The Unreasonable Ineffectiveness of Fisherian "Tests" in Biology, and Especially in Medicine

Deirdre N. McCloskey
University of Illinois at Chicago
Chicago, IL, USA
deirdre2@uic.edu

Stephen T. Ziliak
Roosevelt University
Chicago, IL, USA
sziliak@roosevelt.edu

Abstract

Biometrics has done damage with levels of R or p or Student's t. The damage widened with Ronald A. Fisher's victory in the 1920s and 1930s in devising mechanical methods of "testing," against methods of common sense and scientific impact, "oomph." The scale along which one would measure oomph is particularly clear in biomedical sciences: life or death. Cardiovascular epidemiology, to take one example, combines with gusto the "fallacy of the transposed conditional" and what we call the "sizeless stare" of statistical significance. Some medical editors have battled against the 5% philosophy, as did, for example, Kenneth Rothman, the founder of *Epidemiology*. And decades ago a sensible few in education, ecology, and sociology initiated a "significance test controversy." But, grantors, journal referees, and tenure committees in the statistical sciences had faith that probability spaces can substitute for scientific judgment. A finding of $p < .05$ is deemed to be "better" for variable X than $p < .11$ for variable Y. It is not. It depends on the oomph of X and Y—the effect size, size judged in the light of how much it matters for scientific or clinical purposes. In 1995 a Cancer Trialists' Collaborative Group, for example, came to a rare consensus on effect size: 10 different studies had agreed that a certain drug for treating prostate cancer can increase patient survival by 12%. An 11th study published in the *New England Journal* in 1998 dismissed the drug. The dismissal was based on a t-test, not on what William Gosset (the "Student" of Student's t) had called, against Ronald A. Fisher's machinery, "real" error.[1]

Keywords

Bayesian analysis in medicine, biometrics, Fisher, Gosset, Jeffreys, level of p, levels of t, Rothman, statistical power in medical research, statistical significance, tests of significance

October 2, 2008; accepted September 6, 2009
Biological Theory 4(1) 2009, 44–53.

One wishes to know the probability that a biological or medical hypothesis, H, is true in view of the sadly incomplete facts of the world. It is a problem of inference, inferring the likelihood of a result from the data. If the symptoms of cholera start in the digestive system, then ingestion of something, perhaps foul water, is a probable cause. If cases of cholera in London in 1854 cluster around particular public wells, then bad water is a probable cause.

But, the statistical tests used in many sciences (though not much in chemistry or physics) do nothing to aid such judgments. The tests that were regularized or invented in the 1920s by the great statistician and geneticist Ronald A. Fisher (1890–1962) measure the probability that the facts you are examining will occur assuming that the hypothesis is true. Our point is that by itself, unless in a decision-theoretic context in which the other relevant probabilities and their substantive importance are calculated, such a test is mistaken. The mistake here is known in statistical logic as "the fallacy of the transposed conditional." If cholera is caused not by polluted drinking water but by bad air, then economically poor areas with rotting garbage and open sewers will have large amounts of cholera. They do. So, cholera is caused by bad air. If cholera is caused by person-to-person contagion, then cholera cases will often be neighbors. They are. So, cholera is caused by person-to-person contact. Thus Fisherian science.

If the rebel Chinese general Li Zicheng was in the summer of 1645 attacked by angry peasants from whom he was stealing food, he will be dead. He is dead. Therefore, says the usual procedure of significance testing, he was attacked by peasants. If the biological hypothesis, H, is true, then observations O will be observed with high statistical significance. O is observed. Therefore, H is true. But, of course, being dead is very weak evidence that Li Zicheng was attacked by peasants, considering that by some accounts he committed suicide—and after all there are many ways to die. Statistically speaking, the power of the test of the hypothesis that Li was so attacked is undefined. To be sure, being dead is "consistent with" the hypothesis that Li was attacked by peasants, as the neo-positivist rhetoric of the Fisherian argument has it. But so what? A myriad of other hypotheses, very different from the alleged cause of the general's death, such as committing suicide or catching pneumonia or breaking his neck in a fall from his horse, or dying from heartbreak after losing his campaign against the Manchus, are omitted from Fisherian procedures in the statistics-using sciences, though "consistent with" the fact of his being dead. The Fisherian procedure, at any rate when it proceeds (as it almost always does) without a loss function and a full discussion of Type-II error, neither falsifies nor confirms.

The psychologist and statistician, the late Jacob Cohen, made our point, a very old one, in his aptly entitled article, "The Earth is Round ($p < .05$)." "If a person is an American," Cohen writes, in a parody of the Fisherian logic, "then he is probably not a member of Congress. This person is a member of Congress. Therefore, he is probably not an American" (Cohen 1994: 998). Cohen is pointing out that the illogic of being probably-not-an-American is formally exactly the same as the Fisherian test of significance. And it is mistaken. The structure of the logic is hypothesized that $\mathrm{P}(O \mid H_0)$ is low; observe O in the data; conclude therefore that $\mathrm{P}(H_0 \mid O)$—the transposed conditional of the original hypothesis—is low. The argument appears at least implicitly in article after article in scientific journals, and explicitly in most statistics textbooks. It is wrong.

Cohen applied the logic to an important topic in psychiatry, the misdiagnosis of schizophrenia. In the United States, schizophrenia incidence in adults is about 2%. Like a general attacked by peasants in 1645, it is rare. Let $H_0 =$ the person is normal; $H_1 =$ the person is schizophrenic, and $O =$ the test result on the person in question is positive for schizophrenia. A proposed screening test is estimated to have at least 95% accuracy in making the positive diagnosis (discovering schizophrenia) and about 97% accuracy in declaring a truly normal case "normal." Formally stated, $\mathrm{P}(\text{normal} \mid H_0)$ is approximately 0.97, and $\mathrm{P}(\text{schizophrenic} \mid H_1) > 0.95$.

