Nos. 17-1140(L), 17-1136, 17-1137, 17-1189

IN THE

United States Court of Appeals

FOR THE FOURTH CIRCUIT

IN RE: LIPITOR (ATORVASTATIN CALCIUM) MARKETING, SALES PRACTICES AND PRODUCTS LIABILITY LITIGATION (NO II) MDL 2502

PLAINTIFFS APPEALING CASE MANAGEMENT ORDER 100; JUANITA HEMPSTEAD; PLAINTIFFS APPEALING CASE MANAGEMENT ORDER 99; PLAINTIFFS APPEALING CASE MANAGEMENT ORDER 109,

Plaintiffs-Appellants,

—v.—

PFIZER, INCORPORATED; MCKESSON CORPORATION; GREENSTONE, LLC; PFIZER INTERNATIONAL LLC,

Defendants-Appellees.

On Appeals from the United States District Court
for the District of South Carolina (Charleston),
Nos. 2:14-mn-02502-RMG, 2:14-cv-01879-RMG

BRIEF FOR *AMICI CURIAE* CARL CRANOR, DIERDRE N. McCLOSKEY, AND STEPHEN T. ZILIAK IN SUPPORT OF PLAINTIFFS-APPELLANTS

Christopher J. McDonald
Counsel of Record
Christopher D. Barraza
LABATON SUCHAROW LLP
140 Broadway, 34th Floor
New York, New York 10005
Telephone: (212) 907-0700
Fax: (212) 818-0477
Email: cmcdonald@labaton.com

Counsel for Amici Curiae
Carl Cranor, Dierdre N. McCloskey,
and Stephen T. Ziliak

April 28, 2017

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Page(s)

CASES

STATUTES

OTHER AUTHORITIES

INTEREST OF *AMICI CURIAE*[1]

Amici are professors and academics who teach and write on philosophic issues concerning risks, science and the law, the use of scientific evidence in legal venues, economics, statistics, and the history, philosophy, and rhetoric of economics and statistics as used in business, medicine, and other statistical sciences. *Amici* have no stake in the outcome of this case. They are filing this brief solely as individuals and not on behalf of the institutions with which they are affiliated.

Carl Cranor is a Distinguished Professor of Philosophy and a Faculty Member of the Environmental Toxicology Program at the University of California, Riverside. He has published six books and roughly 80 articles. Two of his books—*Regulating Toxic Substances: A Philosophy of Science and the Law* (Oxford, 1993) and *Toxic Torts: Science, Law and the Possibility of Justice*, (Cambridge, 2006, 2008, 2d Ed. 2016) focus specifically on the use of science in toxic tort law.

In 2014 Professor Cranor was named the national Romanell-Phi Beta Kappa Professor of Philosophy for distinguished achievement and substantial

[1] Pursuant to Rule 29(a)(4)(E), counsel for *amici* represent that no counsel for a party authored this brief in whole or in part and that none of the parties or their counsel, nor any other person or entity other than *amici* or their counsel, made a monetary contribution intended to fund the preparation or submission of this brief. Counsel for *amici* also represent that all parties have consented to the filing of this brief.

contributions to public understanding of philosophy. Professor Cranor has been American Council of Learned Societies Fellow, a Yale Master of Studies in Law Fellow, a Congressional Fellow, and is an elected Fellow of the American Association for the Advancement of Science and the Collegium Ramazzini, an international scientific society "dedicated to advancing the study of occupational and environmental health issues around the world." Twice he was invited as a Gordon Research Conference speaker on Science and Technology Policy. The National Science Foundation and the University of California's Toxic Substances Research and Teaching Program supported his research. The state of California invited his service on science advisory panels for Proposition 65, the Electric and Magnetic Fields Program, the Nanotechnology Program, and the Scientific Guidance Panel of the California Environmental Contaminant Biomonitoring Program. He received his B.A. in mathematics (minor physics) from the University of Colorado, his doctorate from the University of California, Los Angeles, and a Masters of Studies in Law from Yale Law School.

Deirdre Nansen McCloskey is emerita distinguished professor of economics at the University of Illinois at Chicago. Author of 17 books and 400 scientific papers, she has been writing about statistical significance since the mid-1980s, and is well known in the profession and beyond for her advocacy for common sense in using statistics. A Harvard B.A. and Ph.D., she taught for a dozen years at the University of Chicago in economics, and was tenured there.

Stephen T. Ziliak is a Professor of Economics and Faculty Member of the Social Justice Studies Program at Roosevelt University, Chicago, where as a faculty member he served on the Board of Trustees from 2010 to 2013; Conjoint Professor, Faculty of Business and Law, University of Newcastle (Australia); Faculty Affiliate, Graduate Program in Economics, Colorado State University, and Faculty Member of The Angiogenesis Foundation (Cambridge, MA). He has been a Visiting Professor of Economics, Statistics, Law, Rhetoric, Justice, Social Welfare, and Methodology at leading universities of the United States, Belgium, Denmark, England, France, Ireland, Northern Ireland, Turkey, and the Netherlands. He is the lead author of The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives (2008), a best-selling and critically-acclaimed book from the University of Michigan Press (with Deirdre N. McCloskey). His research on practical versus statistical significance has appeared in many leading journals, such as The Lancet, Journal of Economic Literature, Journal of Economic Perspectives, Biological Theory, International Journal of Forecasting, The American Statistician, Review of Behavioral Economics, and Journal of Wine Economics. A member of many international committees, he was a lead author of the historic American Statistical Association "Statement on Statistical Significance and P-Values."

