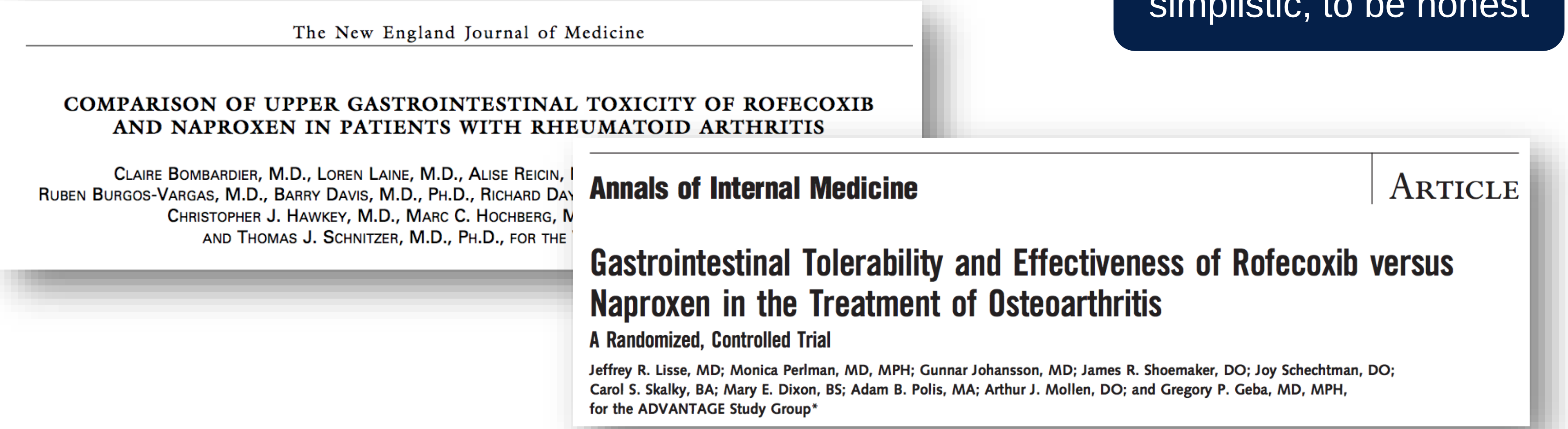
“[Significance testing at the 5% level has] raised economics, psychology, and medicine to the ranks of sciences.” - Fisher, 1930

Where did p<.05 come from?

Overly simplistic, but close enough (see Cowles & Davis)

"Personally, the writer prefers to set a low standard of significance at the 5 per cent point, *and ignore entirely all results which fail to reach this level*." Fisher RA, 1926

Also a bit overly simplistic, to be honest



The New England Journal of Medicine

COMPARISON OF UPPER GASTROINTESTINAL TOXICITY OF ROFECOXIB AND NAPROXEN IN PATIENTS WITH RHEUMATOID ARTHRITIS

Claire Bombardier, M.D., Loren Laine, M.D., Alise Reicin, I
Ruben Burgos-Vargas, M.D., Barry Davis, M.D., Ph.D., Richard Day
Christopher J. Hawkey, M.D., Marc C. Hochberg, M
and Thomas J. Schnitzer, M.D., Ph.D., for the

Annals of Internal Medicine | Article

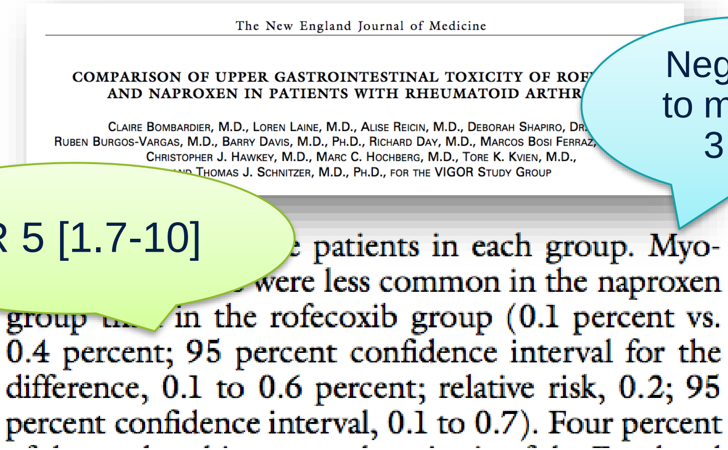
Gastrointestinal Tolerability and Effectiveness of Rofecoxib versus Naproxen in the Treatment of Osteoarthritis

A Randomized, Controlled Trial

Jeffrey R. Lisse, MD; Monica Perlman, MD, MPH; Gunnar Johansson, MD; James R. Shoemaker, DO; Joy Schechtman, DO; Carol S. Skalky, BA; Mary E. Dixon, BS; Adam B. Polis, MA; Arthur J. Mollen, DO; and Gregory P. Geba, MD, MPH, for the ADVANTAGE Study Group*

Fisher RA. The arrangement of field experiments. Journal of the Ministry of Agriculture 1926; 33:503-13.
Cowles M & Davis C. On the origins of the .05 level of statistical significance. Am Psychol 1982 May; 37(5):553-8.

“Ignore entirely all results which fail to reach this level.”