With a positive test for schizophrenia at hand, given the more than 95% assumed accuracy of the test, $\mathrm{P}(\text{schizophrenic} \mid H_0)$ is less than 5%—statistically significant, that is, at $p = 0.05$. In the face of such evidence, a person in the Fisherian mode would reject the hypothesis of "normal" and conclude that the person is schizophrenic. Then he might proceed to do all sorts of good and bad things to the "patient."

But the probability of the hypothesis, given the data, is not what has been tested. The probability that the person is *normal*, given a positive test for schizophrenia, is in truth quite strong—about 60%—*not*, as Fisherians believe, *less than* 3%, because, by Bayes' Theorem

$$
\begin{aligned}
&[\mathrm{P}(H_o \mid O) \\
&\quad = [\mathrm{P}(H_o) \cdot \mathrm{P}(\text{test wrong} \mid H_o)]/\{[\mathrm{P}(H_o) \\
&\qquad \cdot \mathrm{P}(\text{test wrong} \mid H_o)] + [P(H_1) \cdot (\mathrm{P}\ \text{test right} \mid H_1)]\} \\
&\quad = [(.98) \cdot (.03)]/[(.98) \cdot (.03) + (.02) \cdot (.95)] = .607,
\end{aligned}
$$

a humanly important difference from $p = .03$. The conditional probability of a case being "normal" though testing positively as schizophrenic is, Cohen points out, "not small—of the 50 cases testing as schizophrenic [out of an imagined population of 1000 people tested], 30 are false positives, actually normal, 60% of them!" (1994: 999).

The example shows how confused—and humanly and socially damaging—a conclusion from a Fisherian 5% science can be. One of us has a good friend who as a child in the psychiatry–spooked 1950s was diagnosed as schizophrenic.

The friend has shown since then no symptom of the disease. But the erroneous diagnosis—an automatic result of the fallacy of the transposed conditional—has kept him in a state of dull terror ever since. Imagine in other arenas, with similarly realistically low priors, the damage done by the transposed conditional—in scientific work or diet pills or social welfare policy or commercial advertising or the market in foreign exchange. Once one considers the concrete implications of such a large diagnostic error, such as believing that 3% of adults tested for schizophrenia are not-schizophrenic when the truth is that 60% of them are not-schizophrenic, and realizes that, after all, this magnitude of diagnostic error is governing NASA and the departments of cardiovascular disease and breast cancer and HIV health policy, one should perhaps worry.

Part of the problem historically was another campaign of Fisher's, following the elder Pearson, Karl: an attempt to kill off Bayes' Theorem. By contrast, the inventor in 1908 of the t-test for small samples, the Guinness brewer and theoretical statistician William Sealy Gosset, was a lifelong Bayesian. He defended Bayesian methods against all comers—Karl Pearson, Fisher, Karl's son Egon Pearson, Jerzy Neyman (e.g., Gosset 1915, 1922 cited in Pearson 1990: 26–27). Gosset in fact used Bayes' Theorem in his revolutionary papers of 1908, and crucially so in "The Probable Error of a Correlation Coefficient." In 1915 he wrote to the elder Pearson: "If I didn't fear to waste your time I'd fight you on the *a priori* probability and give you choice of weapons! But I don't think the move is with me; I put my case on paper last time I wrote and doubt I've much to add to it" (September 1). Gosset was courageous, but in all his fights mild and self-deprecating, including for Bayes' methods. In the warrior culture of hardboiled-dom in the 1910s and 1920s (the Great War mattered) he was not forceful enough.

Fisher was to a great deal more forceful, and wholly intolerant of "inverse probability" (Fisher 1922, 1926, 1956; cf. Zabell 1989). In Fisher's campaigns for maximum likelihood and his own notion of "fiducial probability" (one of the few campaigns of Fisher's that failed), he tried to kill off prior and posterior probability, and—at least with the mass of research workers as against the few high brows—he succeeded. Egon Pearson and Jerzy Neyman were at first persuaded by Fisher to turn from Bayes' Theorem (Pearson 1966: 9, in David 1966). But Pearson later in life, after Fisher died, reverted to his original position: "Today in many circles," he said, "the current vogue is a neo-Bayesian one, which is of value because it calls attention to the fact that, in decision making, prior information must not be neglected" (Pearson 1990: 110). Of course.

In 1963, the geophysicist, astronomer, and mathematical statistician Harold Jeffreys wrote the following:

> Whether statisticians like it or not, their results are used to decide between hypotheses, and it is elementary that if p entails q, q does not necessarily entail p. We cannot get from "the data are unlikely given the hypothesis" to "the hypothesis is unlikely given the data" without some additional rule of thought. Those that reject inverse probability have to replace it by some circumlocution, which leaves it to the student to spot where the change of data has been slipped in[, in] the hope that it will not be noticed. (Jeffreys 1963: 409)

The Five Percenter longs to find a body of data "significant and consistent with" some hypothesis. The motive is by itself blameless. But Jeffreys noted that the sequence of the Five Percenter's search procedure is backwards and paradoxical (Jeffreys 1963: 409). The Five Percenter is looking at the wrong thing in the wrong way.

In the 1994 volume of the *American Journal of Epidemiology*, David A. Savitz, Kristi-Anne Tolo, and Charles Poole examined 246 articles published in the *Journal* around the years 1970, 1980, and 1990. The articles were divided into three categories: infectious disease epidemiology, cancer epidemiology, and cardiovascular disease epidemiology. Each category contained for each date a minimum of 25 articles. The main findings are presented in a Figure 4, "Percent of articles published in the *American Journal of Epidemiology* classified as partially or completely reliant on statistical significance testing for the interpretation of the study results, by topic and time period" (Savitz et al. 1994: 1050). The findings are not surprising. The study shows that in 1990 some 60% to 70% of all cardiovascular and infectious disease epidemiologists relied exclusively on statistical significance as a criterion of epidemiological importance, as though fit were the same thing as importance. A larger share rely on the fallacy of the transposed conditional. The abuse was worse in 1990 than earlier.