INTRODUCTION

Pfizer, Inc. manufactures Lipitor (atorvastatin calcium). Lipitor is in a class of drugs commonly known as "statins." Statins lower "bad" cholesterol and triglycerides and can raise "good" cholesterol; they are often prescribed to lower the risk of stroke, heart attack, or other heart complications. The scientific community widely recognizes that statins also increase the risk of new-onset diabetes.

At issue is whether the district court properly understood and applied basic statistical principles when excluding the opinions of Plaintiffs' experts on general causation (the capacity of Lipitor to cause diabetes) in all cases, and Plaintiffs' expert on specific causation (whether Lipitor caused a particular plaintiff's diabetes) in the *Hempstead* bellwether case. We respectfully submit that it did not. *Amici* therefore wish to ensure that the Court properly distinguishes 'practical' from mere 'statistical' significance when deciding this appeal.

SUMMARY OF ARGUMENT

The analyses in Case Management Orders ("CMO") 54 and 68—in particular the analysis on page 10 of CMO 68—suggest the district court did not appreciate that statistical significance testing is more complex than simply calculating numbers and determining whether a 5% standard has been met. Courts need to understand the possibilities for errors that can arise from the randomness of biological sampling.

Unless researchers conduct an "ideal" epidemiological study, there are tradeoffs—and the potential for mistakes that could influence the understanding of the results. When performing a statistical significance test, researchers must weigh the costs of accepting false hypotheses against the costs of rejecting true hypotheses. Insisting on a small chance of a *false positive error* (known as a "Type I" error) can increase the chances of a *false negative error* (known as a "Type II" error). There is an inverse relationship between the two.

Because of this relationship, an insistence for statistical significance at α=0.05 could frustrate the revelation of important information from a study. Indeed, there is nothing sacrosanct about statistical significance at 0.05, as several experts and the Reference Manual on Scientific Evidence point out. Unwavering adherence to a 0.05 standard could pose a risk that a "no effect" study result could miss adverse effects that might be present, depending upon sample size and the desired difference to be detected between controls and exposed groups. If an effect is particularly important in practical economic and/or other human terms, then the damage from failing to uncover the significance that exists in truth is particularly grave. Researchers must therefore strike the correct balance between statistical significance and practical importance.

Scientists can overcome constraints posed by fixed and low statistical significance values by using confidence intervals that provide both an idea of the magnitude of the effect and the inherent variability in the estimates. But here, too, scientists do not unwaveringly reject results that lack statistical significance at a 95% confidence level. A study that fails to reach statistical significance at a 95% confidence level may still provide important information. For instance, 90% confidence intervals, equivalent to 0.10 statistical significance values, are not uncommonly utilized to provide greater information about data.

These scientific realities stand in contrast to bright-line tests of statistical significance—like those adopted by the district court in CMOs 54 and 68—that fail to capture important nuances of applied significance testing. By routinely demanding statistical significance at α=0.05 to prevent Type I errors—without factoring in information about sample size, disease rate under investigation, and more subtle analysis—courts risk mistakes that can preclude data about causal relationships, leading to misinformed or uninformed decisions.

ARGUMENT

I. A CLARIFICATION OF HYPOTHESIS TESTING AND OF THE CONCEPT OF STATISTICAL SIGNIFICANCE IS WARRANTED IN THIS MATTER.

Epidemiological studies can be vital in assessing whether causal relationships exist between a defendant's actions or omissions and a plaintiff's injuries. That said, "checklist" emphasis on small statistical significance values can be a barrier to recognizing cause-and-effect relationships. Understanding the district court's error requires a brief explanation of hypothesis testing.

A. A hypothesis test seeks to determine whether underlying data are consistent with a null hypothesis.

Hypothesis testing is one of the most important, if not the most important, concepts in statistical analysis. At a high level, a hypothesis test is performed to determine whether data exhibit certain properties or accord with a specific statistical distribution. When scientists try to identify adverse effects, they seek to discern what the epidemiological studies show about relationships between exposure and disease. For example, they might be interested in whether benzene causes leukemia, a rare disease. If an association is shown in a study, the scientists would further want to know whether or not the association points to a causal relationship between benzene exposure and leukemia. The scientists could use hypothesis testing to determine the level of confidence under which one could conclude that benzene exposure causes leukemia.

A hypothesis test begins by posing a null hypothesis—the hypothesis the scientist wishes to test (for example, the null hypothesis that there is no difference in leukemia rates between people exposed to benzene and people not exposed to benzene). A test statistic is then calculated, assuming that the "null" hypothesis is true. The scientist then determines whether the test statistic falls into one of two subsets of values: a region under which the null hypothesis is rejected (meaning benzene exposure may actually cause increased rates of leukemia) and one under which the null hypothesis cannot be rejected (because there is insufficient evidence to conclude at some level of significance that benzene exposure increases rates of leukemia).

