Washington Legal Foundation
Advocate for freedom and justice®
2009 Massachusetts Avenue, NW
Washington, DC 20036
202.588.0302

Critical Legal Issues: WORKING PAPER SERIES

THE ROLE OF STATISTICAL SIGNIFICANCE IN *DAUBERT*/RULE 702 HEARINGS

By

Kirby Griffis
Hollingsworth LLP

Washington Legal Foundation
Critical Legal Issues WORKING PAPER Series

Number 201
March 2017

TABLE OF CONTENTS

ABOUT WLF'S LEGAL STUDIES DIVISION

Washington Legal Foundation (WLF) established our Legal Studies division in 1986 to address cutting-edge legal issues through producing and distributing substantive, credible publications designed to educate and inform judges, policy makers, the media, and other key legal audiences.

Washington is full of policy centers of one stripe or another. From the outset, WLF's Legal Studies division adopted a unique approach to set itself apart from other organizations in several ways.

First, Legal Studies focuses on legal matters as they relate to sustaining and advancing economic liberty. The articles we solicit tackle legal policy questions related to principles of free enterprise, individual and business civil liberties, limited government, and the Rule of Law.

Second, WLF's publications target a highly select legal policy-making audience. We aggressively market our publications to federal and state judges and their clerks; Members of Congress and their legal staff; Executive Branch attorneys and regulators; business leaders and corporate general counsel; law professors; influential legal journalists, such as the Supreme Court press; and major media commentators.

Third, Legal Studies operates as a virtual legal think tank, allowing us to provide expert analysis of emerging issues. Whereas WLF's in-house appellate attorneys draft the overwhelming majority of our briefs, Legal Studies possesses the flexibility to enlist and the credibility to attract authors with the necessary background to bring expert perspective to the articles they write. Our authors include senior partners in major law firms, law professors, sitting federal judges, other federal appointees, and elected officials.

But perhaps the greatest key to success for WLF's Legal Studies project is the timely production of a wide variety of readily intelligible but penetrating commentaries with practical application and a distinctly commonsense viewpoint rarely found in academic law reviews or specialized legal trade journals. Our eight publication formats are the concise Counsel's Advisory, topical Legal Opinion Letter, provocative Legal Backgrounder, in-depth Working Paper, useful and practical Contemporary Legal Note, informal Conversations With, balanced On the Merits, and comprehensive Monograph.

WLF's Legal Opinion Letters and Legal Backgrounders appear on the Lexis/Nexis® online information service under the filename "WLF," and every WLF publication since 2002 appears on our website at www.wlf.org.

To receive information about previous WLF publications, or to obtain permission to republish this publication, please contact Glenn Lammi, Chief Counsel, Legal Studies, Washington Legal Foundation, 2009 Massachusetts Avenue, NW, Washington, DC 20036, (202) 588-0302, glammi@wlf.org.

ABOUT THE AUTHOR

Kirby Griffis is a Partner in the law firm Hollingsworth LLP in Washington, DC. He tries cases for corporate clients in high-stakes pharmaceutical and medical-device products-liability litigation, and in serial litigation cases of national importance. His most recent trial victory, with Hollingsworth LLP Partner Buffy Mims, was to force a favorable settlement on the last day before closing arguments in a 26-year-old personal-injury case in the Civil District Court for Orleans Parish, Louisiana, an American Tort Reform Foundation-certified hellhole jurisdiction. *Ezeb v. Sandoz Pharm. Corp.*, No. 1992-20622 (La. Civ. D. Ct. June 21, 2016).

Mr. Griffis has also helped to resolve many of the firm's complex matters by engineering *Daubert* victories and other summary dispositions. These include setting up and briefing the first *Daubert* summary judgment in the Parlodel litigation of 1998-2004 and extensive involvement with the "Parlodel trilogy" of appellate-level *Daubert* wins: *Glastetter v. Novartis Pharm. Corp.*, 252 F.3d 986 (8th Cir. 2001), *aff'g* 107 F. Supp. 2d 1015 (E.D. Mo. 2000); *Hollander v. Sandoz Pharm. Corp.*, 289 F.3d 1193 (10th Cir. 2002), *aff'g* 95 F. Supp. 2d 1230 (W.D. Okla. 2000); and *Siharath v. Sandoz Pharm. Corp./Rider v. Sandoz Pharm. Corp.*, 295 F.3d 1194 (11th Cir. 2002), *aff'g* 131 F. Supp. 2d 1347 (N.D. Ga. 2001).

THE ROLE OF STATISTICAL SIGNIFICANCE IN *DAUBERT*/RULE 702 HEARINGS

INTRODUCTION

Since 1975, Rule 702 of the Federal Rules of Civil Procedure has required that proposed expert testimony relate to "scientific … knowledge." For eighteen years, this had relatively little impact on whether expert witnesses could testify, but then in 1993, the US Supreme Court decided *Daubert v. Merrell-Dow Pharmaceuticals*, instructing courts that Rule 702 required them to act as gatekeepers of proposed expert testimony and directing them to ensure that such testimony rested on a "reliable foundation" and was "derived by the scientific method."[1] Rules 702 and 703 were partially rewritten to codify *Daubert*, and now they require that proffered testimony be "based on sufficient facts or data" and be "the product of reliable principles and methods" which the expert has "reliably applied … to the facts of the case."[2]

The Supreme Court fleshed out *Daubert* further with two more decisions. The three became known as the *Daubert* trilogy. In *General Electric v. Joiner*, the Court provided the appropriate standard of review on appeal of a district court's *Daubert* analysis (abuse of discretion) and rejected the US Court of Appeals for the Eleventh Circuit's view that a gatekeeper was to review only the expert's methodology.[3] The Court explained that "conclusions and methodology are not entirely distinct from one another," and that "nothing in either *Daubert* or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert."[4] The Court rejected the argument that the trial court should have confined its gatekeeping inquiry to the methodological question of "whether animal studies can ever be a proper foundation for an expert's opinion,"[5] and explained that a trial court must also consider *how this general methodology was applied* to reach the opinions at issue: "Of course, whether animal studies can ever be a proper foundation for an expert's opinion was not the issue. The issue is whether these experts' opinions were sufficiently supported by the animal studies on which

[1] *Daubert v. Merrell-Dow Pharms.,* 509 U.S. 579, 597, 590 (1993).

[2] FED. R. EVID. 702.

[3] *Gen. Elec. Co. v. Joiner*, 522 U.S. 136, 146 (1997).