The cancer researchers were less enchanted with statistical significance than cardiological and infectious disease researchers were, but did not reach standards of common sense. Savitz, Tolo, and Poole found that after a 60% reliance on a mere statistical significance in the early 1970s, the abuse of p-values by cancer researchers actually fell. We don't know why. Maybe too many people had died. Still, 40% of all the cancer research articles in 1990 relied exclusively on Fisher's Rule of Two (1994: 1050).

In epidemiology, then, the "sizeless stare," as we call it, of statistical significance is relatively recent, cancer research being an exception. In 1970 only about 20% of all articles on infectious disease epidemiology relied exclusively on tests of statistical significance. Confidence intervals and power calculations were of course absent. But epidemiology was not then an entirely statistical science. Only about 40% of all empirical articles in infectious disease epidemiology employed some kind of statistical test. But significance took hold, and by 1980 some 40% relied exclusively on the tests (compare our "Question 16" in economics, where in the 1980s it was about 70%). And by 1990, most subfields of epidemiology had like

economics and psychology become predominately Fisherian. Statistical significance came to mean "epidemiological significance." Statistical *in*significance came to mean "ignore the results."

Douglas G. Altman, a statistician and cancer researcher at the Medical Statistics Laboratory in London has been watching the use of medical statistics, and especially the deployment of significance testing, for 20 years. In 1991 Altman published an article called "Statistics in Medical Journals: Developments in the 1980s." The article appeared in *Statistics in Medicine*. Altman's experience had been similar to ours in economics. At conferences and seminars and the like Altman's colleagues were convinced that the abuse of *t*-testing had by the 1980s abated, and was practiced only by the less competent medical scientists. Any thoughtful reader of the journals knew that such claims were false. To bias the results in favor of the defenders of the status quo Altman examined the first 100 "original articles" published in the 1980s in the *New England Journal of Medicine*. These were new and full-length research articles based on never-before released or published data from clinical studies or other methods of observation. Altman's sample design was meant to replicate for comparative purposes an earlier study by Emerson and Colditz 1983, who studied the matter in 1978–1979 (Altman 1991: 1899).

The Findings

> It is my impression that the trends noted by Felson et al. have continued throughout the 1980s. . . . The obsession with significant *p* values is seen in several other ways:
>
> (1) Reporting of [statistically] significant results rather than those of most importance (especially in abstracts).
> (2) The use of hypothesis tests when none is appropriate (such as for comparing two methods of measurements or two observers).
> (3) The automatic equating of statistically significant with clinically important, and non-significant with non-existent.
> (4) The designation of studies that do or do not "achieve" significance as "positive" or "negative" respectively, and the common associated phrase "failed to reach statistical significance". . . . A review [by other investigators <who>] of 142 articles in three general medical journals found that in almost all cases (1076/1092) researchers' interpretations of the "quantitative" (that is, clinical) significance of their results agreed with statistical significance. Thus across all medical areas and sample size *p* rules, and $p < 0.05$ rules most. It is not surprising if some editors share these attitudes, as most will have passed through the same research phase of their careers and some are still active researchers. (Altman 1991: 1906)

Altman was not surprised when he found in medicine, as we were not surprised in economics, that his colleagues were deluding themselves. "I noted in the first issue of *Statistics in Medicine* that most journals gave much more attention to the format of references in submitted articles than they gave to the statistical content," Altman wrote. "This remains true" (Altman 1991: 1900). Editors are much exercised, he observed with gentle sarcasm, over whether to use "P, p, *P*, or *p* values" (1991: 1902)—but pay no heed to oomph. "It is impossibly idealistic," Altman believed, "to hope that we can stop the misuse of statistics, but we can apply a tourniquet . . . by continuing to press journals to improve their ways" (1991: 1908).

Steven Goodman, in a meaty pieces on the "*p*-value fallacy" published in the *Annals of Internal Medicine*, observed ruefully, "biological understanding and previous research play little formal role in the interpretation of quantitative results." That is, Bayes' Theorem is set aside, as is the total quality management of medical science, the seeing of results in their context of biological common sense. "This [narrowly Fisherian] statistical approach," Goodman writes, "the key components of which are P values and hypothesis tests, is widely perceived as a mathematically coherent approach to inference. There is little appreciation in the medical community that the methodology is an amalgam of incompatible elements (Goodman 1992, 1993, 1999a: 995, 1999b)."

Altman, Savitz, Goodman, and company are not singletons. According to Altman, between 1966 and 1986 fully 150 articles were published criticizing the use of statistics in medical research (Altman 1991: 1897). The studies agreed that R. A. Fisher significance in medical science had become the nearly exclusive technique for making a quantitative decision and that statistical significance had become in the minds of medical writers equated increasingly, and erroneously, with clinical significance.

As early as 1978 the situation was sufficiently dire that two contributors to the *New England Journal of Medicine*, Drummond Rennie and Kenneth J. Rothman, published op-ed pieces in the journal pages about the matter (Rennie 1978; Rothman 1978). Rennie, the deputy editor of the journal—and in 2006 the deputy editor of the *Journal of the American Medical Association*—was not critical of his colleagues' practice. But Rothman, who was a young associate professor at Harvard, and the youngest member of the editorial board, blasted away. In "A Show of Confidence," he made a crushing case for measuring *clinical* significance, not statistical significance. Citing the Freiman et al. (1978) article on "71 Negative Clinical Trials," Rothman argued that the measurement and interpretation of size of effects, confidence intervals, and examination of power functions with respect to effect size (à la Freiman et al. by graphical demonstration) was the better way forward. Rothman—an epidemiologist and biostatistician with a life-long interest in the rhetoric of his fields—wanted secretly to ban the *t*-test altogether. Rennie and the other editors decided on a different solution. Original articles would be subjected to a pre-publication screening by a professional statistician. Rothman was at first hopeful, thinking statistical review would repair the *Journal*. The director of statistical reviews was well chosen—the late Frederick Mosteller

(1916–2006), the founder of Harvard's Statistics Department and a giant of 20th-century data analysis. But Mosteller was only the director, not the worker. Rothman tells us that he as the inside critic and Mosteller as the outside director had not been able to do anything together to raise the standards (Mosteller to Ziliak and McCloskey, University of Chicago, 21 May 2005; KJ Rothman to Ziliak, 30 January 2006). The problem with pre-publication statistical review, of course, is that the articles go not to the Rothmans and Mostellers and Kruskals but out to Promising Young Jones in the outer office dazzled by his recently mastered 5% textbooks. An example nowadays is the "Statistical Analysis Plan" or, aptly acronymized, "SAP," which lays down the minimum statistical criteria considered acceptable by the Food and Drug Administration.