The stylized example above may lead one to believe that hypothesis testing is a simple dichotomous procedure of either rejecting or failing to reject a hypothesis. This is not the case. The randomness of sampling and sample size inject considerable variation that can affect study outcomes. Because statistical studies only consider samples of a larger population, a study showing a positive correlation between exposure and leukemia may be mistaken because of random chance (or other shortcomings). For example, a sample may contain a higher fraction of sick people than the true population, yielding a false positive effect of benzene exposure on leukemia. A study showing a "no effect" outcome—no correlation between exposure and leukemia—might contain a lower fraction of sick people than the true population and mistakenly fail to detect an effect of benzene exposure on leukemia.

These nuances of hypothesis testing must be considered. Because the harm from erring toward either falsely accepting or falsely rejecting a null hypothesis could be significant, these errors must be balanced accordingly.

B. Failing to reject the null hypothesis does not indicate that the effect of interest is meaningless or unimportant.

Early applications of statistical tests tended to focus on deciding whether "'chance' or 'random' error could be solely responsible for an observed association" for decision purposes.[2] Statistical hypothesis testing developed to provide a basis for decision making for research problems in "industrial and agriculture [settings], and typically involved randomized experiments … that formed the basis for a choice between two or more alternative courses of action. Such studies were designed to produce results enabling a decision to be made, and the statistical methods employed were intended to facilitate decision making."[3] From this arose the practice of declaring associations in data as "statistically significant" or "nonsignificant," an arbitrary cutoff that one should not mistakenly reject the null hypothesis more than 5% of the time.

C. There are two types of errors in hypothesis testing: Type I and Type II.

A hypothesis test can have two types of errors. Type I error (often represented by the symbol "α") occurs when one rejects a null hypothesis that is true. Type II error (often represented by the symbol "β") occurs when one does not reject a null hypothesis that is false. The chart below helps to clarify this concept.

[2] Kenneth J. Rothman & Sander Greenland, *Modern Epidemiology* 150 (3d ed. 2008).
[3] *Id*. at 151.

		Experimental Outcome	
		Reject null hypothesis	*Do not reject null hypothesis*
Truth	*Null hypothesis is false*	Correct Decision	Type II Error
	Null hypothesis is true	Type I Error	Correct Decision

The importance of Type I error to this matter is that the level of significance in a hypothesis test defines the probability of making a Type I error. Were one to perform a test comparing a medication to placebo at a 5% level of significance, the probability of *incorrectly* rejecting the null hypothesis that the medication has an effect similar to a placebo would therefore be 5%.

A Type I error, however, is not the only possible "bad" test outcome. Another would be to fail to reject the null hypothesis when in fact the medication works. This is Type II error.

When analyzing the probability of a Type II error, statisticians will sometimes refer to the "power" of the statistical test, which is the probability of correctly rejecting the null hypothesis. Put differently, a test with high power (where more power is good) is one with a low probability of Type II error (where high Type II error is bad). As discussed below, a cost of decreasing Type I error is that Type II error will increase. Therefore, balance must be struck between these two types of error.

D. The balance of Type I and Type II error informs the test one performs and the significance level one chooses.

The issue of Type II error brings into focus two important considerations when performing a hypothesis test. *First*, when one has the ability to select from among multiple statistical tests that could all be used to verify a given null hypothesis, it is best to choose the test with greatest power.[4] *Second*, the natural sacrifice of a reduction in Type I error is an increase in Type II error.[5] Because hypothesis testing involves a balance of two different types of error, the actual application of a test of significance is an important aspect of the test itself.

To see this more clearly, consider the example of the medication above. Suppose that in conducting the hypothesis test, the researcher is overly focused on minimizing Type I error. The researcher is overly concerned about finding a statistical effect when one, in truth, does not exist. Exercising great caution when rejecting the hypothesis that the medication has no effect might seem like a good thing. The cost, however, is that it increases the chance of concluding the

[4] *See*, *e.g.*, Richard J. Larsen & Morris L. Marx, *An Introduction to Mathematical Statistics and its Applications*, at 370 (5th ed. 2012).

[5] *Id*. at 377-80.

medication *does not work* when it is indeed effective. This is also a bad outcome and could result in scrapping perfectly effective medicine.

Medical data, in particular, tend to be characterized by small sample sizes.[6] But even analyses with relatively large sample sizes have been known to fail to uncover true adverse effects in experimental drugs at a 5% level of significance.[7] When a study has smaller samples than would be ideal, this forces tradeoffs in the statistical mistakes that could occur. If the study uses 0.05 as the level of statistical significance, and it shows there is **no effect** between exposure and disease, this might be a true or false negative.

