No. 09-1156

IN THE
Supreme Court of the United States

MATRIXX INITIATIVES, INC., ET AL.,
Petitioners,

v.

JAMES SIRACUSANO AND NECA-IBEW PENSION FUND,
Respondents.

**On Writ of Certiorari
to the United States Court of Appeals
for the Ninth Circuit**

BRIEF OF *AMICI CURIAE* STATISTICS EXPERTS PROFESSORS DEIRDRE N. McCLOSKEY AND STEPHEN T. ZILIAK IN SUPPORT OF RESPONDENTS

EDWARD LABATON
Counsel of Record
IRA A. SCHOCHET
CHRISTOPHER J. MCDONALD
LABATON SUCHAROW LLP
140 Broadway
New York, New York 10005
(212) 907-0700
ELABATON@LABATON.COM

November 12, 2010

TABLE OF CONTENTS

Page

TABLE OF CONTENTS

(continued)

TABLE OF AUTHORITIES

Page(s)

CASES

STATUTES AND REGULATIONS

OTHER MATERIALS

TABLE OF AUTHORITIES

(continued)

Page(s)

INTEREST OF *AMICI CURIAE*[1]

Amici are professors and academics who teach and write on economics, statistics, and the history, philosophy, and rhetoric of economics and statistics as used in business, medicine, and other statistical sciences. *Amici* wish to ensure that the Court properly distinguished 'practical' from mere 'statistical' significance in the context of hypothesis testing when deciding *Matrixx Initiatives, Inc., et al. v. James Siracusano and NECA-IBEW Pension Fund.* *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.

Deirdre N. McCloskey is the Distinguished Professor of Economics, History, English, and Communication at the University of Illinois at Chicago. Previously, she was Visiting Tinbergen Professor (2002-2006) of Philosophy, Economics, and Art and Cultural Studies at Erasmus University of Rotterdam. Since earning her Ph.D. in economics from Harvard University, she has written fourteen books and edited seven more, and has published some three hundred and sixty articles on economic theory, economic history, philosophy, rhetoric, feminism, ethics, and law. Her latest books are *How to be Human – Though an Economist* (Univer-

---

[1] Pursuant to Supreme Court Rule 37.6, 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*, their members, 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, and letters reflecting their blanket consent to the filing of *amicus* briefs have been filed with the Clerk.

sity of Michigan Press 2001), *Measurement and Meaning in Economics* (Stephen T. Ziliak, ed., Edward Elgar 2001), *The Secret Sins of Economics* (Prickly Paradigm Pamphlets, University of Chicago Press, 2002), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives* [with Stephen Ziliak; University of Michigan Press, 2008], and *The Bourgeois Virtues: Ethics for an Age of Capitalism* (University of Chicago Press, 2006). Before *The Bourgeois Virtues* her best-known books were *The Rhetoric of Economics* (University of Wisconsin Press 2d ed. 1998) and *Crossing: A Memoir* (University of Chicago Press 1999), which was a *New York Times* Notable Book.

Stephen T. Ziliak is Faculty Trustee and Professor of Economics at Roosevelt University. He has held academic appointments at the University of Iowa, Bowling Green State University, Emory University, and the Georgia Institute of Technology. Professor Ziliak's areas of expertise include the study of welfare and poverty; economic history, rhetoric, and philosophy; and history and philosophy of science and statistics. He has written extensively on economic and statistical methods, having published more than three dozen articles, book chapters, and books on the origin, development, and real-world use of statistical significance and hypothesis testing. Professor Ziliak has published articles in *Journal of Economic Literature*, *Journal of Economic Perspectives*, *Journal of Economic Methodology*, *Journal of Socio-Economics*, *Journal of Economic History*, *Lancet*, *Biological Theory*, *Proceedings of the Joint Statistical Meetings*, *Quarterly Review of Economics and Finance*, and *Review of Social Economy*. He authored *The Cult of Statistical Significance: How the*

Standard Error Costs Us Jobs, Justice, and Lives (with Deirdre McCloskey, University of Michigan Press, 2008) and he is Associate Editor of *Historical Statistics of the United States: Colonial Times to the Present* (Cambridge University Press, 2006). Professor Ziliak earned his Ph.D. in economics and his Ph.D. Certificate in the rhetoric of the human sciences at the University of Iowa.