Like Gosset, Jeffreys, and Zellner, Rothman doubted the philosophical grounding of *p* values (Rothman 1990: 334). As Jeffreys put the following:

> If P is small, that means that there have been unexpectedly large departures from prediction [under the null hypothesis]. But why should these be stated in terms of P? The latter gives the probability of departures, measured in a particular way, equal to *or greater than* the observed set, and the contribution from the actual value [of the test statistic] is nearly always negligible. *What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure. On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law [or null hypothesis], not against it. The same applies to all the current significance tests based on P integrals. (Jeffreys 1961, quoted by Zellner 1984: 288; emphasis in original; editorial insertions by Zellner)

Rothman complained in his editorial in the *New England Journal* that Fisherian "testing . . . is equivalent to funneling all interest into the precise location of one boundary of a confidence interval" (Rothman 1978: 1363). In 1986 the situation was the same: "Declarations of 'significance' or its absence can supplant the need for any real interpretation of data; the declarations can serve as a mechanical substitute for thought, promulgated by the inertia of training and common practice" (Rothman 1986: 118).

Rothman then became assistant editor of the *American Journal of Public Health.* The chief editor of the *American Journal of Public Health* "seemed to be sympathetic" with Rothman's views—Rothman recalls one time when the editor backed him up in a little feud with a well-placed statistician. Still, Rothman's views hardly set journal policy, and it shows in the journal. Rothman finally found his chance when in 1990, after 15 years of quiet struggle, he started his own journal, *Epidemiology.* His editorial letter to potential authors was unprecedented:

> When writing for *Epidemiology*, you can . . . enhance your prospects if you omit tests of statistical significance. . . . In *Epidemiology*, we do not publish them at all. . . . We discourage the use of this type of thinking in the data analysis, such as in the use of stepwise regression. We also would like to see the interpretation of a study based not on statistical significance, or lack of it, for one or more study variables, but rather on careful quantitative consideration of the data in light of competing explanations for the findings. For example, we prefer a researcher to consider whether the magnitude of an estimated effect could be readily explained by uncontrolled confounding or selection biases, rather than simply to offer the uninspired interpretation that the estimated effect is "significant.". . . Misleading signals occur when a trivial effect is found to be "significant," as often happens in large studies, or when a strong relation is found "nonsignificant," as often happens in small studies. (Rothman 1990: 334)

Rothman concluded the letter by offering advice on how to publish quantitatively, epidemiologically significant figures, such as odds ratios on specific medical risks, bounded by confidence intervals.

Now with his own journal, Rothman was going to get it right. In January 1990 he and the associate editors Janet Lang and Cristina Cann published another luminous editorial, "That Confounded *P*-Value" (Lang et al. 1998). They "reluctantly" (p. 8) agreed to publish *p*-values when "no other" alternative was at hand. But they strongly suggested that authors of submitted manuscripts illustrate "size of effect" (p. 7) in "figures"—in plots of effect size lines against well-measured components.

Rothman and his associates were and are not alone, even in epidemiology. The statistician James O. Berger (2003) has recently shown how epidemiologists and other sizeless scientists go wrong with *p*-values. Use of Berger's *applet*, a public-access program, shows Rothman's skepticism to be empirically sound (http://www.stat.duke.edu/∼berger). The program simulates a series of tests, recording how often a null hypothesis is "true" in a range of different *p*-values. Berger cites a 2001 study by the epidemiologists Sterne and Davey Smith, which found that "roughly 90% of the null hypotheses in the epidemiology literature are initially true." Berger reports that even when *p* "is near 0.05, at least 72%—and typically over 90%" of the null hypotheses will be true (Sterne and Davey Smith 2001; Berger 2003: 4). Berger agrees with Rothman and the authors here that on the contrary "true" is a matter of judgment—a judgment of *epidemiological*, not mere statistical, significance. It is about the quality of the water from the wells.

Rothman's letter itself elicited no response. This is our experience, too: Many of the Fisherians, to put it bluntly, seem to be less than courageous in defending their views. Hardly ever have we seen or heard an attempt to provide a coherent—or indeed any—response to the case against null-hypothesis testing for "significance." The only published response that

Rothman can recollect in epidemiology came from J. L. Fleiss, a prominent biostatistician, in the *American Journal of Public Health* published in 1986. But Fleiss merely complained that "an insidious message is being sent to researchers in epidemiology that tests of significance are invalid and have no place in their research" (Fleiss 1986: 559). He gave no actual *arguments* for giving Fisherian practices a place in research. This is similar to our experience. Kevin Hoover and Mark Siegler offered in 2005 (published 2008, with our detailed reply) the only written response to our complaints in economics that we have seen. Courageous though it was for them to venture out in defense of the Fisherian conventions, a sterling exception to the spinelessness of their colleagues, they could offer no actual arguments (though they did catch us in a most embarrassing failure to take all the data from the *American Economic Review* in the 1990s). Hoover and Siegler merely wax wroth for many pages against our strictures.