The New England Journal of Medicine

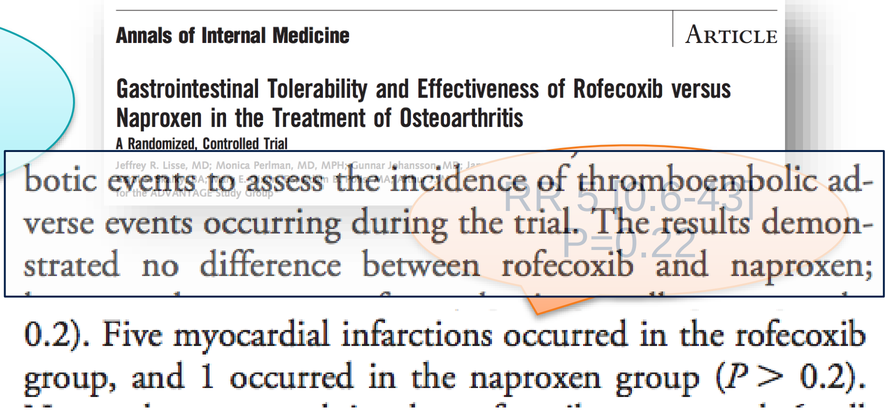
COMPARISON OF UPPER GASTROINTESTINAL TOXICITY OF ROF
AND NAPROXEN IN PATIENTS WITH RHEUMATOID ARTHR

CLAIRE BOMBARDIER, M.D., LOREN LAINE, M.D., ALISE REICIN, M.D., DEBORAH SHAPIRO, DR.
RUBEN BURGOS-VARGAS, M.D., BARRY DAVIS, M.D., PH.D., RICHARD DAY, M.D., MARCOS BOSI FERRAZ,
CHRISTOPHER J. HAWKEY, M.D., MARC C. HOCHBERG, M.D., TORE K. KVIEN, M.D.,
THOMAS J. SCHNITZER, M.D., PH.D., FOR THE VIGOR STUDY GROUP

e patients in each group. Myo- were less common in the naproxen group in the rofecoxib group (0.1 percent vs. 0.4 percent; 95 percent confidence interval for the difference, 0.1 to 0.6 percent; relative risk, 0.2; 95 percent confidence interval, 0.1 to 0.7). Four percent

RR 5 [1.7-10]

Neglected to mention 3 MIs



Annals of Internal Medicine | ARTICLE

Gastrointestinal Tolerability and Effectiveness of Rofecoxib versus Naproxen in the Treatment of Osteoarthritis

A Randomized, Controlled Trial

Jeffrey R. Lisse, MD; Monica Perlman, MD, MPH; Gunnar Johansson, MD; Ja
for the ADVANTAGE Study Group

botic events to assess the incidence of thromboembolic adverse events occurring during the trial. The results demonstrated no difference between rofecoxib and naproxen;

0.2). Five myocardial infarctions occurred in the rofecoxib group, and 1 occurred in the naproxen group (*P* > 0.2).

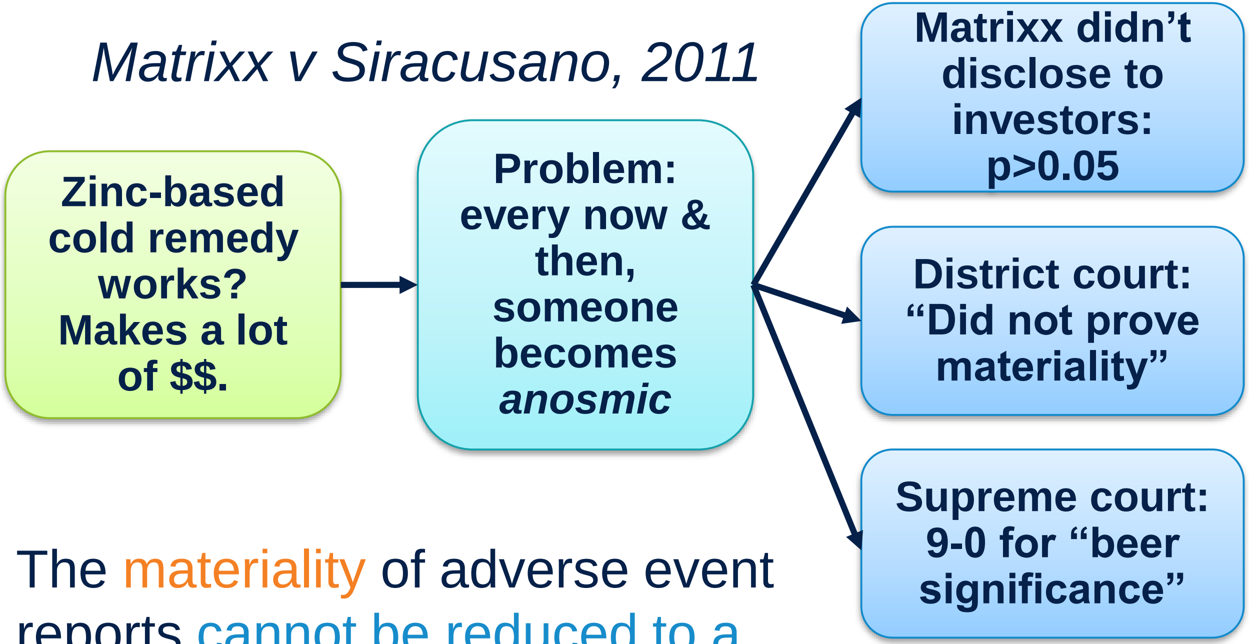
RR 5 [0.6-43]
P=0.22

- May 1999: Vioxx approved
- Mar 2000: Merck shadiness
- Nov 2000: NEJM paper
- Oct 2003: Ann Int Med paper
- Sept 2004: Vioxx withdrawn
- Nov 2007: $4.85b settlement

Bombardier C et al. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. NEJM 2000 Nov; 343:1520-1528.
Lisse JR et al. Gastrointestinal tolerability and effectiveness of rofecoxib versus naproxen in the treatment of osteoarthritis. Ann Int Med 2003 Oct; 139(7):539-46.
Ziliak & McCloskey. The Cult of Statistical Significance: Ch. 1

Even the Supreme Court has now weighed in

Matrixx v Siracusano, 2011



Zinc-based cold remedy works? Makes a lot of $$.