Consider the benzene and leukemia example. If the exposed and unexposed groups each contained roughly 77,000 people (impractical and prohibitively expensive) and researchers set the α level at 0.05 and β at 0.20, there would be a 20% chance (1 chance out of 5) of failing to detect leukemia even if benzene

[6] Douglas G. Altman, *Statistics and Ethics in Medical Research III: How Large a Sample?*, 281 Brit. Med. J. 1336, 1336-37 (1980), -

[7] *See*, *e.g.*, Stephen T. Ziliak, *The Art of Medicine: The Validus Medicus and a New Gold Standard*, 376 The Lancet 324, 325 (2010) (discussing the study of Vioxx); Jeffrey R. Lisse *et al.*, *Gastrointestinal Tolerability and Effectiveness of Rofecoxib versus Naproxen in the Treatment of Osteoarthritis*, 139 Annals Internal Med. 539, 543-45 (2003) (discussing that the 5 heart attacks of the Rofecoxib group was statistically different from the one heart attack in the control group at only a 20% level of significance). There were more than 5,000 individuals in this particular Vioxx study. *Id*.

caused it at a relative risk of 3.[8] In other words, they would have no better than an 80% chance of detecting that benzene caused leukemia at a relative risk of 3. Smaller sample sizes with a fixed 0.05 relative risk value would reduce further the chances of detecting leukemia at a relative risk of 3.[9]

Consequently a balance must be struck between these errors, and a singular focus on statistical significance can indeed be inappropriate.

II. STATISTICAL SIGNIFICANCE SHOULD BE WEIGHED AGAINST PRACTICAL IMPORTANCE

The concept of practical importance relates to either the magnitude of the effect being studied or the social significance of the effect itself. When a particular result or effect has a high level of practical importance, the cost of Type II error is magnified. As a result, significance testing must be conducted with particular care to avoid eschewing important results simply because they do not meet a particular level of statistical significance. [10]

[8] *Id.* The *power* of a study is 1- β, which from the numbers above would be 1-.20 = 0.80.

[9] *Id.* at 38.

[10] Practical importance can trump statistical significance outright when it comes to public safety. Recommended industry standards are that adverse events should be pursued diligently whether they are significant or insignificant in a statistical sense. *See* FDA, Center for Drug Evaluation and Research, Guidance for Industry: Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessments (Mar. 2005), at 4 (stating that "[i]t is possible that even a single well-documented case report can be viewed as a signal, particularly if the report describes a positive re challenge or if the event is extremely rare in the absence of drug use."). In addition, the FDA does not require a statistically significant association between a

A. Practical importance exists when the magnitude or consequence of the effect being studied is meaningfully large.

A variable or an effect has practical importance when the size of that effect is meaningfully large.[11] How the balance is struck is a matter of scientific, ethical, and political deliberation which needs to happen before the experiment and throughout the research and evaluation process, finally becoming the main determinant of judgments of the goodness or badness of the resultant product—as in the benzene and leukemia example. Researchers have noted, however, that statistical significance has been overemphasized at the expense of practical importance.[12] Researchers can become too focused on rejecting or failing to reject (in a statistical sense) a particular hypothesis rather than also analyzing whether the importance of the effect in question is itself large.

drug and a given effect to warrant a label change such as a precaution or warning. *See* 21 C.F.R. § 201.57(e) ("The labeling shall be revised to include a warning as soon as there is reasonable evidence of an association of a serious hazard with a drug; a causal relationship need not have been proved.").

[11] Donald N. McCloskey, *The Insignificance of Statistical Significance*, Sci. Am., Apr. 1995, at 32-33.

[12] *Id*; Deirdre N. McCloskey & Stephen T. Ziliak, *The Standard Error of Regressions*, 34 J. Econ. Lit. 97, 109-11 (1996); Deirdre N. McCloskey & Stephen T. Ziliak, *The Unreasonable Ineffectiveness of Fisherian "Tests" in Biology, and Especially in Medicine*, 4 Biological Theory 44 (2009) (summarizing instances in applied statistical analysis of medicine in which practical or clinical importance inappropriately lost to statistical significance); Roger E. Kirk, *Practical Significance: A Concept Whose Time Has Come*, 56 Educ. & Psychol. Measurement 746, 746-59 (1996); Steven Goodman, *A Dirty Dozen: Twelve P-Value Misconceptions*, 45 Seminars in Hematology 135, 136-37 (2008) (stating that statistical significance and clinical importance are not synonymous).

Because of the trade-off that may need to occur between statistical significance and practical importance, a bright-line rule of statistical significance would be poor practice. It would be impossible to construct an across-the-board rule that could take into account the necessary case-by-case balancing between practical importance and statistical significance.

B. The potential harm from Type II error is large when practical importance is high.

As stated above, using a too-low significance level increases the likelihood of mistakenly failing to reject the null hypothesis. This problem is magnified when the effect that is being studied is of great practical importance. If a phenomenon in question would, if true, result in great social or economic impact, then the potential harm from disregarding it on the basis of a hypothesis test is heightened.

This particular issue has been identified with regard to statistical analysis of medical data. For example, because clinical trials are often performed with small samples, it may be particularly difficult to garner statistical significance at the 5% level.[13] Moreover, an ethical problem exists in selecting a significance level that is too low when studying a clinical trial that involves elements of great practical

[13] *See*, *e.g.*, Douglas G. Altman, *Statistics and Ethics in Medical Research III: How Large a Sample?*, 281 Brit. Med. J. 1336, 1336-37 (1980); G.T. Lewith & D. Machin, *Change the Rules for Clinical Trials in General Practice*, J. Royal C. Gen. Prac. 239 (Apr. 1984).

importance.[14] Therefore, applying a significance test to medical data requires caution, as an appropriate balance must be struck between Type I and Type II error.