[4] *Ibid.*

[5] *Id.* at 144.

they purport to rely."[6]

Kumho Tire v. Carmichael completed the *Daubert* trilogy, making clear that the *Daubert* standard applied to technical testimony as well as scientific.[7] The Court reiterated that the four factors it had set forth in *Daubert* itself were not mandatory factors,[8] but also made clear that it was not just up to a district court judge to pick criteria to apply to a *Daubert* hearing based on their fancy: they were to consider whether the expert "employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field."[9] In other words, a proffered expert witness needs to be applying the same standards in performing a causation (or other) analysis as people in the same field would apply to the same problem outside of the courtroom.

Skirmishing over the meaning of *Daubert* began immediately, and—once the legal community grasped the full potential practical impact of the decision—the skirmish lines developed into full-scale battles that have continued to this day. This WORKING PAPER addresses one facet of that battle, which is of considerable practical importance to those in the trenches—the role of statistical significance in the *Daubert*/Rule 702 test. This WORKING PAPER will chiefly be concerned with identifying how statistical significance comes up in fights over admissibility and what to look for in working up *Daubert* challenges involving statistical significance. It is not primarily concerned with describing the contours—sometimes uneven and inconsistent—of jurisprudence over statistical significance. In applying the lessons of this WORKING PAPER to a particular case in a given jurisdiction, allowances will of course have to be made for the applicable precedent, and a new approach urged where previous courts have gotten it wrong.

The examples provided here are from pharmaceutical and product-liability litigation, but the concepts have broader application.

Daubert is not just a subject of concern for people interested in the rules of evidence. In his concurring opinion in *Joiner*, Justice Stephen Breyer wrote that gatekeeping by district courts was needed to ensure that "the powerful engine of tort liability, which can generate strong financial incentives to reduce, or to eliminate, production, points toward the right substances and does not destroy the wrong

[6] *Ibid.*

[7] 526 U.S. 137, 145 (1999).

[8] *Id.* at 145-46.

[9] 526 U.S. at 152.

ones."[10] This observation points to a very important reality: the battle over junk science is about a lot more than whether a particular company must pay money to a particular plaintiff. It is about whether helpful or life-saving pharmaceuticals, chemicals, and other products will be available at all, and whether they will be used as widely as they should be. Active litigation currently targets dozens of products, including antidepressants taken by millions (Zoloft, Paxil, Lexapro, etc.); a leading antinausea medication prescribed to everyone from kids who need to keep down their first dose of antibiotics to cancer patients (Zofran); the herbicide generally credited as revolutionizing low-impact, no-till agriculture in the US and abroad (Roundup); and one of the classic chemotherapy drugs (Taxotere). And new products join the plaintiffs' bar's target list all the time.

The chilling impact of junk-science-fueled litigation cannot be denied. Multiple products have been removed from the market and then later shown *not* to be associated with the conditions that plaintiffs accused them of causing. Others are underused because public fear generated by the litigation—and plaintiffs'-bar advertising—deters use at an appropriate level. It has been well-documented, for example, that publicity over claims that antidepressants can increase suicide has led to a substantial drop in the use of such pharmaceuticals by people who need them. This drop, of course, is associated with a *spike* in suicides due to undertreatment of depression.[11] Outbreaks of measles and rubella in the children of parents avoiding vaccines due to well-publicized junk science are another example (albeit one mostly occurring outside the court system).

Daubert itself arose out of Bendectin litigation, a notorious legal debacle that provides yet another example of the harm junk science can impose. Bendectin was a morning sickness drug that was removed from the market by its manufacturer, Merrell Dow, due to the burgeoning cost of litigation. The plaintiffs' bar filed hundreds of lawsuits alleging that Bendectin had caused birth defects, despite the fact that multiple epidemiology studies—and multiple regulatory agencies—had found there to be no such association. Fear generated by the litigation led many women to have abortions, and hospital admissions for prepartum vomiting rose significantly once the drug was no longer available. It is now generally accepted that Bendectin does not cause birth defects, but millions of women with morning sickness were denied this safe treatment for three decades. In 2013, FDA approved Diclegis, a drug containing

[10] *Joiner*, 522 U.S. at 148-49 (Breyer, J., concurring).

[11] Robert D. Gibbons *et al.*, *Early Evidence on the Effects of Regulators' Suicidality Warnings on SSRI Prescriptions and Suicide in Children and Adolescents*, 164 AM. J. PSYCHIATRY 1356 (2007).

the exact some molecule as Bendectin.[12]

I. *DAUBERT* FORCES COURTS TO EVALUATE WHETHER SCIENTISTS ARE DOING THEIR JOBS

The reluctance of many judges to do what they consider to be sitting in judgment of science is, of course, the reason for the *Daubert* decision in the first place. That reluctance found immediate expression in the opinion of the Ninth Circuit on remand of *Daubert* itself:

> Our responsibility, then, unless we badly misread the Supreme Court's opinion, is to resolve disputes among respected, well-credentialed scientists about matters squarely within their expertise, in areas where there is no scientific consensus as to what is and what is not 'good science,' and occasionally to reject such expert testimony because it was not 'derived by the scientific method.' Mindful of our position in the hierarchy of the federal judiciary, we take a deep breath and proceed with this heady task.[13]

The Ninth Circuit—breath held—promptly did the right thing and excluded the proffered testimony.[14] This decision marked the death knell for Bendectin litigation. In fact, the Bendectin cases were *not* a particularly difficult challenge for a judicial gatekeeper court to manage: the science was strongly negative. Once the Supreme Court showed the way, the Ninth Circuit easily performed the analysis necessary to evaluate—and exclude—the proffered expert causation opinions. The Bendectin science consisted primarily of multiple negative studies and was anything but a dispute vigorously carried out in the pages of scientific journals between groups of respected, well-credentialed scientists on each side. If it had been, most judges would have found it very easy to conclude that the proposed testimony was admissible under *Daubert*.

In *Daubert,* the Supreme Court—informed by *amicus* briefs from various interested parties in the legal and scientific communities—set forth these non-mandatory factors district courts should consider when assessing the scientific reliability of proposed testimony:

[12] *See* Press Release, Food and Drug Administration, FDA Approves Diclegis for Pregnant Women Experiencing Nausea and Vomiting (Apr. 8, 2013).

[13] *Daubert v. Merrell-Dow Pharms.,* 43 F.3d 1311, 1316 (9th Cir. 1995).