Problem: every now & then, someone becomes *anosmic*

Matrixx didn't disclose to investors: $p>0.05$

District court: "Did not prove materiality"

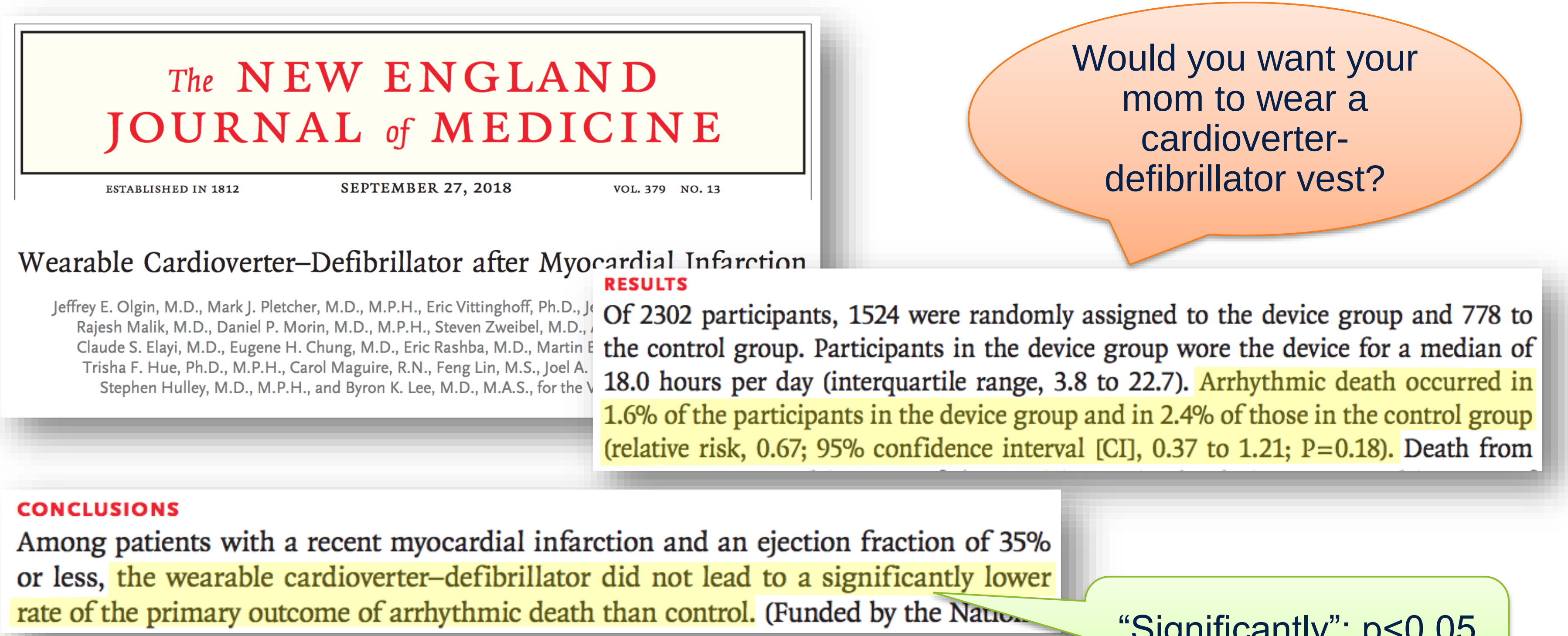
Supreme court: 9-0 for "beer significance"

The materiality of adverse event reports cannot be reduced to a bright-line rule." - Justice Sotomayor

Notice: now zinc-free…

Ziliak & McCloskey. Lady Justice v. Cult of Statistical Significance. *In* Oxford Handbook on Professional Economic Ethics. Oxford UP, 2014.
Ziliak ST. Matrixx v. Siracusano and Student v. Fisher. Significance 2011 Sept; 8(3):131-4.

But we’re WAY more sophisticated now.



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 SEPTEMBER 27, 2018 VOL. 379 NO. 13

Wearable Cardioverter–Defibrillator after Myocardial Infarction

Jeffrey E. Olgin, M.D., Mark J. Pletcher, M.D., M.P.H., Eric Vittinghoff, Ph.D., Je
Rajesh Malik, M.D., Daniel P. Morin, M.D., M.P.H., Steven Zweibel, M.D., A
Claude S. Elayi, M.D., Eugene H. Chung, M.D., Eric Rashba, M.D., Martin E
Trisha F. Hue, Ph.D., M.P.H., Carol Maguire, R.N., Feng Lin, M.S., Joel A.
Stephen Hulley, M.D., M.P.H., and Byron K. Lee, M.D., M.A.S., for the V

RESULTS

Of 2302 participants, 1524 were randomly assigned to the device group and 778 to the control group. Participants in the device group wore the device for a median of 18.0 hours per day (interquartile range, 3.8 to 22.7). Arrhythmic death occurred in 1.6% of the participants in the device group and in 2.4% of those in the control group (relative risk, 0.67; 95% confidence interval [CI], 0.37 to 1.21; P=0.18). Death from

CONCLUSIONS

Among patients with a recent myocardial infarction and an ejection fraction of 35% or less, the wearable cardioverter–defibrillator did not lead to a significantly lower rate of the primary outcome of arrhythmic death than control. (Funded by the Natio