A particularly relevant example of this is the analysis of adverse events relating to Rofecoxib (brand name Vioxx). An initial analysis of the effectiveness of Vioxx, an anti-inflammatory medication, found that individuals in the Vioxx group suffered an increased rate of adverse heart-related events (such as infarction and stroke) relative to the control.[15] This increased rate of adverse events, however, was not deemed statistically significant.[16] Therefore, the risk of Vioxx toward heart-related adverse events was downplayed when the drug first entered the market. Once it became evident that heart-related risks did indeed exist, Vioxx was withdrawn from the market.[17] The two-pronged effect of this was that (1) individuals took the drug without fully understanding the potential heart-related risks, and (2) the drug, which could be used by some at acceptable risk[18] was no

[14] *See generally*, Lewith *et al.*, *supra* note 18.

[15] Jeffrey R. Lisse *et al.*, *Gastrointestinal Tolerability and Effectiveness of Rofecoxib versus Naproxen in the Treatment of Osteoarthritis*, 139 Annals Internal Med. 539, 543-44 (2003).

[16] *Id.* (discussing that the incidence of heart attacks in the Rofecoxib group was significant at 20%).

[17] *Merck Withdraws Vioxx; FDA Issues Public Health Advisory*, FDA Consumer, Nov.-Dec. 2004, at 38.

[18] Ricardo Alonso-Zaldivar, FDA *Gives Painkillers a Pass*, Newsday, Feb. 19, 2005, at A03.

longer available. Both of these problems could have been avoided had less attention been paid to p-values, which address the degree of consistency between data and the null hypothesis but reveal nothing about either the magnitude or variability of the effect.

C. The origins of significance testing further reveal the harm of disregarding practical importance.

As a corollary to the research on statistical significance versus practical importance, applied statisticians reexamined the origins of significance testing. The findings of this analysis have revealed that the common 5% statistical significance bench-mark came into prominence because it was an initial suggestion of preference of Ronald Fisher, one of the fathers of the hypothesis test.[19] In particular, Fisher preferred a 5% rule because at a critical value of 1.96—the critical value for a normally distributed test statistic using the 5% rule—one would reject the null hypothesis were one's result more than 1.96 standard deviations from the mean. Put differently, Fisher's preference was to categorize as significant results that were more than two standard deviations from expectation under the null hypothesis.[20]

The fact that this standard exists because an early developer of the test deemed it appropriate highlights an underlying problem of blindly applying this

[19] Lisse *et al.*, *supra* note 20, at 543-44.

[20] *Id.*

standard in all contexts. Indeed, a group of epidemiologists expressed a similar opinion in a brief submitted to the Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals*.[21] There, several professors in support of petitioners expressed their displeasure with the use of statistical significance testing as the only acceptable method of showing scientific validity in the field of epidemiology.[22] Moreover, these professors noted that significance testing is often mistaken as a fundamental input of scientific analysis.[23]

III. THERE IS NOTHING SACROSANCT ABOUT THE 0.05 STATISTICAL SIGNIFICANCE LEVEL

If courts consider only whether or not an epidemiological study is statistically significant at the 0.05 level, they invite numerous mistakes and their rulings may fail to consider important information.

A. Fixating only on low and fixed statistically significant rates precludes and obscures important study information.

Rothman and Greenland, using Freiman *et al.*'s examples,[24] reinforce how a demand for statistical significance at **α**= 0.05 for a study can frustrate the revelation of important information about relations in the world.

[21] *Daubert v. Merrell Dow Pharms*, *Inc*., 509 U.S. 579 (1993).

[22] Br. Amici Curiae of Professors Kenneth Rothman, Noel Weiss, James Robins, Raymond Neutra and Steven Stellman, in Supp. of Pet'rs, *Daubert v. Merrell Dow Pharms.*, *Inc*., 509 U.S. 579 (1993) (No. 92-102), 1992 WL 12006438, at *5.

[23] *Id*. at 3.

[24] Jennie A. Freiman, Thomas C. Chalmers, Harry Smith, Jr., and Roy R. Kuebler, *The Importance of Beta, the Type II Error and Sample Size in the Design and*

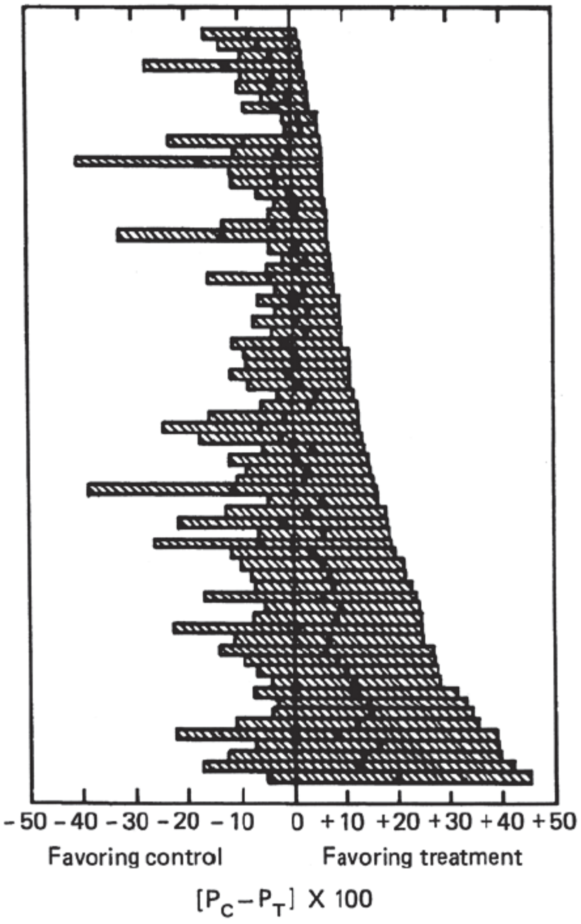
> In a classic review of 71 clinical trials that reported no "significant" difference between the compared treatments, Freiman *et al*. (1978) found that "in the great majority of such trials the data either indicated or at least were consistent with a moderate or even reasonably strong effect of the new treatment". [See Figure 1 below]. [However], the original investigators interpreted their data as indicative of no effect because the P-value for the null hypothesis was not "statistically significant." The misinterpretations arose because the investigators relied solely on hypothesis testing for their statistical analysis rather than on estimation. On failing to reject the null hypothesis, the investigators in these 71 trials inappropriately accepted the null hypothesis as correct, which probably resulted in type II error for many of these so-called negative studies.[25]