Even the rare courageous Fisherians, in other words, do not deign to make a case for their procedures. They merely complain that the procedures are being criticized. "Other defenses of [null hypothesis significance testing]," Fidler et al. observed, "are hard to find" (Fidler et al. 2004: 121). The Fisherians, being comfortably in control, appear inclined to leave things as they are, *sans* argument. One can understand. If you don't have any arguments for an intellectual habit of a lifetime, perhaps it is best to keep quiet.

Rothman's campaign did not succeed. Fidler et al. (2004) found, as we and others have found in economics and psychology and in other fields of medicine, that epidemiology is getting worse, despite Rothman's letter. Over 88% of more than 700 articles they reviewed in *Epidemiology* (between 1990 and 2000) and the *American Journal of Public Health* (between 1982 and 2000) failed, they find, to distinguish and interpret substantive significance. In the *American Journal of Public Health*, some 90% confused a statistically significant result with an epidemiologically significant result, and equated statistical *in*significance with substantive unimportance. Epidemiology journals, in other words, performed worse than the *New England Journal of Medicine*, Rothman's training-ground as an editor.

Fidler and her coauthors (2004) observe that for decades "advocates of statistical reform in psychology have recommended confidence intervals as an alternative (or at least a supplement) to *p* values." The American Psychological Association *Publication Manual* called them in 2001 "the best reporting strategy," though few seem to be paying attention (APA *Manual* 2001: 22 in Fidler et al. 2004: 119; Fidler 2002). Since the mid-1980s, confidence intervals have been widely reported in medical journals. Unhappily, requiring the calculation of confidence intervals does not guarantee that effect sizes will be interpreted more carefully, or indeed at all. Savitz et al. find that even though 70% of articles in the *American Journal of Epidemiology* report confidence intervals "inferences are made regarding statistical significance tests, often based on the location of the null value with[out] respect to the bounds of the confidence interval" (1994: 1051). In other words, say Fidler and her coauthors, confidence intervals "were simply used to do [the null hypothesis testing ritual]" (Fidler et al. 2004: 120).

Fidler and her coauthors (2004) attempted as we have to assemble outside allies. They "sought lessons for psychology from medicine's experience with statistical reform by investigating two attempts by Kenneth Rothman to change statistical practices." They examined 594 *American Journal of Public Health* articles published between 1982 and 2000 and 110 *Epidemiology* articles published in 1990 and 2000:

> Rothman's editorial instruction to report confidence intervals and not *p* values was largely effective: In *AJPH*, sole reliance on *p* values dropped from 63% to 5%, and confidence interval reporting rose from 10% to 54%; *Epidemiology* showed even stronger compliance. However, compliance was superficial: Very few authors referred to confidence intervals when discussing results. The results of our survey support what other research has indicated: Editorial policy alone is not a sufficient mechanism for statistical reform. (Fidler et al. 2004: 119)

Rothman himself has said of his attempt to reduce *p*-value reporting in his *Epidemiology* that "my revise-and-resubmit letters . . . were not a covert attempt to engineer a new policy, but simply my attempt to do my job as I understood it. Just as I corrected grammatical errors, I corrected what I saw as conceptual errors in describing data" (quoted in Fidler et al. 2004: 121).

Fidler's team studied the *American Journal of Public Health* and *Epidemiology* before, during, and after Rothman's editorial stints; before and after the International Committee of Medical Journal Editors creation of statistical regulations encouraging the analysis of effect size; and before and after the changes to the *AJPH*'s "Instructions to Authors" encouraging the use of confidence intervals. Rothman as an assistant editor, of course, did not make policy at the journal. He made his own preferences known to authors, but ultimately he "carried out the editor's policy," which only occasionally overlapped with Rothman's ideal (Rothman to Ziliak, email communication, 27 January 2006).

Fidler et al. counted a statistical practice "present," such as what we call "asterisk biometrics," the ranking of coefficients according to the size of the *p*-value, if an article contained at least one instance of it. Their full questionnaire is similar to ours in economics (Ziliak and McCloskey 2008: 62–92), focusing on substantive as against statistical significance testing. Did "significant" mean "epidemiologically important" or "statistically significant"? Practice was recorded as ambiguous if the author or authors did not preface "significant"

with "statistically," follow the statement of significance directly with a *p*-value or test statistic, or otherwise differentiate between statistical and substantive interpretations. "Explicit power" in their checklist means "did a power calculation." "Implicit power" means some mention of a relationship between sample size, effect size, and statistical significance was made—for example, a reference to small sample size as perhaps explaining failure to find statistical significance. The results, alas, "Of the 594 AJPH articles, 273 (46%) reported NHST. In almost two thirds of the cases 'significant' was used ambiguously. Only 3% calculated power and 15% reported 'implied power.' . . . An overwhelming 82% of NHST articles had neither an explicit nor implicit reference to statistical power, even though all reported at least one non-significant result."

Fifty-four percent of *American Journal of Public Health* articles reported confidence intervals; 86% did in *Epidemiology*. But "Table 2 shows that fewer than 12% of *AJPH* articles with confidence intervals interpreted them and that, despite fully 86% of articles in *Epidemiology* reporting confidence intervals, interpretation was just as rare in that journal" (Fidler et al. 2004: 122). The situation, they find, did not improve with the years. The authors usually did not refer in their texts to the width of their confidence intervals, and did not discuss what is epidemiologically or biologically or socially, or clinically significant in the size of the effect. In other words, during the past two decades more than 600 of some 700 articles published in the leading journals of public health and epidemiology showed no concern with epidemiological significance. Thus too economics, sociology, population biology, and other Fisherian fields.