[14] *Id.* at 1322.

1. Whether it involves "generating hypotheses and testing them to see if they can be falsified"—*i.e.*, whether the conclusion offered by the expert is one that is *testable* and *tested;*
2. Whether it has been subjected to peer review and publication;
3. The known or potential rate of error; and
4. How widely it is accepted in the relevant scientific community.[15]

As will be discussed below, these criteria come straight from the scientific method itself. Although the Court described the factors as non-mandatory, it made clear that district court judges are to assess proffered scientific testimony by the standards of the applicable scientific discipline. The court must ensure that the expert "employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field."[16] In practice—since the factors are part of the scientific method—that means the factors must be applied in virtually every case.

It is essential that judges conduct the screening required by *Daubert* and Rule 702. However unpleasant it may be to delve into and make a decision about science, a lay jury is certainly far less able to make such a determination at trial than a judge is at a *Daubert* hearing. As the Eleventh Circuit put it:

> The *Daubert* trilogy, in shifting the focus to the kind of empirically supported, rationally explained reasoning required in science, has greatly improved the quality of the evidence upon which juries base their verdicts. Although making determinations of reliability may present a court with a difficult task of ruling on matters that are outside of its field of expertise, this is less objectionable than dumping a barrage of scientific evidence on a jury, who would likely be less equipped than the judge to make reliability and relevancy determinations.[17]

II. THE SCIENTIFIC METHOD

Posing *Daubert* challenges requires a sound understanding of the scientific method and how statistical significance fits into it.

[15] *Daubert*, 509 U.S. at 493-94.

[16] *See, e.g., Kumho Tire,* 526 U.S. at 152.

[17] *Rider v. Sandoz Pharm. Corp.*, 295 F.3d 1194, 1197 (11th Cir. 2002) (citations & quotations omitted).

Science proceeds by intelligent observation of the world. The scientific method involves the formulation of a hypothesis, the comparison of that hypothesis to measurements of observed phenomena, and the modifying or discarding of the hypothesis if it does not accurately predict the measurements taken of the phenomena. Science frequently considers a hypothesis to be *disproved* because of its failure to comport with the observed data, and far more rarely—after a great many additional sets of observations—to be *proved*. This disparity is in part because one or a handful of possible associations may turn out to be mere statistical noise, as detailed at some length below.

The layperson's idea that studies *prove things* comes from a misunderstanding of the scientific method. A layperson may believe—and many jurors do—that scientists come up with a hypothesis, design a study to test it, and then if the study comes out a certain way deem the hypothesis to be *confirmed*. This makes it very easy for a plaintiffs' lawyer to tell a jury that a "positive" result in a study designed to assess whether X and Y are associated *proves* that X causes Y. Real scientists do not talk that way. Scientists come up with a hypothesis about the world and then design studies that will test whether their hypothesis is or is not consistent with observed data.

Even an association between two variables that is sufficiently persistent across multiple sets of data to be accepted as a valid association may not be accepted as a causal association. The finding in any particular study of an association between a substance and an injury is not equivalent to causation.[18] There are three reasons that a positive association may be observed in an epidemiological study: (1) chance, (2) bias, and (3) real effect.[19] Bias does not refer only to observational prejudices in those collecting the data (though it includes this), but also to a plethora of other simple and technical sources of error. Furthermore, even a "real effect" (*i.e.*, variable X and variable Y truly are associated with one another) could be real for a reason very different than that in the minds of the researchers.

For example, let us suppose that, in 1972, researchers at UCLA posit that marijuana causes hair growth, and they perform a study that purports to find an association between marijuana use and long hair in men. Their study could have a number of different meanings. **First**, the observation may be due to mere chance. A

[18] *See* Michael D. Green *et al.*, *Reference Guide on Epidemiology*, REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 552 (3d ed. 2011).

[19] *See Magistrini*, 180 F. Supp. 2d at 591; *Caraker v. Sandoz Pharms. Corp.*, 188 F. Supp. 2d 1026, 1032 (S.D. Ill 2001); *see also* Eddy A. Bresnitz, *Principles of Research Design*, in GOLDFRANK'S TOXICOLOGIC EMERGENCIES 1827-28 (Goldfrank *et al.* eds. 6th ed. 1998).

look at the strength of the association measured by statistical tools and repeated studies will help assess this possibility. **Second**, the observation could be due to one or more sources of bias, such as the graduate students who are conducting the study unconsciously skewing their selection of subjects and their subjective assessments of hair as "long" or "not long" in favor of their hypothesis. **Third**, the effect could be a real one (*i.e.*, men who smoke marijuana really are more likely to have long hair), but the hypothesis about marijuana causing the hair growth could be false. It may be that men with long hair are more likely to smoke marijuana, that the same propensities that lead men to choose to wear their hair long also lead many of them to consume marijuana, and so on. Finally, of course, it may be that marijuana really does cause hair growth—but no real scientist would so conclude without far more evidence.

If scientists suspect Bendectin to be teratogenic (*i.e.* causative of embryonic and fetal malformations), they would test that theory with an experiment that examines thousands of births in women who were exposed to Bendectin and compares them to a set of similar births in women who were not, comparing the incidence of birth defects in both groups. If the results show no meaningful differences between the two groups (as was the case with the Bendectin studies)—and particularly when that result is repeated—scientists would have to conclude that the experiment failed to demonstrate a causal connection between Bendectin and birth defects. The existence of a causal connection, put more bluntly, has been refuted. Science classically proceeds not by confirmations but by refutations, which is why one of the seminal works in the history and philosophy of science is entitled *Conjectures and Refutations*.[20] This principle is enshrined in the first *Daubert* factor, which asks whether a proposed causal relationship is falsifiable, and, if so, whether it has been falsified. *Daubert*, 509 U.S. at 493-94.

The most important type of scientific evidence regarding human causation is epidemiology. There are two categories of epidemiological studies: experimental studies and observational studies. The "gold standard" in experimental epidemiology is the double-blind, randomized controlled clinical trial (RCT), the type of experimental study that FDA requires before approving a drug as safe and effective.[21] In an RCT, scientists test a predetermined hypothesized association by exposing a group of

[20] Karl Popper, CONJECTURES AND REFUTATIONS: THE GROWTH OF SCIENTIFIC KNOWLEDGE (5th ed. 1989). The Supreme Court cited *Conjectures and Refutations* in *Daubert*, making it one of a very short list of philosophy books ever to achieve that distinction.