Figure 1: **Ninety Percent Confidence Limits for the True Percentage Difference for the 71 Trials**.[26]

Interpretation of the Randomized Control Trial—Survey of 71 Negative Trials, 299 New England J. of Med., 690, 694 (1978).

[25] Rothman and Greenland, *supra* note 2, at 154.

[26] Reproducing Figure 10-1 in Rothman, *Modern Epidemiology*, at 155, *supra* note 2.



Freiman, *et al.*, point out, "[F]or a reduction of 25 per cent in mortality from the control mortality rate [a beneficial effect], 50 per cent of the trials had betas in excess of 74%."[27] For those studies there was about a 75% chance of missing a beneficial outcome. "Only four of the trials (5.63 per cent) were large enough to ensure a beta ≤0.10, the usually accepted standard for clinical trials."[28] Rothman and Greenland offer explanations for why errors can occur:

> This failure to reject the null hypothesis can occur either because the effect is small, the observations are too few, or both, as well as from biases. More to the point,

[27] Freiman *et al.*, *supra* note 29, at 692.
[28] *Id.*

> however, is that type I and type II errors arise because the investigator has tended to dichotomize the results of a study in the categories "significant" or "not significant." Since this degradation of the study is unnecessary, an "error" that results from an incorrect classification of the study result is also unnecessary. …Declarations of significance or it absence can supplant the need for more refined interpretations of data; the declarations can serve as a *mechanical substitute for thought*, promulgated by the inertia of training and common practice. The neatness of an apparent clear-cut result may appear more gratifying to investigators, editors, and readers than a finding that cannot be immediately pigeonholed.[29]

Moreover, Freiman, *et al.*, argue that the reason the 71 clinical trials were rejected as not "significant" was "that most studies included too few patients to provide reasonable assurance that a clinically meaningful 'difference' (*i.e.,* therapeutic effect) would not be missed. Published reports gave few details of the prior planning, so that it is unclear to what extent the hazards of insufficient trial size were taken into account."[30]

For purposes of establishing causation in litigation, balancing between statistical significance and practical importance may require a case-by-case reassessment of the status quo.

[29] Rothman & Greenland, *supra* note 2, at 154 (emphasis added).
[30] Freiman, *et al.*, *supra* note 29, at 690.

B. Other levels of statistical significance may be appropriate in litigation.

Statistical significance need not be set at 0.05. It can be higher: Freiman, *et al.*, and Rothman and Greenland among others suggest that 0.10 could be appropriate.[31] For example, "forms of ionizing radiation have long been known as carcinogenic in many human organ systems," yet epidemiological studies "[b]ased on atomic bomb survivors … [are established] with 90 percent confidence [or 0.10 statistical significance]…."[32] The Federal Judicial Center's Reference Guide on Epidemiology in the *Reference Manual on Scientific Evidence* points out that statistical significance at values other than 0.05 can be appropriate,[33] noting "in its study of the effects of second-hand smoke, the U.S. Environmental Protection Agency (EPA) used a 0.10 standard for significance testing."[34]

[31] Freiman, *et al.*, *supra* n. 29, at 692; Rothman and Greenland, *supra* note 2, at 189.

[32] Carl F. Cranor, *Toxic Torts: Science, Law, and the Possibility of Justice* 229-230 (2d. ed. 2016), citing H. Kato, "Cancer Mortality," in Cancer in Atomic Bomb Survivors, ed. I. Shi-gematsu and A. Kagan (1986), quoted in Arthur K. Sullivan, "Classification, Pathogenesis, and Etiology of Neoplastic Diseases of the Hematopoietic System," in Wintrobe's Clinical Hematology, ed. G. R. Lee *et al.*, 1725, 1750 (9th ed. 1993); *see* Julius C. McElveen and Chris Amantea, "*Risk Symposium: Legislating Risk Assessment*," 63 U. Cin. L. Rev. 1553, 1556 (1995).

[33] Green, *et al*. "Reference Guide on Epidemiology," at 577-578 (Noting that although .05 is often the significance level selected, other levels can and have been used.)