When in 2000 Rothman left his post as editor of *Epidemiology*, confidence-interval reporting remained high—it had become common in medical journals. But in the *American Journal of Public Health* reporting of unqualified *p* "again became common." Rothman's success at *Epidemiology* appears to have been longer lasting. Still, interpretation in other journals of epidemiology is rare. "In both journals [Fidler et al. should add 'but not in *Epidemiology*'] . . . when confidence intervals were reported, they were rarely used to interpret results or comment on [substantive] precision. This rather ominous finding holds even for the most recent years we surveyed" (Fidler et al. 2004: 123). Fidler and her team confirm in thousands of tests what Savitz et al. (1994) found in the *American Journal of Epidemiology* in tens of thousands of tests and what Rossi found in 39,863 tests in psychology and speech and education and sociology, and management (Rossi 1990: 648).

The historian of medicine Richard Shyrock argued in an early paper that instruments such as the stethoscope and the X-ray machine saved some parts of medicine from the Fisherian pitfall. If one can see or hear the problem, one does not need to rely on correlations (Shyrock 1961: 228). Since 1961, though, doctors have lost many of their skills of physical assessment, even with the stethoscope (and certainly with their hands), and have come to rely on a medical literature deeply infected with Fisherianism. Shyrock's piece appeared in a special issue of *Isis* on the history of quantification in the sciences, mostly celebrating the statistical side of it. Puzzlingly, none of the contributors to the symposium mentioned the Gosset-Fisher-Neyman-Pearson-Jeffreys-Deming-Savage complex. Fisher-significance, the omission suggests, was not to be put on trial. The inference machines remained broken.

By 1988 the International Committee of Medical Journal Editors had been sufficiently pressured by the Rothmans and the Altmans to revise their "uniform requirements for manuscripts submitted to biomedical journals." "When possible," the Committee wrote, "quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid sole reliance on statistical hypothesis testing, such as the use of *p* values, which fail to convey important quantitative information" (ICMJE 1988: 260). The formulation is not ideal. The "error" in question is tacitly understood to be sampling error alone, when after all a good deal of error does not arise from the smallness of samples. "Avoid sole reliance" on the significance error should be "don't commit" the significance error. The "important quantitative information" is effect size, which should have been mentioned explicitly. Still, it was a good first step, and in 1988 among the sizeless sciences was amazing.

The Requirements—on which at a formative stage Rothman among others had contributed an opinion—were widely published. They appeared for instance in the *Annals of Internal Medicine*—where later the Vioxx study was published—and in the *British Medical Journal*. More than 300 medical and biomedical journals, including the *American Journal of Public Health*, notified the International Committee of their willingness to comply with the manuscript guidelines (Fidler et al. 2004: 120). But the Requirements have not helped.

The essence of the problem of reform—and the proof that we need to change academic and institutional incentives, including criteria for winning grants—is well illustrated in a study of "temptation to use drugs" published in the *Journal of Drug Issues*. The study was financed by the Centers for Disease Control. It was authored by two professors of public health at Emory University (one of them was an Associate Dean for Research), and a third professor, a medical sociologist at Georgia State University. The study was conducted in Atlanta between August 1997 and August 2000. Its subjects were African-American women—mothers and their daughters—living in low-income neighborhoods of Atlanta (Klein et al. 2003: 167). The dependent variable was "frequency-of-[drug] use and times-per-day" multiplied for each drug type and summed by month. In the 125 women studied the value of the dependent variable ranged from zero

to 910, that is, from zero to an appalling 30 drug doses a day.

Statistical Significance Decides Everything

Initially, each of the temptations-to-use drugs variables was entered into simple regression equations, to determine if they were *statistically significant* predictors of the outcome measure. Next, those found to be *related to amount of drug use* reported were entered simultaneously into a stepwise multiple regression equation. . . . Next, the bivariate relationships between the other predictor variables listed earlier were examined one by one, using Student's t tests whenever the independent variable was dichotomous. . . . Items that were found to be marginally—or statistically—significant predictors in these bivariate analyses were selected for entry into the multivariate equation. (Klein et al. 2003: 169, 170)

The authors do at least report mean values of the temptations to use drugs—a first step in determining substantive significance. For example, they report that women were "least tempted to use drugs when they were: talking and relaxing (74.0%), experiencing withdrawal symptoms (73.3%), [and] waking up and facing a difficult day (70.7%). And they would be tempted "quite a bit" or "a lot" when they were "with a partner or close friend who was using drugs (38.5%)" or when "seeing another person using and enjoying drugs (36.1%)" (Klein et al. 2003: 170). Here is how they presented their findings:

When examined in bivariate analyses, 15 of the 16 temptations-to-use drugs items were found to be associated [that is, the authors assert, statistically significantly related with; *not* substantively significantly related] with actual drug use. These were: while with friends at a party ($p < .001$), while talking and relaxing ($p < .001$), while with a partner or close friend who is using drugs ($p < .001$), while hanging around the neighborhood ($p < .001$), when happy and celebrating ($p < .001$), when seeing someone using and enjoying drugs ($p < .05$), when waking up and facing a tough day ($p < .001$), when extremely anxious and stressed ($p < .001$), when bored ($p < .001$), when frustrated because things are not going one's way ($p < .001$), when there are arguments in one's family ($p < .05$), when in a place where everyone is using drugs ($p < .001$), when one lets down concerns about one's health ($p < .05$), when really missing the drug habit and everything that goes with it ($p < .010$), and while experiencing withdrawal symptoms ($p < .01$). (Klein et al. 2003: 171–172)

"The only item that was not associated with the amount of drugs women used," the article concluded, "was 'when one realized that stopping drugs was extremely difficult' " (Klein et al. 2003: 172). This is surely a joke, some will think, perhaps a belated retaliation for the 1990s *Social Text* scandal, in which a scientist posed as a postmodern theorist in order to expose its intellectual pretense. It's not. It's normal science in biology, medicine, psychiatry, economics, psychology, sociology, education, and many other fields. But what is the scientific or policy oomph of such a temptations-to-use-drugs study? Everything is "significant."