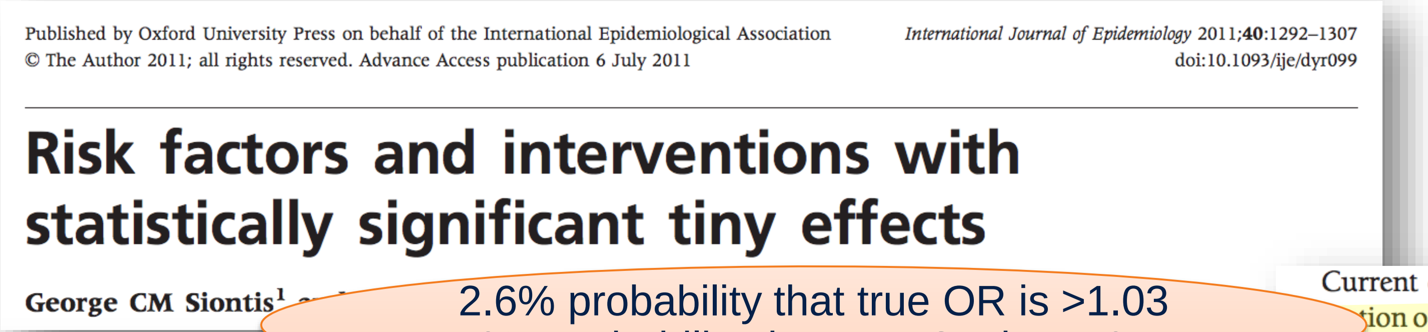
Would you want your mom to wear a cardioverter-defibrillator vest?

“Significantly”: $p<0.05$ or 1/3 reduction?

Olgin JE et al. Wearable cardioverter-defibrillator after myocardial infarction. NEJM 2018 Sep; 379(13): 1205-15.

How does this play out with very large samples?

“You can make the *p*-value as small as you can afford”



Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2011; all rights reserved. Advance Access publication 6 July 2011

International Journal of Epidemiology 2011;**40**:1292–1307
doi:10.1093/ije/dyr099

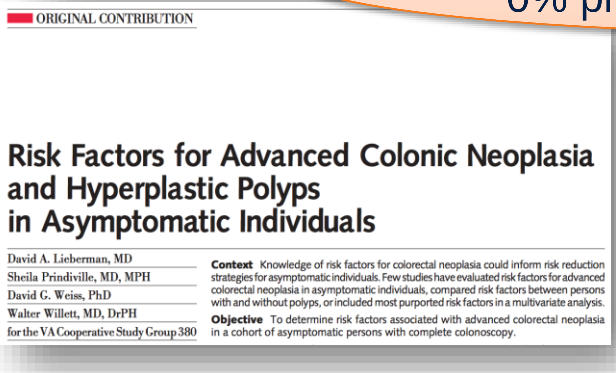
Risk factors and interventions with statistically significant tiny effects

George CM Siontis[1]

RR 0.95-1.05
& p<0.05

NEJM, JAMA, Lancet

2.6% probability that true OR is >1.03
0% probability that true OR is >1.05



ORIGINAL CONTRIBUTION

Risk Factors for Advanced Colonic Neoplasia and Hyperplastic Polyps in Asymptomatic Individuals

David A. Lieberman, MD
Sheila Prindiville, MD, MPH
David G. Weiss, PhD
Walter Willett, MD, DrPH
for the VA Cooperative Study Group 380

Context Knowledge of risk factors for colorectal neoplasia could inform risk reduction strategies for asymptomatic individuals. Few studies have evaluated risk factors for advanced colorectal neoplasia in asymptomatic individuals, compared risk factors between persons with and without polyps, or included most purported risk factors in a multivariate analysis.

Objective To determine risk factors associated with advanced colorectal neoplasia in a cohort of asymptomatic persons with complete colonoscopy.

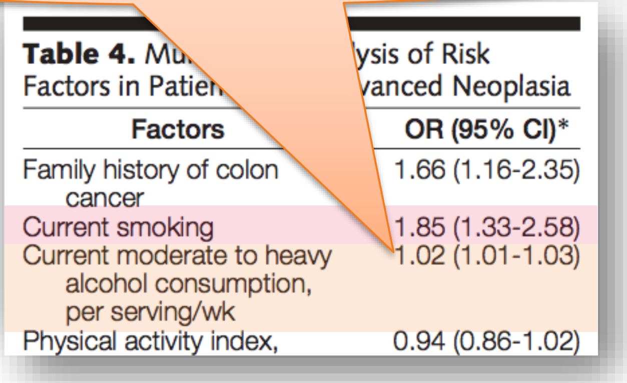
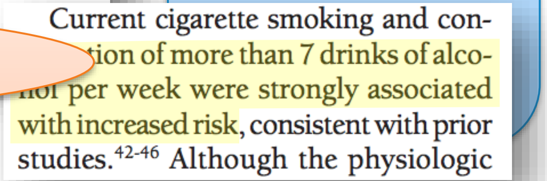
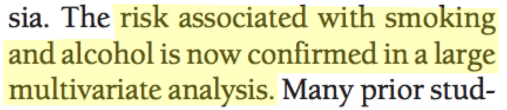


Table 4. Mu ysis of Risk Factors in Patie vanced Neoplasia

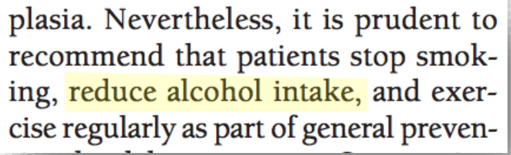
Factors	OR (95% CI)*
Family history of colon cancer	1.66 (1.16-2.35)
Current smoking	1.85 (1.33-2.58)
Current moderate to heavy alcohol consumption, per serving/wk	1.02 (1.01-1.03)
Physical activity index,	0.94 (0.86-1.02)