[21] *See Reference Guide on Epidemiology*, *supra* note 18, at 555 ("Such a study design is often used to evaluate new drugs or medical treatments and is the best way to ensure that any observed difference in outcome between the two groups is likely to be the result of exposure to the drug or medical treatment.").

randomly-assigned individuals in a clinical setting either to the studied treatment or a placebo and then following them, measuring any differences in the outcome at interest.

In the absence of RCTs, the most scientifically reliable evidence of causation in humans comes from observational epidemiology. In observational studies, scientists seek to infer associations from exposures that occur in non-controlled settings, either by comparing the incidence of disease among individuals exposed to an agent with an unexposed group ("cohort studies") or by comparing the frequency of prior exposures in individuals who have a disease as compared to a group of individuals who do not have the disease ("case control studies").[22]

In both cohort and case-control studies, scientists compare two populations to determine if an association exists between an exposure and a disease. In a cohort study, scientists compare individuals with an exposure to individuals without an exposure. If a greater percentage of individuals with an exposure subsequently develop a disease than do those without the exposure, the study will report a positive association. Likewise, a case-control study will report a positive association if a greater percentage of individuals with a disease (cases) report a given exposure in their past than do healthy individuals (controls).

III. STATISTICAL SIGNIFICANCE AND HOW IT WORKS

Statistical significance is a mathematical tool that was developed in the 1920s as a way of assessing the meaning of a particular set of observed results. Although the mathematics is somewhat daunting (even to many scientists), the concept is not: statistical significance measures how likely it is that repeated data sets of similar size would yield a similar outcome. This is widely viewed—incorrectly—as being the same thing as how likely it is that a particular set of data could be obtained merely by chance. If a particular set of data is fairly likely to occur merely by the operation of chance, then the outcome is normally considered to be negative—*i.e.*, not indicative of an association.

Consider perhaps the simplest possible scientific experiment: flipping a coin to determine if it is biased. If a person flips a coin two times and obtains heads both times, he may begin to entertain the hypothesis that the coin is biased in favor of heads. However, the data does not yet support any such hypothesis: the chance that an *unbiased* coin flip will yield heads twice in a row is ½ x ½, or 25%. Three heads in a

[22] *See Magistrini v. One Hour Martinizing Dry Cleaning*, 180 F. Supp. 2d 584, 590-91 (D.N.J. 2002)*, aff'd*, 68 Fed. Appx. 356 (3d Cir. 2003).

row will occur 12.5% of the time (½ x ½ x ½), and so on. The trend is not statistically significant until it crosses some threshold of likeliness—most commonly 5%, or 1/20 (the selection of an appropriate threshold will be discussed *infra*). For the simple experiment, the chance of a series of heads crosses the line at 5 flips:

# of heads in a row	Calculation	Likelihood of repetition of result
2	$(½)^2$	25%
3	$(½)^3$	12.5%
4	$(½)^4$	6.25%
5	$(½)^5$	3.125%

Coins are not the only things that can be biased: scientists can be, too, as can their experimental subjects, their hypotheses, and their manipulations of the data. Scientists are often wrong, and often wrong because of their own biases and prejudices. One notorious example is decades of refusal by surgeons to accept evidence that basic hygiene would reduce infections and mortality rates in their patients (gentlemen could not have unclean hands). Litigation (and being paid to testify in litigation) is another such source of bias. Science has long recognized that scientists can—even with pure hearts and the best of intentions—see in the data that which really is not there, due to their own wishes, preconceptions, or errors in thinking. Science has developed many tools to reduce the size of the field in which bias can operate. Double-blinding is one such tool, removing some of the opportunities that researchers and research subjects otherwise would have to influence outcomes, inadvertently or otherwise.

Statistical significance is another such tool. For example, a properly applied test of statistical significance can tell a scientist that the trend that he thought he was seeing in his experimental animals supporting his pet hypothesis is in fact perfectly consistent with mere chance. Statistical significance is ordinarily wielded against far more complex data sets than a few coin tosses, of course, such as the test scores of a hundred students who ate breakfast compared to those of a hundred who did not.

IV. A BRIEF STATISTICS AND STATISTICAL-SIGNIFICANCE GLOSSARY

Assessing statistical significance in a complicated data set requires advanced training and education in the selection of appropriate statistical tools and how to apply them, a subject far beyond the scope of this monograph and the training of many scientists. An understanding of the basic terms and concepts used in such assessments is well within the grasp of any educated person, though. Some of the most important concepts follow.

An **association** or a **correlation** is an observed relationship between two variables whereby, when one is present, the other is at least somewhat more likely to be present as well. Thus, for example, smoking is associated with lung cancer; it is also associated with higher alcohol intake. An association can be a **causal association** (one of the factors causes or contributes to the other) or not. For example, smoking does not cause higher alcohol intake or vice versa; instead, there is overlap in the socioeconomic and personality traits that lead a person to smoke and those that lead a person to consume alcohol. It is a scientific maxim that **correlation does not equal causation**; among the devices used to determine whether an association is causal are those of statistical significance.

Statistical significance is closely related to several other statistical measures used to assess the reliability of associations found in the data.

P-value is expressed as a percentage (usually 95% or 99%, corresponding to 0.05 or 0.01). The p-value test is used to determine if a result is statistically significant. Formally speaking, a p-value measures the chance that a repeated experiment of the same size would generate data as strong or stronger than the current data against the null hypothesis (that there is no association), and *not* the chance that the association is true. The distinction is one that is rather likely to be lost on anyone who is not a statistician. P-values are almost universally translated, in the press and in the courtroom, into direct measurements of the likelihood that a claimed association is true.

It is common to see in scientific studies statements that a particular association was statistically significant, or just "significant" (the latter is just a short-form version of the former), and in the data charts an indication of measured significance such as "P<0.01."

Relative risk is a way of expressing the strength of a particular association (*i.e.*, the likelihood that it will manifest). For example, saying that the relative risk of lung cancer associated with smoking two packs a day is 7.5 means that, all other things being equal, a two-pack-a-day smoker is 7.5 times more likely than a nonsmoker to develop lung cancer.