In September 1978 Jennie A. Freiman, Thomas C. Chalmers, Harry Smith, Jr., and Roy R. Kuebler, doctors and statistical researchers at Mt. Sinai in New York, published in the *New England Journal of Medicine* a study entitled "The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial." The abstract reads as follows:

Seventy-one "negative" randomized control trials were re-examined to determine if the investigators had studied large enough samples to give a high probability (>0.90) of detecting a 25 per cent and 50 per cent therapeutic improvement in the response. *Sixty-seven of the trials had a greater than 10 per cent risk of missing a true 25 per cent therapeutic improvement, and with the same risk, 50 of the trials could have missed a 50 per cent improvement.* Estimates of 90 per cent confidence intervals for the true improvement in each trial showed that in 57 of these "negative" trials, a potential 25 per cent improvement was possible, and 34 of the trials showed a potential 50 per cent improvement. Many of the therapies labeled as "no different from control" in trials using inadequate samples have not received a fair test. Concern for the probability of missing an important therapeutic improvement because of small sample sizes deserves more attention in the planning of clinical trials. (Freiman et al. 1978: 690; italics supplied)

Freiman, who is a specialist in obstetrics and gynecology, and her colleagues, in other words, had reanalyzed 71 articles in medical journals. Heart and cancer-related treatments dominated the clinical trials under review. Each of the 71 articles concluded that the "treatment"—for example, "chemotherapy" or "an aspirin pill"—performed no better in a clinical sense than did the "control" of nontreatment or a placebo. That is, the treatments were "insignificant."

Freiman et al. (1978) found that if the authors of the original studies had considered the power of their tests—the probability of rejecting the null hypothesis "[treatment] no different from control" as the treatment effect moves in the direction of "vast improvement"—and in conjunction with effect size, the experiments would not have ended "negatively." That is, the clinicians conducting the original studies would have found that indeed the treatment therapy was capable of producing "important therapeutic improvement."

Specifically, Freiman et al. (1978) found that if fully 50 of the 71 trials had paid attention to power and effect size and not merely to a one-sided, qualitative, yes/no interpretation of "significance," they would have *reversed* their conclusions. Astonishingly, they would have found up to "50 per cent improvement" in "therapeutic effect." The Fisherian tests of significance, the only tests employed by the original authors of the 71 studies, literally could not see the beneficial effects of the therapies under study, though staring at them.

The precise standard of improvement—the minimum standard of oomph the authors set—is a "reduction in mortality

from the control [group] mortality rate," a baseline rate of 29.7 per cent (Freiman et al. 1978: 691). They realize, it is not a very strict standard of medical oomph. They are bending over backwards not to find their colleagues mistaken. Like Gosset, they want to give their Fisherian colleagues the benefit of the doubt.

Yet, they found that 70% of the alleged "negative" trials were stopped, missing an opportunity to reduce the mortality of their patients by up to 50%. Of the patients who were prescribed sugar pills or otherwise dismissed, in other words, about 30% died unnecessarily. In one typical article the authors in fact missed at $\alpha = 0.05$ a 25% reduction in mortality with probability of about 0.77 and, at the same level of Type-I error, a 50% reduction with probability about 0.42 (Freiman et al. 1978: 691).

Each of the 71 experiments was shut down on the belief that a 30% death rate was equally likely with the sugar pill (or whatever the control was) and with the treatment therapy, spurning opportunities to save lives. The article shows that in the original experiments as few as 15% of the patients receiving the treatment therapy would have died had the experiment continued—half as many as actually died.

We agree with Rothman that the article seems in the end to lose contact with the effect size, at times advising that power be treated "dichotomously" and rigidly irrespective of effect size (Rothman and Ziliak, personal interview, 30 January 2006). "Important information can be found on the edges," as Rothman put it. But overall, Rothman and we agree that it's a crushing piece. The oomph-laden content of their work is exemplary. Freiman and her colleagues note that the experiments and 71 oomph-less, premature truncations were conducted by leading medical scientists. Such premature results were published in *Lancet*, the *British Medical Journal*, the *New England Journal of Medicine*, the *Journal of the American Medical Association*, and other elite journals. Effective treatments for cardiovascular and cancer, and gastrointestinal patients were abandoned because they did not attain statistical significance at the 5% or better level.

In 1995 the authors of 10 independent and randomized clinical trials involving thousands of patients in treatment and control groups had come to an agreement on an effect size. Consensus on a mere direction of effect—up or down, positive or negative—is rare enough in science. After four centuries of public assistance for the poor in the United States and Western Europe for example, economists do not speak with one voice on the direction of effect on labor supply exerted by tax-financed income subsidies (Ziliak and Hannon 2006). Medicine is no different. Disagreement on the direction of effect—let alone the size of effect—is more rule than exception.

So the Prostate Cancer Trialists' Collaborative Group was understandably eager to publicize the agreement. Each of the 10 studies showed that a certain drug "flutamide"—for the treatment of prostate cancer—could increase the likelihood of patient survival by an average of 12% (the 95% confidence interval in the pooled data put an upper bound on flutamide-enhanced survival at about 20% [Rothman et al. 1999]). Odds of 5 in 100 are not the best news to deliver to a prostate patient. But if castration followed by death is the next best alternative, a noninvasive 12% to 20% increase in survival sounds good.

But in 1998 the results of still another, eleventh trial were published in the *New England Journal of Medicine* (Eisenberger et al. 1998: 1036–1042). The authors of the new study found a similar size effect. But when the two-sided p-value for their odds ratio came in at 14, they dismissed the efficacious drug, concluding "no clinically meaningful improvement" (pp. 1036, 1039). Kenneth Rothman, Eric Johnson, and David Sugano (1999) examined the individual and pooled results of the 11 separate studies, including the study conducted by Eisenberger et al.