INTRODUCTION

Matrixx Initiatives, Inc. is a drug manufacturer that produces Zicam cold remedies. Respondents (Plaintiffs below) have alleged that Matrixx and several of its executives (Petitioners) violated federal securities laws by failing to disclose adverse events linking certain Zicam products with anosmia (the loss of the sense of smell), and issuing misleading statements about Zicam products. The District Court dismissed Respondents' complaint finding that the number of adverse events of which Respondents were aware was not statistically significant. The District Court reasoned that because the adverse events Petitioners withheld from public disclosure could not meet a standard of statistical significance, Respondents failed to allege the element of materiality.

This decision was overturned by the Ninth Circuit Court of Appeals, which found that the District Court erred in applying a strict, bright-line standard of statistical significance to determine materiality. In particular, the Court of Appeals found that a determination of materiality should instead be left to the trier of fact.

In its Supreme Court brief, Petitioners have argued that a standard of statistical significance should be applied when weighing the materiality

of undisclosed adverse events in Section 10(b) cases. Petitioners argue that a statistical significance standard is a commonly used method to determine whether an observed effect is due to chance or real association—be it cause or correlation.[2] Moreover, Petitioners claim that a standard of statistical significance ought to define whether undisclosed information is material to the financial marketplace in securities litigation.[3] As academic and practicing statisticians, we respectfully disagree.

SUMMARY OF ARGUMENT

Statistical significance testing is more complex than simply calculating numbers and determining whether a 5 percent standard has been met. For example, when performing a test of statistical significance, a researcher must weigh the costs of accepting hypotheses that are false against the costs of rejecting hypotheses that are true. To reduce the chance of the latter error (a "Type I" error), researchers can lower their standards of statistical significance, but this would result in an increase in the former type of error (a "Type II" error). The balance must be made by the investigating party in each case.

Problems in significance testing have also been compounded by researchers who prefer to analyze only statistical significance at the expense of practical importance. 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

---

[2] Br. for Pet'r at 34, *Matrixx Initiatives, Inc. v. Siracusano* (No. 09-1156).

[3] *Id.* at 42-44.

particularly grave. As a result, a researcher must strike the correct balance between statistical significance and practical importance. Consequently, a bright-line test of statistical significance would fail to capture important nuances of applied significance testing.

That Petitioners want to place the responsibility of balancing statistical significance and practical importance in the hands of drug manufacturers themselves is troubling. When faced with a particularly important set of undisclosed adverse events, a drug manufacturer may have incentive to avoid disclosing these events until it has received data sufficient to meet a particular standard. In a nutshell, an ethical dilemma exists when the entity conducting the significance test has a vested interest in the outcome of the test.

Finally, it bears mention that significance testing may be particularly cumbersome in the case of adverse events. Because adverse events may gradually trickle in over time, and because it may be implausible to determine the number of individuals actively taking a medication within a given time period, hypothesis tests can be subject to error. Indeed, this is why clinical tests are generally preferred when one is attempting to determine whether certain adverse events are indeed drug-related. As such, a bright-line standard of statistical significance makes little practical sense when determining whether undisclosed adverse events are material.

ARGUMENT

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

To better understand why the Ninth Circuit's opinion should be affirmed, a basic understanding of hypothesis testing is necessary. Below, we explain what a hypothesis test is, the way the results can be interpreted after such a test is performed, and the two types of errors that a researcher can make when performing a hypothesis test.

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 when a researcher seeks to determine whether data exhibit certain properties or accord with a specific statistical distribution. For example, suppose that a researcher had data on fuel efficiency of a specific model automobile driven at constant speed when outfitted with two different tires. The researcher could use hypothesis testing to determine the level of confidence under which one could conclude that one tire results in improved fuel efficiency.

To provide further detail behind the process, a hypothesis test is first performed by posing a null hypothesis, which is the hypothesis that the statistician wishes to test (for example, the null hypothesis that there is "nil" difference between the two front tires of a motor vehicle in terms of

contributions to fuel efficiency). A test statistic is then calculated, assuming that the "null" hypothesis is true. The testing procedure then involves the determination of whether the test statistic falls into one of two subsets of values: a region under which the null hypothesis is rejected (meaning the tires may actually result in different levels of fuel efficiency) and one under which the null hypothesis cannot be rejected (that is, there is insufficient evidence currently to conclude at some level of significance that the two tires result in different levels of fuel efficiency).