No good scientist believes that their relative-risk calculation is the precise real-world value; they will express their results along with a **confidence interval**, which is the range of possible relative-risk values consistent with the data to a particular degree of statistical significance, usually 95% (often expressed as 0.05 in charts) or 99% (0.01). A confidence interval is critical for putting a relative risk into perspective: if the range of a confidence interval includes 1.0 (*i.e.,* the value consistent with no association whatsoever between the two variables being studied), the result is often considered to be a negative one. 1.0 is the **null value**—the value consistent with the "null hypothesis" that X and Y are not associated with one another at all. Scientists often say that such a result is "not significant."

The chart below shows confidence intervals for three hypothetical studies, each with a relative risk of 7.5.

Study 1	RR 7.5 (99% CI 6.3-8.9)
Study 2	RR 7.5 (95% CI 1.2-13.7)
Study 3	RR 7.5 (95% CI 0.4-35.6)

Based on these results alone (and ignoring any considerations of the methodology, biases, and so on of the study), the reported result from Study 1 would be generally considered a very strong one. Translated into English, it reports that there is a 99% chance that a repeated experiment would find a relative risk value between 6.3 and 8.9. It is a strong result because it uses a robust measure of confidence (99%) and the interval is narrow. Such results are suggestive of a large study and a strong association.

The Study 2 result is of a sort more common in the published literature: a less robust measure of confidence (95%), a wider confidence interval, and a lower bound, near the null-value result of 1.0. A savvy reader will immediately wonder if the 95% value was chosen because it—unlike the 99% one—yielded a range that excluded 1.0 and therefore could be labeled as "statistically significant."

The Study 3 result probably would not be published on its own, but may well be reported along with more robust data yielded from a study. The relative risk of 7.5 is the same as the other two, but the confidence interval is very wide and is consistent both with a null hypothesis (1.0) and with the two variables under investigation being *negatively* associated with one another (values <1.0). Nevertheless, plaintiffs' lawyers all too often tout such a result as showing that X is 7.5 times more likely to cause Y, leaving it to the defense to try to educate a judge—or a jury (if the *Daubert* gatekeeper has already failed)—as to why that risk value is essentially meaningless.

Odds ratio is an alternative way of expressing the strength of a particular association; it directly compares the occurrence of the variable under study in the experimental group to that in the control group. An odds ratio of 1.3, for example, means that the variable in question occurred 1.3 times as often in the experimental group as in the controls. Odds ratios also should be expressed with confidence intervals.

V. FINDING THE FLAWS FOR A *DAUBERT* CHALLENGE

Statistical significance has been an important *Daubert* issue from the beginning, in part because it is inherent in the "known or potential rate of error" factor. The Bendectin litigation that gave rise to *Daubert* soon collapsed in large part because of the absence of statistically-significant epidemiology supporting causation.

As the US Supreme Court has explained, epidemiological research cannot provide a scientifically reliable basis for an affirmative causation opinion if it is statistically insignificant or inadequately controlled for bias.[23] In *Joiner*, the Court took on statistical significance directly. It held that "[a] court may conclude that there is simply too great an analytical gap between the data and opinion proffered,"[24] and concluded that the research cited by the plaintiff's experts did not validate their conclusions in part because the epidemiological studies did not report a statistically significant causal link between PCBs and lung cancer, lacked proper controls, and examined substances other than PCBs.[25]

An exhaustive list of possible flaws in a body of scientific evidence with which an advocate or a judge may be asked to contend—and corresponding lines of attack that may be developed—are too ambitious for this WORKING PAPER, but some major themes that are directly relevant to statistical significance are set out below.

A. The 1-in-20 Rule and Cherry Picking

A P-value of 0.05 (corresponding to a 95% confidence level) sounds good, particularly when it is incorrectly expressed—as it likely will be to judge and jury—as reflecting a 95% chance that the claimed association is a true one. But 95% reflects a 1/20 chance that the association is spurious. *Random* data will generate a spurious association between two variables to a 95% confidence interval one time out of 20.

[23] *See General Electric Co. v. Joiner*, 522 U.S. 136, 145-46 (1997).

[24] *Id.* at 144.

[25] *Id.* at 145-46.

This is of high importance when, as is often the case, there are many negative studies and few positive studies (or just one) purporting to find an association. All things being equal, even if there is *no* association between two variables, one out of every 20 studies will nevertheless find an association to a 95% confidence interval. When the bias against publication of negative results is considered, the likely ratio of negative to positive studies in the literature becomes even smaller, even when there is no association at all.

Plaintiffs' experts frequently pull selected findings from the literature and leave the rest, without scientific justification for their selections. In the Prempro litigation, for example, the plaintiffs' epidemiologist and cell biologist testified that Premarin, an estrogen-only hormone replacement therapy, caused breast cancer. To support this conclusion, she relied on subgroups in various studies that had shown a positive association with Premarin, "while discounting subgroups where EHRT had no statistically significant effect."[26] She downplayed negative results from larger studies and relied on multiple studies that were not statistically significant.

Another example comes from Zoloft litigation, in which the judge assessed proffered testimony of Dr. Anick Bérard, a pharmacoepidemiologist who claimed to find that Zoloft caused more than a dozen different kinds of birth defects and other problems. Her methodology included making a "forest plot" of odds ratios from many different studies of different antidepressant drugs and different conditions and purporting to discern trends in odds ratios among those studies.[27] The court noted that if there were a class effect of antidepressants causing teratogenic effects, one would expect *consistent* associations—similar drugs causing similar effects in similar populations.[28] Unsurprisingly, a body of literature involving multiple drugs and many conditions will—by the operation of mere chance—include a number of statistically-significant results that are not true causal associations. The court correctly understood this in excluding Dr. Bérard.

The 1-in-20 rule does not just apply to whole studies, but also to particular data analyses. For example, drugs and chemicals are often tested with large screening tests, in which they are evaluated against dozens of different outcomes at the same time. For example, a study may assess a large body of health data in a large group of patients to see if patients who were exposed to possible toxin "T" were more likely to develop any of 40 different cancers. In such a study, one should expect a *false-positive*

[26] *In re Prempro Products Liability Litigation,* 738 F. Supp. 2d 887, 892 (E.D. Ark. 2010).

[27] *In re Zoloft*, 26. F. Supp. 3d 449, 455 (E.D. Pa. 2014).

[28] *Id.* at 458.

association to be found with two of those 40 cancers to a 95% confidence interval (at a 95% confidence interval, there will be a false positive one out of every 20 times, or twice with 40 different outcomes examined). Studies like this are properly used to generate hypotheses for further research. If T shows a statistically-significant positive association with non-Hodgkin's lymphoma and testicular cancer, then larger studies investigating those specific outcomes may be warranted. It would *not* be appropriate to conclude that T in fact causes either form of cancer outcome based on this study.