Current cigarette smoking and con- tion of more than 7 drinks of alcohol per week were strongly associated with increased risk, consistent with prior studies.[42-46] Although the physiologic



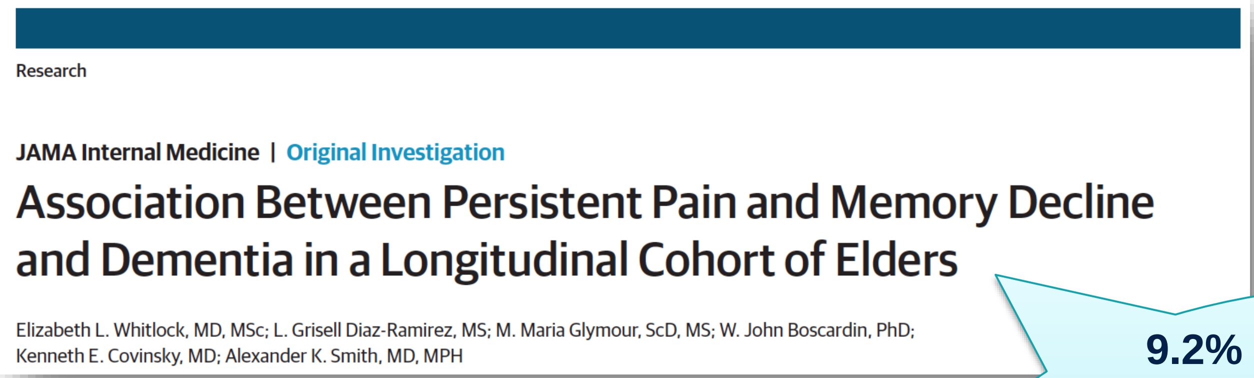
sia. The risk associated with smoking and alcohol is now confirmed in a large multivariate analysis. Many prior stud-



plasia. Nevertheless, it is prudent to recommend that patients stop smoking, reduce alcohol intake, and exercise regularly as part of general preven-

Demidenko E. The *p*-value you can’t buy. Am Stat 2016 Mar; 70(1):33-8.
Siontis GCM & Ioannidis JPA. Risk factors and interventions with statistically significant tiny effects. Int J Epidemiol 2011 Jul; 40:1292-1307.
Lieberman et al. Risk factors for advanced colonic neoplasia and hyperplastic polyps in asymptomatic individuals. JAMA 2003 Dec; 290:2959-67.

I’d never do this. Never.

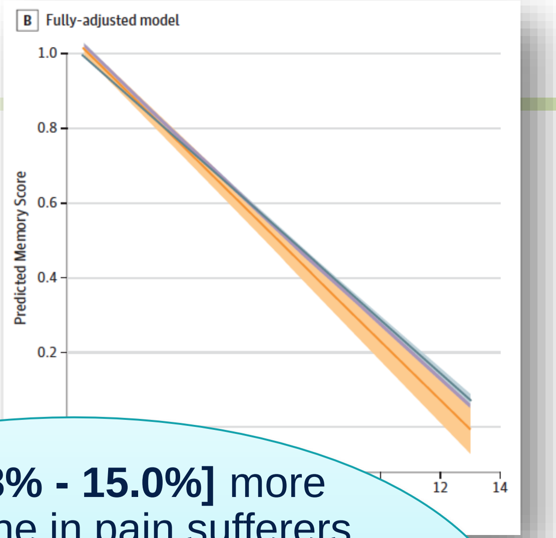


Research

JAMA Internal Medicine | Original Investigation

Association Between Persistent Pain and Memory Decline and Dementia in a Longitudinal Cohort of Elders

Elizabeth L. Whitlock, MD, MSc; L. Grisell Diaz-Ramirez, MS; M. Maria Glymour, ScD, MS; W. John Boscardin, PhD; Kenneth E. Covinsky, MD; Alexander K. Smith, MD, MPH



9.2% [2.8% - 15.0%] more rapid decline in pain sufferers vs controls!

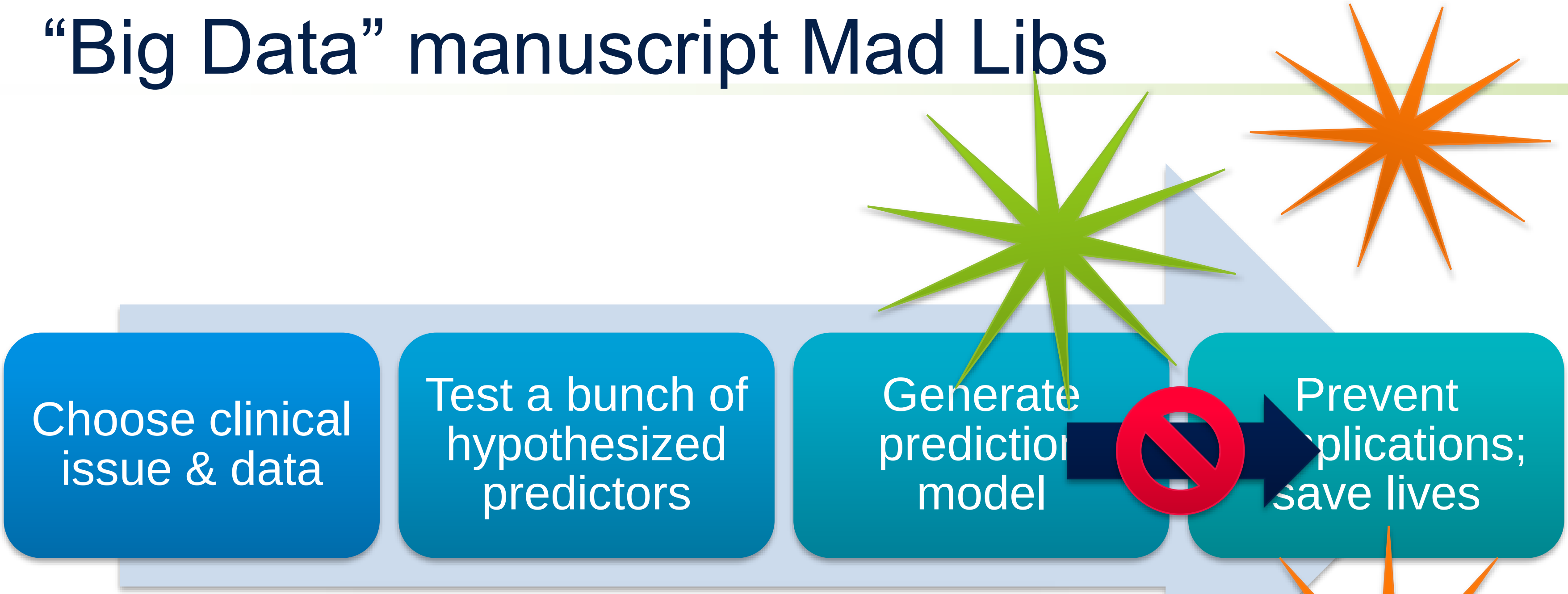
“Patients reporting ongoing pain may be at higher risk for current and incident cognitive impairment…”