> One might suspect that [Eisenberger et al.'s] findings were at odds with the results from the previous ten trials, but that is not so. From 697 patients randomized to flutamide and 685 randomized to placebo, Eisenberger and colleagues found an OR of 0·87 (95% CI 0·70–1·10), a value nearly identical to that from the ten previous studies. Eisenberger's interpretation that flutamide is ineffective was based on absence of statistical significance. (Rothman et al. 1999: 1184)

Rothman and his coauthors depict the flutamide effect graphically in a manner consistent with a Gosset-Jeffreys-Deming approach. That is, they pool the data of the separate studies and plot the flutamide effect (measured by an odds ratio, or the negative of the survival probability in a hazard function) against a p-value function. With the graphical approach, Rothman and his coauthors are able to show pictorially how the p-values vary with increasingly positive and increasingly negative large effects of flutamide on patient survival. And what they show is substantively significant:

> Eisenberger's new data only reinforce the findings from the earlier studies that flutamide provides a small clinical benefit. Adding the latest data makes the p value function narrower, which is to say that the overall estimate is now more precise, and points even more clearly to a benefit of about 12% in the odds of surviving for patients receiving flutamide.

Rothman et al. (1999) conclude, "the real lesson" from the latest study is "that one should eschew statistical significance testing and focus on the quantitative measurement of effects."

That sounds right. Statistical significance is spoiling biological science, is undermining medical treatment, and is killing people. It is leaving a great deal, shall we say, unexplained.

Note

1. This paper is a revision of chapters 14–16 in Ziliak and McCloskey 2008.

References

Altman DG (1991) Statistics in medical journals: Developments in the 1980s. Statistics in Medicine 10: 1897–1913.

American Psychological Association (APA) 1952 to 2001 [revisions] Publication Manual of the American Psychological Association. Washington, DC: APA.

Berger JO (2003) Could Fisher, Jeffreys, and Neyman have agreed on testing? Statistical Science 18: 1–32.

Cohen J (1994) The earth is round ($p < 0.05$). American Psychologist 49: 997–1003.

David FN, ed (1966) Research Papers in Statistics: Festschrift for J. Neyman. London: Wiley.

Eisenberger MA, Blumenstein BA, Crawford ED, Miller G, McLeod DG, Loehrer PJ, Wilding G, Sears K, Culkin DJ, Thompson IM, Bueschen AJ, Lowe BA (1998) Bilateral orchiectomy with or without flutamide for metastatic prostate cancer. New England Journal of Medicine 339: 1036–1042.

Fidler F (2002) The fifth edition of the APA Publication Manual: Why its statistics recommendations are so controversial. Educational and Psychological Measurement 62: 749–770.

Fidler F, Thomason N, Cumming G, Finch S, Leeman J (2004) Editors can lead researchers to confidence intervals but they can't make them think: Statistical reform lessons from medicine. Psychological Science 15: 119–126.

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society A 222: 309–368.

Fisher RA (1926) Bayes' Theorem. Eugenics Review 18: 32–33.

Fisher RA ([1956] 1959) Statistical Methods and Scientific Inference, 2nd ed. New York: Hafner.

Fleiss JL (1986) Significance tests do have a role in epidemiological research: Reaction to AA Walker. American Journal of Public Health 76:559–600.

Freiman JA, Chalmers T, Smith H, Kuebler RR (1978) The importance of beta, the type II error and sample design in the design and interpretation of the randomized control trial: Survey of 71 negative trials. New England Journal of Medicine 299: 690–694.

Goodman S (1999a) Toward evidence-based medical statistics. 1: The p-value fallacy. Annals of Internal Medicine 130: 995–1004.

Hoover K, Siegler M (2008) Sound and fury: McCloskey and significance testing in economics. Journal of Economic Methodology 15: 1–37.

International Committee of Medical Journal Editors (ICMJE) (1988) Uniform requirements for . . . statisticians and biomedical journal editors. Statistics in Medicine 7: 1003–1011.

Jeffreys H (1963) Review of L. J. Savage, et al., The Foundations of Statistical Inference (Methuen, London and Wiley, New York, 1962). Technometrics 5: 407–410.

Klein H, Elifson KW, Sterk CE (2003) Perceived temptation to use drugs and actual drug use among women. Journal of Drug Issues 33: 161–192.

Lang JM, Rothman KJ, Cann CI (1998) That confounded p-value. Epidemiology 9: 7–8.

Pearson ES (1990) [posthumously published by Plackett RL, Barnard GA, eds] 'Student': A Statistical Biography of William Sealy Gosset. Oxford: Clarendon Press.

Rennie D (1978) Vive la Difference ($p < 0.05$). New England Journal of Medicine 299: 828–829.

Rossi J (1990) Statistical power of psychological research: What have we gained in 20 years? Journal of Consulting and Clinical Psychology 58: 646–656.

Rothman KJ (1978) A show of confidence. New England Journal of Medicine 299: 1362–1363.

Rothman KJ (1986) Modern Epidemiology. New York: Little, Brown.

Rothman KJ (1990) Writing for epidemiology. Epidemiology 9: 333–337.

Rothman KJ, Johnson ES, Sugano DS (1999) Is flutamide effective in patients with bilateral orchiectomy? Lancet 353: 1184.

Savitz DA, Tolo K, Poole C (1994) Statistical significance testing in the American Journal of Epidemiology, 1970–1990. American Journal of Epidemiology 139: 1047–1052.

Shyrock RH (1961) The history of quantification in medical science. Isis 52: 215–237.

Sterne JAC, Davey Smith G (2001) Sifting the evidence—What's wrong with significance tests? British Medical Journal 322: 226–231.

Zabell S (1989) R. A. Fisher on the history of inverse probability. Statistical Science 4: 247–263.

Zellner A (1984) Basic Issues in Econometrics. Chicago: University of Chicago Press.

Ziliak ST, Hannon J (2006) Public assistance: Colonial times to the 1920s. In Historical Statistics of the United States. (Carter SB, Gartner SS, Haines MR, Olmstead AL, Sutch R, Wright G, eds). New York: Cambridge University Press.

Ziliak ST, McCloskey DN (2008) The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. Ann Arbor, MI: University of Michigan Press.