B. All the Data

Courts should certainly look askance at proposed expert testimony that does not take into account studies that fail to confirm an expert's hypothesis. Scientists draw causation conclusions—once enough evidence is available—by taking into account all of the evidence available. They look at all of the evidence, assess what weight it deserves given objective scientific criteria (such as that epidemiology studies are of more value than animal studies in assessing causation in humans, that a study with a small sample size may be unable to provide a reliable measure of a supposed association, and so on). The evidence points to a conclusion. With regard to causation, several outcomes are possible, such as: no causal association; some evidence of causation but not enough to draw any conclusion; or strong and consistent evidence of causation. Negative findings must be explained and accounted for.

It is quite common for expert witnesses to take the reverse approach, starting with the conclusion that they have been asked to defend and then reasoning backwards. Plaintiffs' experts often try to rely on just the studies that support their position (statistically-significant or otherwise), and find reasons to ignore those that are unsupportive. The lawyer's job is to expose the unscientific nature of this selectiveness.

Courts should reject experts who cherry-pick isolated epidemiologic findings and fail to explain why they ignored contrary epidemiologic findings.[29] The federal district court's meticulous exclusion of proffered testimony from plaintiffs' expert witness Dr. Nicholas Jewell in the *In Re Zoloft* litigation is currently on appeal to the

[29] *See Arias v. DynCorp*, 928 F. Supp. 2d 10, 17 (D.D.C. 2013) (rejecting testimony that ignored contrary studies without adequate justification); *see also Cano v. Everest Minerals Corp.*, 362 F. Supp. 2d 814, 850 (W.D. Tex. 2005) (rejecting testimony of expert who "sifted through the literature to pick and choose positive relative risks between ionizing radiation (of any type, source, and dose) and a particular Plaintiff's cancer").

Third Circuit.[30] Dr. Jewell's mistreatment of the Zoloft science is of the sort that *Daubert* practitioners see all the time. Dr. Jewell—whose testimony the district court allowed after prior experts had been excluded—claimed to find a true causal association between Zoloft and cardiac birth defects based on several studies that found a statistically-significant association and that Dr. Jewell described as non-overlapping and consistent with one another. The court noted that Dr. Jewell ignored a later and more comprehensive study that largely subsumed the same data as the prior studies into a larger pool of data—and found no increased risk.[31] The larger study in question "includ[ed] virtually all the data from the earlier Danish studies" that Dr. Jewell had relied upon.[32] Yet while those earlier Danish studies had found a tripling of the relative risk of a cardiac birth defect, the later study found "no association between Zoloft use and cardiac birth defects."[33]

Outside of the courtroom, when a scientist is confronted with a new and larger study that raises questions about his conclusions, the scientist needs to explain *why* this new study does not undermine those earlier conclusions.[34] Yet Dr. Jewell was unable "to provide any methodological or statistical explanation for why this larger, later study failed to replicate the findings of the earlier study, or why the earlier studies should be considered more reliable."[35]

Of course, scientists can and do assign differing weight to different scientific studies, deeming some to be persuasive while discounting others for various reasons. But they must be consistent and scientific in the criteria that they use to do so. When scientists cannot articulate an objective, principled basis for their selection of evidence, the gatekeeper is justified in concluding that they are engaged in something other than science. Rule 702 and *Daubert* require gatekeepers to exclude such

[30] Ed. note: Washington Legal Foundation filed an *amicus* brief in support of the Respondent in *In re Zoloft*, *available at* http://www.wlf.org/upload/litigation/briefs/WLFAmicusBrief-InreZoloft-.pdf.

[31] *See, e.g.*, *In re: Zoloft*, 2015 WL 7776911 at *16.

[32] *See In re: Zoloft*, 2015 WL 7776911 at *7.

[33] *Ibid.*

[34] *See id.* ("Scientists are expected to address and reconcile data that does not support their opinions, and not simply rely upon data which does.").

[35] *Ibid; see also Joiner*, 522 U.S. at 146 ("[N]othing in either *Daubert* or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert.").

evidence as not scientifically reliable.[36]

Sometimes, the problem is not that the other side is ignoring studies that exist but do not support their conclusion; rather, there is not much evidence on the relevant subject in the first place. If the evidence available to experts is scant, they may try to argue that they have simply done the best that they can with the available scientific data and that this should pass *Daubert* muster. But reputable scientists do not conclude that a drug causes a malady or a chemical causes cancer based on a few scraps of evidence; they conclude that sufficient evidence is lacking. *Daubert* requires the exclusion of proffered testimony that is not "based on sufficient facts or data."[37] As Judge Richard Posner of the Seventh Circuit explained, "the courtroom is not the place for scientific guesswork, even of the inspired sort. Law lags science, it does not lead it."[38] The trilogy of cases involving the drug Parlodel are excellent illustrations of the exclusion of witnesses (and termination of a large litigation) because a theory of causation cobbled together from disparate and insufficient elements was, after much litigation, deemed insufficient to pass *Daubert* muster.[39]

C. P-Hacking

Scientists are under strong pressure to publish and produce, and negative results are not as likely to be published or to advance careers as positive ones. These pressures can lead to many scientific errors barely short of outright fraud. One is colloquially called "P-hacking," which involves various manipulations of data analysis to achieve statistically-significant results. The selection of the particular statistical methodology to calculate significance is one way this is done. Scientists achieve negative results by one method, and then they reanalyze the data using a different method until they achieve a positive result.

P-hacking may be hard to detect if done by the original scientists, but it is fairly common in "reanalyses" performed by hired experts who purport to discover a

[36] *See, e.g.*, *Arias*, 928 F. Supp. 2d at 25 (excluding expert testing where expert failed to "explain why he decided to credit [one study's] results and dismiss [another study's] results").

[37] FED. R. EVID. 702.

[38] *Rosen v. Ciba-Geigy Corp.*, 78 F.3d 316, 318-19 (7th Cir. 1996) (affirming exclusion of expert's causation testimony where expert failed to explain how a nicotine overdose could precipitate a heart attack and failed to cite scientific or medical literature in support of theory because such testimony "was not real science" and "lacks scientific rigor").