Perhaps a “tiny effect” transgression.

But I also did *something else* stemming from “significance = *significance*” that obstructs science…

Database research, odds ratios, and prediction

“Big Data” manuscript Mad Libs



Choose clinical issue & data

Test a bunch of hypothesized predictors

Generate prediction model

Prevent complications; save lives

9% faster decline!

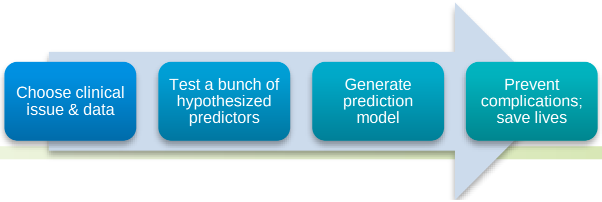
Research

JAMA Internal Medicine | Original Investigation

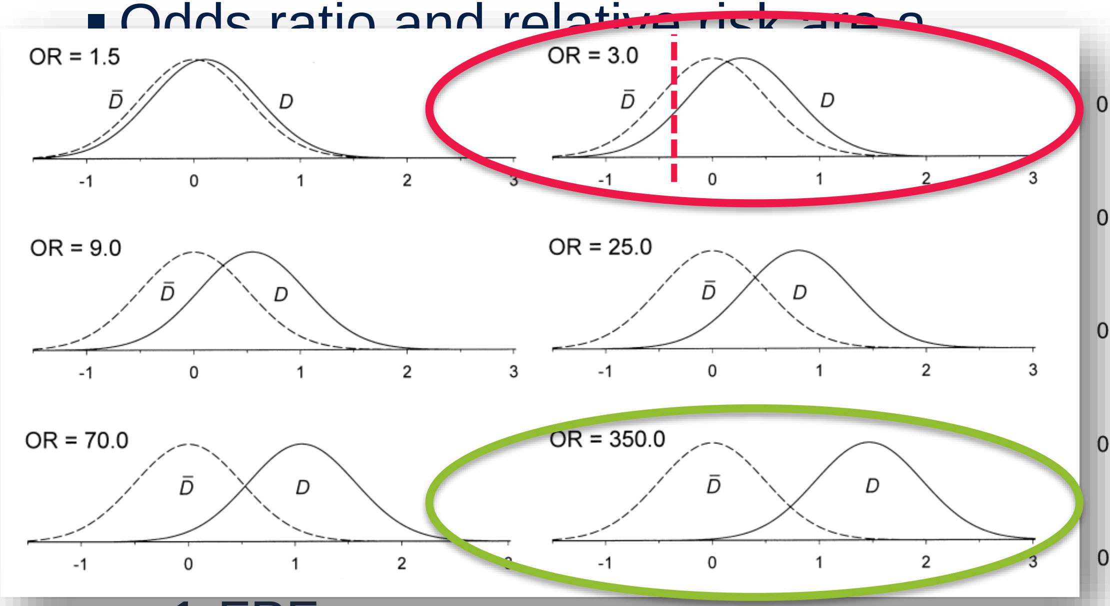
Association Between Persistent Pain and Memory Decline and Dementia in a Longitudinal Cohort of Elders

Elizabeth L. Whitlock, MD, MSc; L. Grisell Diaz-Ramirez, MS; M. Maria Glymour, ScD, MS; W. John Boscardin, PhD; Kenneth E. Covinsky, MD; Alexander K. Smith, MD, MPH