The stylized example above may lend 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. As we explain below, there are various nuances of hypothesis testing that must be considered. A failure to reject a null hypothesis does not mean that one then accepts it as the truth. 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

Confusion often surrounds the correct interpretation of the data after a significance test is performed. A rejection of the null hypothesis does not necessarily mean that one accepts the alternate hypothesis.[4] Similarly, failing to reject the null hypothesis does not mean that one should

---

[4] *See, e.g.*, David R. Anderson, Dennis J. Sweeney, & Thomas A. Williams, Modern Business Statistics 351 (3d ed. 2006).

then accept it.[5] In particular, if one estimates the size of an effect by calculating a statistical parameter but is unable to reject the null hypothesis that this parameter is equal to zero, one does not then accept the null hypothesis and apply a value of zero to this parameter. Instead the best estimate of the size of the effect is the value of the parameter that one has calculated.[6]

To further illustrate this point, consider the following example. In economics research of some national importance, one estimates a simple consumption function which presumes that consumer spending in the economy is a function of national income, measured by gross domestic product (GDP). Upon estimation, the so-called marginal propensity to consume is derived from the estimated parameter on the *GDP* variable. Suppose that one estimates that the marginal propensity to consume is 0.7—that is, a one dollar increase in GDP then results in a 70 cent increase in consumer spending. Further suppose, however, that this parameter has a *p*-value less than 0.05, which means that it is "statistically insignificant at 5 percent." On the basis of this test, one should not then assume that the marginal propensity to consume is zero. Indeed, one's best estimate of the marginal propensity to consume is still 0.7. Put simply, if an empirical model is correctly specified, the best estimate of an effect is the one derived from that model, regardless of

---

[5] *Id.*

[6] Steven Goodman, *A Dirty Dozen: Twelve P-Value Misconceptions*, 45 Seminars in Hematology 135, 136 (2008) (stating that "[t]he effect best supported by the data from a given experiment is always the observed effect, regardless of its significance").

whether a particular level of statistical significance is achieved.

C. There Are Two Types of Errors in Hypothesis Testing: Type I and Type II

Errors are of potential concern when a hypothesis test is performed, and two types of errors can stem from such a test. Type I error occurs when one rejects the null hypothesis when the null hypothesis is true. Type II occurs when one does not reject the null hypothesis when it is false. The chart below helps to clarify this concept.

		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

As the table above illustrates,[7] Type I error occurs when the researcher rejects the null hypothesis accidentally. Alternatively, Type II error occurs when the researcher is unable to reject the null hypothesis, when in fact she should.

The importance of Type I error to this matter is that when one identifies the level of significance in a hypothesis test, one has also identified the probability of making a Type I error. For example, suppose that one is testing whether a particular medication lowers cholesterol relative to a place-

---

[7] Tables similar to this are commonly presented in statistics texts. *See, e.g.*, Richard J. Larsen & Morris L. Marx, An Introduction to Mathematical Statistics and its Applications, at 299 (2d ed. 1986).

bo. In this test, one could construct the following null and alternate hypotheses.

Null hypothesis: *Experimental medication has effect on subject cholesterol equivalent to that of a placebo.*

Alternate hypothesis: *Effect on subjects' cholesterol from experimental medication is different than the effect from a placebo.*

Were one to perform this test at, for example, a 5 percent 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 percent.

A Type I error, however, is not the only "bad" outcome when performing the above test. From a testing perspective, another poor outcome would be to fail to reject the null hypothesis when in fact the experimental 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 we explain 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.[8] *Second,* the natural sacrifice of a reduction in Type I error is an increase in Type II error.[9] 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 experimental medication above. Suppose that in conducting the hypothesis test, the researcher is overly focused on minimizing Type I error. That is, 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 experimental medication has no effect might seem like a good thing. The cost, however, is that it increases the chance of concluding the medication *does not work* when it is indeed effective. This is also a bad outcome and could result in scrapping perfectly effective medicine. Consequently a balance must be struck between these errors, and a singular focus on statistical significance can indeed be inappropriate.

This is discussed in more detail in Section II below.

---

[8] *See, e.g., id.* at 303.

[9] *Id.* at 300-03.

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.

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.[10] How the balance is struck, the determination made, is a matter of scientific, ethical, and political deliberation which needs to happen prior to 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 tire and fuel efficiency example. Researchers have noted, however, that too much emphasis has been placed solely on statistical significance at the expense of practical importance.[11] The balance

---

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

[11] *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

puts no weight on the practical importance of the outcome. That is, researchers can become too focused on simply 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. As such, too much emphasis could be placed on results found to be statistically significant that have little practical importance, and results or phenomena that are of large practical importance could be ignored or rejected simply because they do not meet a particular criterion of statistical significance.

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. That is, it would be impossible to construct an across-the-board rule that could take into account the case-by-case balancing between practical importance and statistical significance that may be desired. Moreover, as we explain in more detail below, one would only worsen this problem by placing the responsibility of conducting such a test in the hands of companies that have a vested interest in certain test outcomes.