[34] *Id.* (*citing*, U.S. Environmental Protection Agency, Respiratory Health Effects of Passive Smoking: Lung Cancer and Other Disorders (1992).)

Rothman and Greenland recognize that chances of false positives can exceed 0.05.[35] Some courts concur.[36]

C. Courts risk scientific and legal error by requiring statistical significance at 0.05.

The above discussion highlights a potential legal problem. Demands for low (0.05) statistical significance can hide causal information. If important causal information that would assist the fact finder in identifying a causal relationship is excluded, issues of causation cannot be fairly adjudicated. Unfortunately, courts often do not seem sensitive to the possibility of false negatives, *i.e.*, whether or not a study might yield a mistaken "no effect" or negative outcome.[37]

Given the critical role that causation plays, when reviewing a causation expert's proffered opinions, courts must root their analysis in an accurate understanding of statistical significance. Austin Bradford Hill, whose methodology

[35] Rothman & Greenland, *supra* note 2, at 187.

[36] *Turpin v. Merrell Dow Pharms., Inc*., 959 F.2d 1349, 1353-54, n.1 (6th Cir. 1992), *cert. denied*, 506 U.S. 826 (1992).

[37] There are, however, rare and welcome exceptions. In *Ambrosini v. Labarraque*, 101 F.3d 129, 136 (D.C. Cir. 1996), the Court recognized that negative studies can lack sufficient statistical power to be considered conclusive. *See id.* ("Statistical power . . . represents the ability of a study, based on its sample size, to detect a causal relationship. Conventionally, in order to be considered meaningful, negative studies, that is, those which allege the absence of a causal relationship, must have at least an 80 to 90 percent chance of detecting a causal link if such a link exists; otherwise, the studies cannot be considered conclusive."). A study such as the one in *Ambrosini* still allows for a 20% to 10% chance of a false negative, both asymmetrically higher than the typical odds of a false positive. *See also* Cranor, *Toxic Torts*, at 238.

for inferring causation from statistical studies is widely accepted by courts, argued that statistical significance was overrated in revealing causal relationships. Hill articulated "nine different viewpoints" that would assist researchers in inferring causation from associations that appeared in epidemiological studies. These "viewpoints" would help:

> [T]o make up our minds on the fundamental questions—is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect? No formal tests of significance can answer those [causal] questions. Such tests can, and should, remind us of the effects that the play of chance can create, and they will instruct us in the likely magnitude of those effects. Beyond that they contribute nothing to the "proof" of our [causal] hypothesis.

Courts, including the district court in this matter, have acted as if statistical significance were more important than Hill suggested.

IV. SCIENTISTS' USE OF 90% CONFIDENCE INTERVALS ALSO HIGHLIGHTS THAT 5% STATISTICAL SIGNIFICANCE IS NOT A BRIGHT LINE TEST

As discussed in Section I.B above, statistical significance was first used in commercial settings to make binary choices; a decision cutoff point—even an arbitrary one of 5% statistical significance—served a useful purpose. As Rothman and Greenland have observed, however, determining causation in epidemiology requires an inquiry more nuanced than applying a bright line 5% statistical significance threshold.[38] Put another way, if one seeks to understand the causal relationship between an exposure and disease (or a treatment and beneficial effects), one must find a more subtle—and more appropriate—way to *assess* the causal contribution; despite courts' familiarity with the 5% threshold in other contexts, treating a study as a causation on-off switch is not the answer. "Practices that reduce data analysis or scientific inference to mechanical 'bright-line' rules (such as '$p < 0.05$') for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making."[39]

To discern what epidemiological studies can reveal, Rothman has urged the use of confidence intervals (and other methods not addressed here). The confidence interval does more than assess the extent to which the null hypothesis is compatible with the data; it provides an idea of the magnitude of the effect and the inherent variability in the estimate. The p-value, on the other hand, indicates only the degree of consistency that exists between the data and the null hypothesis and reveals nothing about either the magnitude of the effect or its variability.

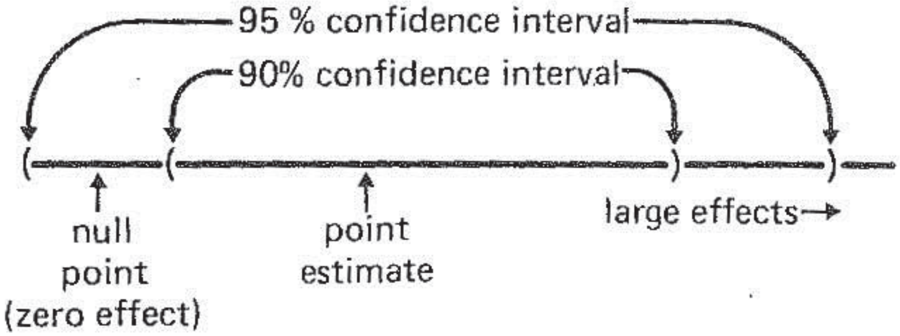
Moreover, even if one end of a confidence interval includes a "no effect" point (suggesting that the study shows "no effect"—a point judges may endorse or

[38] Rothman & Greenland, *supra* note 2, at 188, 189-192.