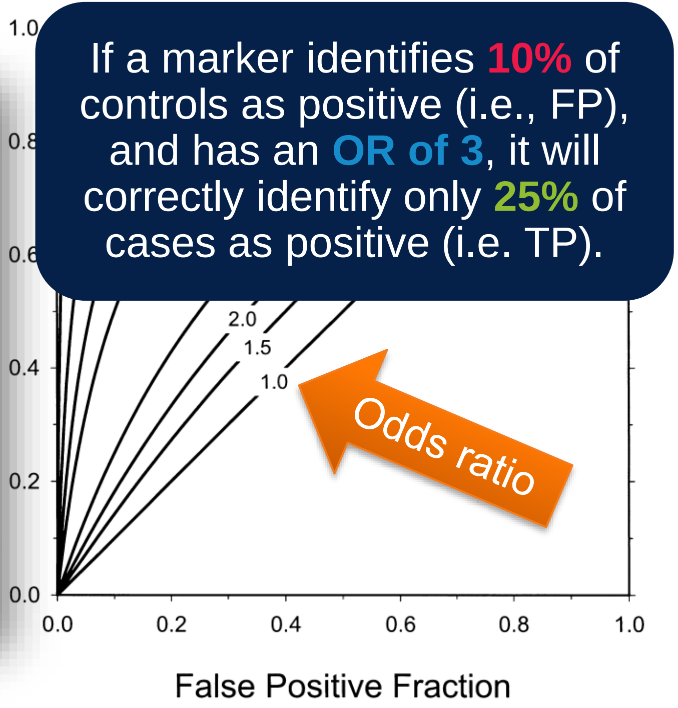
What’s wrong with that?


Choose clinical issue & data
Test a bunch of hypothesized predictors
Generate prediction model
Prevent complications; save lives

- Odds ratio and relative risk are a



$$\frac{1\text{-FPF}}{\text{FPF}}$$

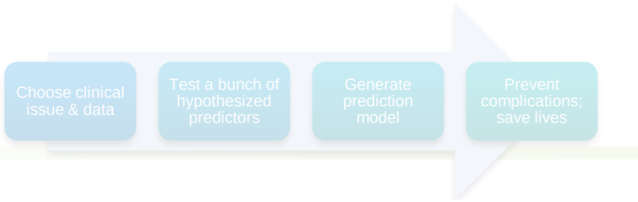


If a marker identifies 10% of controls as positive (i.e., FP), and has an OR of 3, it will correctly identify only 25% of cases as positive (i.e. TP).

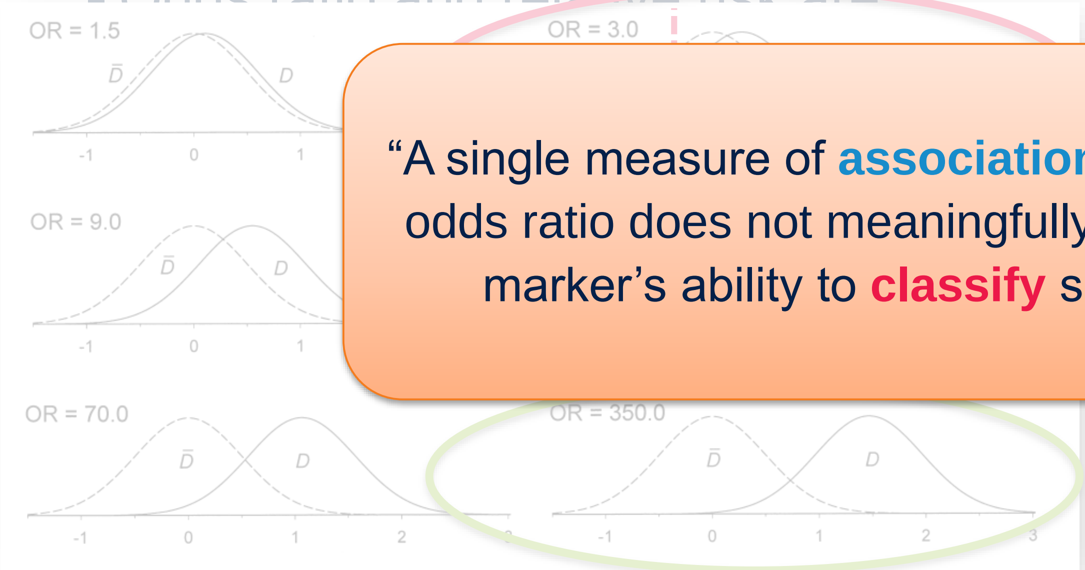
Odds ratio

Pepe MS et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epid 2004 159(9):882-90.

What's wrong with that?


Choose clinical issue & data
Test a bunch of hypothesized predictors
Generate prediction model
Prevent complications; save lives

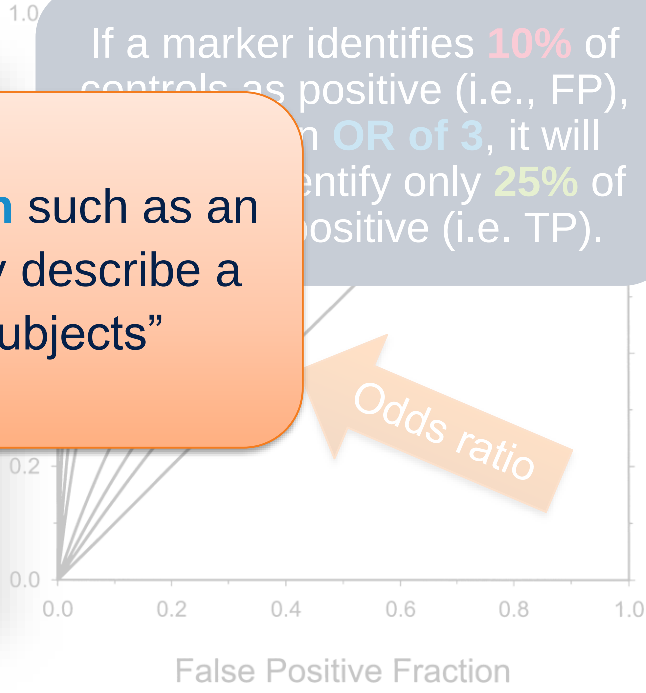
- Odds ratio and relative risk are a



$$\frac{1\text{-FPF}}{\text{FPF}}$$

"A single measure of **association** such as an odds ratio does not meaningfully describe a marker's ability to **classify** subjects"

If a marker identifies **10%** of controls as positive (i.e., FP), n **OR of 3**, it will entify only **25%** of ositive (i.e. TP).