---

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).

B. The Potential Harm from Type II Error Is Large When Practical Importance Is High

As stated above, the problem of using a significance level that is too low is that one is more likely to fail to reject the null hypothesis when it would be appropriate to do so. (In the earlier example of an experimental cholesterol medication, this amounts to concluding the medication does not work when in fact it does.) This problem is magnified when the effect that is being studied is of great practical importance. That is, 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 five-percent level.[12] 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 importance.[13] Therefore, care must be taken in applying a significance test to medical data, as an appropriate balance must be struck between Type I and Type II error.

A particular example that is relevant to this discussion is the analysis of adverse events relating to Rofecoxib (brand name Vioxx). An initial analy-

---

[12] *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).

[13] *Id.*

sis of the effectiveness of Vioxx, an anti-inflammatory medication, found that individuals in the Vioxx group suffered from an increased rate of adverse heart-related events (such as infarction and stroke) relative to the control.[14] This increased rate of adverse events, however, was not deemed statistically significant.[15] Therefore, the risk of Vioxx toward heart-related adverse events was downplayed when the drug was first introduced to the market. Once it became evident that heart-related risks did indeed exist, Vioxx was withdrawn from the market.[16] 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[17] was no longer available. Both of these problems could have been avoided had less attention been paid to p-values.

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 have begun to reexamine the origins of significance testing. The findings of this analysis have revealed that the most commonly held

---

[14] 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).

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

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

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

B. The Potential Harm from Type II Error Is Large When Practical Importance Is High

As stated above, the problem of using a significance level that is too low is that one is more likely to fail to reject the null hypothesis when it would be appropriate to do so. (In the earlier example of an experimental cholesterol medication, this amounts to concluding the medication does not work when in fact it does.) This problem is magnified when the effect that is being studied is of great practical importance. That is, 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 five-percent level.[12] 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 importance.[13] Therefore, care must be taken in applying a significance test to medical data, as an appropriate balance must be struck between Type I and Type II error.

A particular example that is relevant to this discussion is the analysis of adverse events relating to Rofecoxib (brand name Vioxx). An initial analy-

---

[12] *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).

[13] *Id.*

sis of the effectiveness of Vioxx, an anti-inflammatory medication, found that individuals in the Vioxx group suffered from an increased rate of adverse heart-related events (such as infarction and stroke) relative to the control.[14] This increased rate of adverse events, however, was not deemed statistically significant.[15] Therefore, the risk of Vioxx toward heart-related adverse events was downplayed when the drug was first introduced to the market. Once it became evident that heart-related risks did indeed exist, Vioxx was withdrawn from the market.[16] 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[17] was no longer available. Both of these problems could have been avoided had less attention been paid to p-values.

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 have begun to reexamine the origins of significance testing. The findings of this analysis have revealed that the most commonly held

---

[14] 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).

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

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

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

benchmark of statistical significance—that is, five percent—came into prominence because it was an initial suggestion of preference of Ronald Fisher, one of the fathers of the hypothesis test.[18] In particular, Fisher preferred a five-percent rule because at a critical value of 1.96—that is, the critical value for a normally distributed test statistic using the five-percent rule—one would reject the null hypothesis were one's result more than 1.96 standard deviations from the mean. Put differently, Fisher's own preference was to categorize results as significant when they were more than two standard deviations from expectation under the null hypothesis.[19]

Although applied statisticians commonly use the five-percent level when performing significance tests, the fact that this standard exists because an early developer of the test deemed it appropriate highlights an underlying problem of blindly applying this 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*.[20] 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.[21] Moreover, these professors noted

---

18 Lisse et al., *supra* note 14, at 543-44.

19 *Id.*

20 *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993).

21 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.

that significance testing is often mistaken as a fundamental input of scientific analysis.[22]

D. If the Balance between Type I and Type II Error Should be Left to the Researcher, a Statistical Significance Standard Would Create a Conflict of Interest

The balance between Type I and Type II error is typically determined by the researcher that is conducting the test in question.[23] Therefore, a statistical significance standard in Section 10(b) cases, such as the one sought by Petitioners in this matter,[24] would potentially create a conflict of interest. In particular, the party that may be in the best position to analyze the statistical significance and practical importance of adverse event reports ("AERs") is the drug manufacturer itself, because the manufacturer has data on AERs that have been reported to it by consumers, physicians, and pharmacists. However, the drug manufacturer's incentives are not necessarily aligned with those of either its customers or its investors.