[39] *See Glastetter v. Novartis Pharm. Corp.*, 252 F.3d 986 (8th Cir. 2001); *Hollander v. Sandoz Pharms. Corp.*, 289 F.3d 1193 (10th Cir. 2002); *Rider v. Sandoz Pharm. Corp.*, 295 F.3d 1194 (11th Cir. 2002).

significant association where the original authors did not. One such case involved an expert pathologist who claimed that chlorpyrifos, an insecticide, caused non-Hodgkin's lymphoma, based on a study that found a relative risk of 1.6 in the highest quartile of exposure, with a 95% confidence interval of 0.74-3.53 (in other words, not statistically significant). The expert lowered the confidence interval to 80%, which (barely) qualified the study as statistically significant. The district court judge had no problem finding this to be unscientific.[40]

The danger of P-hacking comes when scientists (or experts hired to prove a questionable association) secretly analyze data repeatedly until they find the set of assumptions that yields the result they want, finally publicly presenting these results as scientific evidence. In litigation involving Viagra, a plaintiff's expert witness had authored a study linking Viagra to a vision problem. The defendants obtained the data and records involved in the study by subpoena and demonstrated to the district court that the study had multiple data and statistical analysis flaws.[41]

Another species of P-hacking occurs with **subgroup analysis**, whereby the researchers run statistical tests on various slices and subgroups of data in search of a statistically significant association. For example, a study may investigate whether exposure to a particular chemical is associated with breast cancer in women. This study may report that the association was negative in general, but it was statistically significant for the subgroup of premenopausal women. This sort of finding—particularly if the p-value is unimpressive and the confidence interval wide—is suspicious. Subgroup analysis can be perfectly appropriate—such as when a study proposes to perform a particular subgroup analysis that is a logical part of the hypothesis under investigation and *discloses that intention in advance* and when the study is designed to be large enough to yield meaningful results even with regard to a reduced subset of the data. When done after the fact, subgroup analysis is far more likely to lead to error.

Study authors who do not wish to be accused of P-hacking will set out their methodology—including the statistical tests that they will use and which, if any, subgroups will be analyzed—*before* collecting the data.

Another excellent example of P-hacking comes from litigation involving Lipitor. The plaintiffs provided their proffered expert statistician, Dr. Nicholas Jewell, with a large amount of data from Parke-Davis's New Drug Application for Lipitor. Dr. Jewell mined the data, extensively analyzing different variables in different ways until he

[40] *Pritchard v. Dow Agro Sciences,* 705 F. Supp. 2d 471, 488 (W.D. Pa. 2010).

[41] *In re Viagra Products Liability Litigation,* 658 F. Supp. 2d 936, 946 (D. Minn. 2009).

purported to find a statistically-significant association between Lipitor and elevated blood-glucose levels. Dr. Jewell did not, of course, publish in advance a plan of analysis, but he instead engaged in a "whole lot" of analyses, excluding from his expert report multiple analyses that he "didn't believe … supported … the kinds of opinions [he] wanted to put in [his] summary."[42] Dr. Jewell did not retain the analyses that had not yielded the results that he wanted.[43] The court identified multiple errors in Dr. Jewell's *selection of data* for consideration: for example, he relied on single elevated glucose readings as evidence that Lipitor could produce diabetes without a sound clinical basis to do so, and he applied different standards in different studies to select which groups of data to look at (sometimes using single elevated glucose measurements, sometimes ignoring glucose-measurement data entirely).[44]

Experts must be able to justify *why* they chose to rely on some data and exclude other data from consideration, and they must particularly explain any decisions made to filter data in one way in one study and in another way in another. Playing with what data will or will not go into a calculation is, however, only one way to alter the results of the calculation. A second way to change results is to alter the formula used in the calculation. The criteria employed in deciding which statistical tool to use to measure data are technical and complex. Factors such as the size of the data set, the number of possible confounding variables, and how common the outcome under investigation is in the background population affect which tool is best. In any event, the best practice is to select a formula appropriate for the kind of data expected *in advance*. Deciding this only after the data is collected is questionable, and it is just bad science to try out multiple tools to see which, if any, yield a significant result.

It can be hard to tell if a scientist has tried out multiple statistical tools in analyzing data from a published study. Sometimes—particularly with large epidemiological studies—prior to data collection, one or more publications will be done laying out the methodology to be employed, including the statistical analysis. Often no evidence will arise regarding how a particular tool was selected; the study will merely give a result, perhaps naming the tool in question (such as Fisher's exact test, Barnard's exact test, the Cochran-Mantel-Haenszel test, etc.), perhaps not.

Defendants must probe whether the expert witness being challenged has fished around in the data with different tools. Dr. Jewell—seemingly an inexhaustible

[42] *In re Lipitor*, MDL No. 2:14-mn-02502-RMG (D.S. Car. CMO 54 Nov. 20, 2015) at 6.

[43] *Ibid.*

[44] *Id.* at 7-8, 13.

source of *Daubert* anecdotes—once again serves as an example here. In assessing Lipitor data, even after all of the liberties that he took with *selecting* data, he still could not get a statistically-significant result employing a Fisher's exact test, so he switched to another test called a mid-p test, which generated a (barely) statistically-significant result. The MDL court rejected that—not because of the selection of the mid-p test, but because that selection was made in an unscientific way:

> It is important to note that using the mid-p approach, standing alone, does not render Dr. Jewell's analysis unreliable. The mid-p approach is used by some statisticians and can be a valid methodology. (See Dkt. No. 972-10 at 13 n.16). For instance, if Dr. Jewell thought the mid-p approach a better approach than the Fisher exact test, pre-specified the use of the mid-p approach from the outset, and consistently used it in all of his analyses, his use of it may be considered reliable.
>
> The problem with Dr. Jewell's use of the mid-p test is that his use of it was results driven. He only used this test once the Fisher exact test returned a non-significant result.[45]

The impact of the selection of statistical tools on the results cannot be understated. Some people have a naïve ideal about scientists that they are slaves to their observations and their data. So long as scientists can be trusted to refrain from concocting data, the ideal goes, they conduct their experiments, gradually accrue a set of data which are firm and unyielding as iron, and then assess the data. Whether the outcome is or is not what the scientist expected, the data and the mathematical tools used to assess it are inflexible, and they must accept what they get. This is very far from the truth. Scientists, especially those who testify as experts in litigation, can get a result that they like (without just making up data) by:

1. Excluding some data from consideration, ideally data that shares some characteristic creating an excuse to exclude it;
2. Regrouping the data with data from another study (or subsets of data from the two studies) and analyzing that;
3. Breaking up the data into subgroups and testing each subgroup until one is found that yields a significant outcome;
4. Cycling through different statistical tools until one yields a significant result; or
5. Foregoing statistical significance and purporting to identify a "trend" or "tendency" in the data.