[39] ASA Statement on Risk-Limiting Post Election Audits, American Statistical Association (2010), http://www.amstat.org/policy/pdfs/Risk-Limiting_Endorsement

for which defendants might argue), if the other end of the confidence interval is asymmetrically beyond the point estimate of the study, it is compatible with the study showing an adverse effect. This is illustrated in Figure 2 below.

Figure 2: Two nested confidence intervals, with the wider one including the null hypothesis.[40]



In this figure, whereas a 95% confidence interval according to some would mean "there is 'no effect' from exposure," the asymmetry in the uncertainty distribution around the point estimate suggests an effect. Moreover, the smaller 90% confidence interval (analogous to a 0.10 statistical significance value) is consistent with an adverse effect; it excludes the "no effect" point. This represents a 10% chance that a positive study might be a statistical anomaly compared with a 5% chance when 0.05 is used for statistical significance. Rothman argues that the two confidence intervals in this case together show that "[b]ecause a 95% interval incudes the null point and a 90 percent interval does not, it can be inferred that the

[40] Reproducing Figure 9.1 in Rothman, *Modern Epidemiology*, at 120 (1986). The "null point (zero effect)" is typically represented as the value 1.0.

P-value is greater than 0.05 and less than 0.10."[41] In this example a researcher is not taking great chances with either odds of a false positive.

The 90% confidence interval in this example, analogous to statistical significance at 0.10, reinforces the point that courts should not routinely reject higher statistical significance values. Reviewing data with 90% confidence intervals can tell researchers more about what data show. Indeed, calculating confidence intervals "does much more than assess the extent to which a hypothesis is compatible with the data. It provides simultaneously an idea of the likely direction and magnitude of the underlying association and the random variability of the point estimate."[42] Random and non-random errors, and misunderstanding how to measure and interpret so called random error are of particular importance here.

As Rothman and Greenland point out, "Freiman *et al*. used confidence limits for the risk differences to reinterpret the findings from these studies. These confidence limits indicated that many of the treatments under study were in indeed beneficial, as seen in [Figure 1]."

[41] Rothman and Greenland, *supra* note 2, at 157.

[42] Rothman, *Modern Epidemiology*, at 158. *See also id.* at 157 ("The process of calculating the confidence interval is an example of the process of interval estimation.").

This discussion shows that the district court might not have understood how confidence intervals can reveal important information. For example, in CMO 68, the district court quoted the following with approval from *In re Bextra & Celebrex*, 524 F. Supp. 2d at 1174:

> Because the confidence interval includes results which do not show any increased risk, and indeed, show a decreased risk, that is, it includes values less than 1.0, we would say the study does not demonstrate a 'statistically significant' increased risk of an adverse outcome.

However, the above from Rothman along with Freiman *et al*. shows respectable scientists using confidence intervals, one end of which is less than 1.0 (somewhat beyond the null point), as a means of revealing information from statistical studies. Thus, as noted, a confidence interval that is asymmetric around a point estimate can show that data are consistent with an adverse or beneficial effect (whichever one is interested in). Moreover, 90% confidence intervals (based on α=0.10) are acceptable in the scientific community.

The relationship between statistical significance and confidence intervals is such that a study *with* a "statistically significant" p-value will also have a confidence interval that *does not* contain the null point (zero effect) of 1.0, and a study with a p-value that *is not* statistically significant *will* have a confidence interval that contains, or "crosses" 1.0 (*i.e.*, the lower bound of the confidence interval will be less than 1.0 and the upper bound will be greater than 1.0).

However, as the discussion above demonstrates, even non-significant confidence intervals can be better suited to assessing epidemiological associations. Courts should therefore embrace their utility in assessing causation.

CONCLUSION

The judgment of the district court should be reversed.

Respectfully submitted,

Christopher J. McDonald
Christopher J. McDonald
Christopher D. Barraza
LABATON SUCHAROW LLP
140 Broadway
New York, New York 10005
(212) 907-0700
cmcdonald@labaton.com

Counsel for Amici Curiae
Carl Cranor, Deirdre N. McCloskey,
and Stephen T. Ziliak

April 28, 2017

CERTIFICATE OF COMPLIANCE WITH TYPE-VOLUME LIMIT

Pursuant to Federal Rule of Appellate Procedure 32(g)(1), this brief complies with the Court's Type-Volume Limit for Briefs allocating 6,500 words. Excluding the parts exempted by Federal Rule of Appellate Procedure 32(f), this brief contains 6,474 words.

This brief complies with the typeface and type style requirements because this brief was prepared using a proportionally spaced typeface using Microsoft Word 2010 (Times New Roman, 14 point). This certificate was prepared in reliance on the word-count function of the word-processing system (Word 2010) used to prepare this brief.

/s/ Christopher J. McDonald
Christopher J. McDonald

CERTIFICATE OF SERVICE

I hereby certify that, on April 28, 2017, I electronically filed the foregoing Brief for Amicus Curiae Carl Cranor, Deirdre N. McCloskey, and Stephen T. Ziliak in Support of Plaintiffs-Appellants with the Clerk of the Court for the United States Court of Appeals for the Fourth Circuit by using the appellate CM/ECF system. All participants are registered CM/ECF users and will be served by the appellate CM/ECF system.

/s/ *Christopher J. McDonald*
Christopher J. McDonald