Pepe MS et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epid 2004 159(9):882-90.

What are we left with?

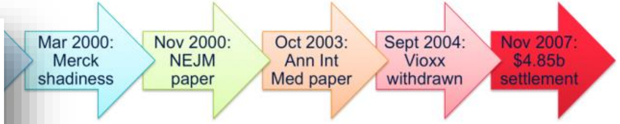
Can't
- Assume a decent OR will give you a useful prediction model or save lives

Can't
- Assume a highly significant point estimate will be meaningful

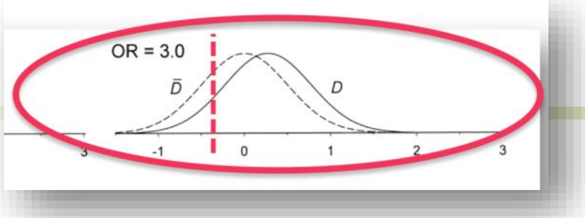
Can't
- Assume a non-significant point estimate isn't meaningful

Can't
- Really do **anything** with a p-value alone *anyway*, apparently

- "Personally, the writer prefers to set a low standard of significance at the 5 per cent point, *and ignore entirely all results which fail to reach this level*." Fisher RA, 1926


Mar 2000: Merck shadiness → Nov 2000: NEJM paper → Oct 2003: Ann Int Med paper → Sept 2004: Vioxx withdrawn → Nov 2007: $4.85b settlement

Materiality cannot be reduced to a bright-line rule.



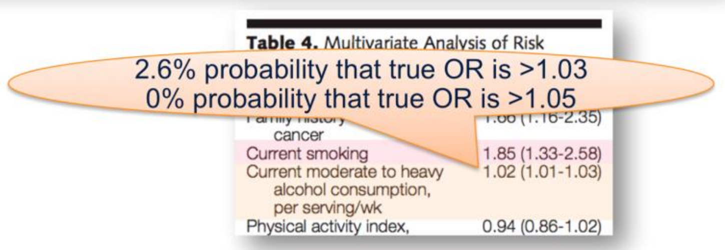
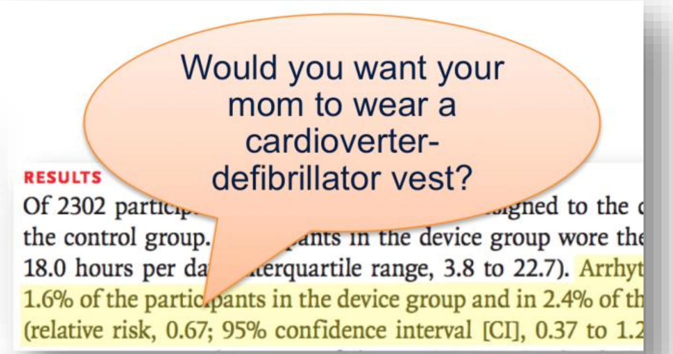


Table 4. Multivariate Analysis of Risk

Family history cancer	1.66 (1.16-2.35)
Current smoking	1.85 (1.33-2.58)
Current moderate to heavy alcohol consumption, per serving/wk	1.02 (1.01-1.03)
Physical activity index,	0.94 (0.86-1.02)

2.6% probability that true OR is >1.03
0% probability that true OR is >1.05


Would you want your mom to wear a cardioverter-defibrillator vest?

RESULTS

Of 2302 particip… …signed to the c… the control group. …ants in the device group wore the… 18.0 hours per da… …terquartile range, 3.8 to 22.7). Arrhyt… 1.6% of the participants in the device group and in 2.4% of th… (relative risk, 0.67; 95% confidence interval [CI], 0.37 to 1.2…

~~Now I feel frightened and powerless.~~

- Start to use our guts – our judges of “beer significance” – or more complex analyses translating effects into costs, lives, or function
- Demand effect sizes!
  - Change your language: “How much”, not “Does it”

Things we can “feel” and understand

- Give as much attention to bias and study design as you do to point estimates and confidence intervals
  - Were the design and analysis good? Statistical significance becomes *secondary.*

Works cited

- Historical literature
  - Fisher RA. The arrangement of field experiments. Journal of the Ministry of Agriculture 1926; 33:503-13.
  - Fisher RA. Statistical methods for research workers. Oliver & Boyd (Edinburgh), 1925
- Recent primary literature
  - Siontis GCM & Ioannidis JPA. Risk factors and interventions with statistically significant tiny effects. Int J Epidemiol 2011 Jul; 40:1292-1307.
  - Pepe MS et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epid 2004 159(9):882-90.
- Analysis
  - Ziliak ST. The *Validus Medicus* and a new gold standard. Lancet 2010 Jul 31; 376(9738): 324-5
  - Ziliak ST & McCloskey DN. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. Univ of Michigan Press, 2008
  - Wasserstein & Lazar, Am Stat 2016 70(2):129-133, 2016
  - Cowles M & Davis C. On the origins of the .05 level of statistical significance. Am Psychol 1982 May; 37(5):553-8.
  - Ziliak & McCloskey. Lady Justice v. Cult of Statistical Significance. *In* Oxford Handbook on Professional Economic Ethics. Oxford UP, 2014.
- Textbook for “estimation approach”
  - Cumming G & Calin-Jageman R. Introduction to the New Statistics. Routledge (New York, NY), 2017.

Not available, but presumably coming soon: The American Statistician special issue *Statistical Inference in the 21st Century: A World Beyond P<0.05*

UCSF