[45] *Id.* at 15-16.

D. Methodology Hacking

This category refers to something that defendants will see as much from the plaintiffs' lawyers themselves as from their expert witnesses. In methodology hacking, plaintiffs experts misappropriate quotations from scientific and scientific-legal sources to justify an unscientific methodology.

It is very common, for example, for plaintiffs who lack repeated statistically-significant studies in support of a causation conclusion to cite an epidemiology textbook by Kenneth Rothman (such as his MODERN EPIDEMIOLOGY) to the effect that statistical significance is overrated and that it is appropriate to consider non-statistically-significant results.

More recently, the plaintiffs' bar has been touting a recent statement by the American Statistical Association that "The widespread use of 'statistical significance' (generally interpreted as 'p≤0.05') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process."[46] *In re Zoloft* plaintiffs seeking to protect Dr. Jewell's unscientific opinions and excuse their failure to identify statistically-significant evidence have already presented this statement to the Third Circuit.[47]

The problem with this kind of misappropriation of quotes is that it does not match up with anyone's actual real-world methodology, and it totally misrepresents the views of the scientific community in order to excuse unscientific analyses. When someone like Kenneth Rothman or the ASA inadvertently provides the plaintiffs' bar with a sound bite to use against statistical significance, that person is in fact advocating for equally rigorous methods of getting at the truth. He is not merely proclaiming that a weak association is good enough for science.

For example, the ASA paper quoted by the *In re Zoloft* plaintiffs in defense of Dr. Jewell makes clear that p-values are not the only valid way of assessing the significance of a set of data. Other ways may be equally legitimate: prediction intervals, Bayesian methods, decision-theoretic modeling, and so on.[48] In fact, in the same paper, the ASA drives a dagger into the heart of Dr. Jewell's methods (though he was no doubt far from their minds when developing their statement):

[46] Wasserstein & Lazar, *The ASA's Statement on P-Values*, THE AMERICAN STATISTICIAN 131 (Oct. 19, 2016).

[47] *Adams v. Wolters Kluwer Health Inc.*, No. 16-2247 (Appellant's Opening Brief, 3d Cir. Aug. 10, 2016).

[48] *Id.* at 132.

> Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference, and 'p-hacking,' leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided.[49]

What constitutes correct scientific method has always been a subject of controversy. In the literature of the philosophy of science, for example, claims have ranged from asserting that science develops new "facts" in a completely non-incremental fashion[50] to advocacy of essential anarchy in scientific methodology.[51] These more-or-less armchair critiques have more practical counterparts in a decades-old fight about the utility of p-values and statistical significance[52] and in contemporary critiques of the accuracy of most of what is published in scientific and medical journals. Much of the current dialogue—improperly being exploited by the plaintiffs' bar to suggest that science has very low standards—arises from concerns that *positive* associations are being improperly generated and overreported.[53] In other words, the misappropriated quotations come from methodological debates as much concerned with the overstatement of causal associations as with the understatement of them.

Thus, what might otherwise be an in-house argument by scientists about their own techniques has been corrupted by the desire of the plaintiffs' bar to avoid *Daubert*'s impact on their inventories of cases. For example, look to the old debate over p-values, which has always been between those who believe they are useful as a primary indicator of reliability and those who believe that *other tools are better*. One crucial issue motivating all participants is their perception that science reports far too many *false positives*, *i.e.*, allegedly statistically-significant associations that will never be confirmed by repeated studies because they are not true.

The plaintiffs' bar quotes the critics of p-values out of context to make an argument that the critics themselves would never make: that bad scientific results should be accepted despite failing the various tests of significance to which their

[49] *Id.* at 131-32.

[50] Thomas Kuhn, STRUCTURES OF SCIENTIFIC REVOLUTIONS (3d ed. 1996).

[51] Paul Feyerabend, AGAINST METHOD (4th ed. 2010).

[52] *See, e.g.,* R. Nuzzo, *Scientific Method: Statistical Errors*, NATURE, Feb. 12, 2014.

[53] *See, e.g.,* J. Ioannidis, *Why Most Published Research Findings Are False*, PLOS MED, Aug. 2005.

authors subjected them. The plaintiffs' bar does not (for example) propose that p-testing be replaced with a rigorous Bayesian analysis, but instead contends that the whole matter is one of "weight" that should be presented to a jury.

Methodology-hacking should be exposed to the courts for what it is: misappropriation of phrases rather than identification of a genuine scientific methodology. In the *In re Zoloft* litigation, for example, the pharmacoepidemiologist expert justified reliance on non-significant results by quoting Kenneth Rothman and claiming that there had been "an evolution of the thinking of the importance of statistical significance."[54] The court very properly saw through this, holding that there was in fact no recognized methodology in the field of epidemiology consistent with what the expert was trying to do.

CONCLUSION

The cost of keeping junk science off the witness stand at trial is a daunting one: it compels lawyers to delve deeply into the science in order to understand and evaluate it, take detailed depositions to develop a record, and then—most challenging of all—boil all of that information down into a compact presentation for *Daubert* briefing and (ideally) a hearing. Judges must devote the time and effort that it takes to understand the challenges presented to them and to rule on them. If the lawyers do their job well, and file excellent briefs, the judge's job is made much easier.

Those who practice in this area have seen an increase in scientific literacy in the bar since 1993, at least among those engaged in *Daubert* practice. Judges are more likely not to shrink from terms like "statistically-significant epidemiology," and even relatively inexperienced plaintiffs' lawyers are more likely to know that their experts need to at least appear to have followed a scientific methodology in reaching their conclusions. This is all to the good, as it is in everyone's interest for the science presented in court to be genuine science, and not junk.

Statistical significance has been fought over from the start of the *Daubert* era, with the current battleground being an unjustified attack on the notion that scientists care about it at all. It has been a fundamental scientific tool since it was developed, and that remains unchanged today. With a good grounding in the basic principles of statistical significance and in the particular principles of the scientific disciplines at issue in a particular case, any lawyer can be very well-equipped to bring and win a *Daubert* challenge. Likewise, any judge with a good grounding in those same principles can be prepared to hear a *Daubert* challenge and decide it competently.

[54] *In re Zoloft*, 26 F. Supp. 3d 449, 456 (E.D.Pa. 2014).