Petitioners in this matter state that a significance standard of five-percent is a general standard and that the Supreme Court has used this standard in the past.[25] However, applied toward the reporting of AERs, such a standard would place an incredible amount of discretionary power in the hands of a party that has particular incentive *not to reject the null hypothesis of the test.* For example, the drug manufacturer might ignore

---

[22] *Id.* at 3.

[23] *See, e.g.,* David R. Anderson, Dennis J. Sweeney, & Thomas A. Williams, *Modern Business Statistics* 351 (3d ed. 2006).

[24] Br. for Pet'r at 33, *Matrixx Initiatives, Inc., v. Siracusano* (No 09-1156).

[25] *Id.* at 35.

AERs that, currently, test at p-values of, say, eight or nine percent by deciding that a five percent significance rule is appropriate.

To further see that *any* bright-line standard of statistical significance would be problematic, consider the following. The 5 percent significance rule insists on 19 to 1 odds that the measured effect is real.[26] There is, however, a practical need to keep wide latitude in the odds of uncovering a real effect, which would therefore eschew any bright-line standard of significance. Suppose that a p-value for a particular test comes in at 9 percent. Should this p-value be considered "insignificant" in practical, human, or economic terms? We respectfully answer, "No." For a p-value of .09, the odds of observing the AER is 91 percent divided by 9 percent. Put differently, there are 10-to-1 odds that the adverse effect is "real" (or about a 1 in 10 chance that it is not). Odds of 10-to-1 certainly deserve the attention of responsible parties if the effect in question is a terrible event. Sometimes odds as low as, say, 1.5-to-1 might be relevant.[27] For example, in the case of the Space Shuttle Challenger disaster, the odds were thought to be extremely low that its O-rings would fail. Moreover, the Vioxx matter discussed above provides an additional example. There, the p-value in question was roughly 0.2,[28] which

---

[26] At a 5 percent p-value, the probability that the measured effect is "real" is 95 percent, whereas the probability that it is false is 5 percent. Therefore, 95 / 5 equals 19, meaning that the odds of finding a "real" effect are 19 to 1.

[27] Odds of 1.5 to 1 correspond to a p-value of 0.4. That is, the odds of the measured effect being real would be 0.6 / 0.4, or 1.5 to 1.

[28] Lisse et al., *supra* note 14, at 543-44.

equates to odds of 4 to 1 that the measured effect—that is, that Vioxx resulted in increased risk of heart-related adverse events—was real. The study in question rejected these odds as insignificant, a decision that was proven to be incorrect.

One might also consider an example in which a bright-line standard provides a specific benchmark of significance—say, 5 percent. As more AERs gradually become evident, a company may watch a p-value for the significance of the rate of adverse events fall from, say, 10 percent to 8 percent and then to 7 percent. The company might predict that with sufficient time, the 5 percent standard will indeed be met. However, until that standard is met, the result is not significant and no action is needed. Consequently, a bright-line standard of statistical significance in this setting could create a conflict of interest in that a drug manufacturer might have incentive to respond differently to tests of significance than would an impartial researcher.

III. A STATISTICAL SIGNIFICANCE STANDARD WOULD REJECT VALID ADVERSE EVENTS

Problems particular to the analysis of adverse events and to adverse event reporting render significance testing in the area difficult if not unreliable. For example, practitioners in the field of medical research have noted that the problem of sample size can be particularly acute. That is, if a drug does indeed cause an adverse event with greater frequency than a placebo, a very large sample size may be required to detect this difference statistically.[29] Medical data, in particular,

---

[29] Altman, *supra* note 12, at 1336-37.

tends to be characterized by small sample sizes.[30] Moreover, even analyses with relatively large sample sizes have been known to fail to uncover true adverse effects in experimental drugs at a five-percent level of significance.[31]

This said, the issue of sample size in uncovering adverse effects still exists. As a result, recommended industry standards are that adverse events should be pursued diligently whether they are significant or insignificant in a statistical sense.[32] In addition, the FDA does not require a statistically significant association between a drug and a given effect to warrant a label change such as a precaution or warning.[33] The sample size problem described above could be compounded by practical problems that exist in adverse event reporting. To see this, first consider the manner

---

30 *Id.*

31 *See, e.g.*, Ste phen 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); Lisse et al., *supra* note 14, at 543-44 (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 percent level of significance). There were more than 5,000 individuals in this particular Vioxx study. *Id.*

32 *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 rechallenge or if the event is extremely rare in the absence of drug use.").

33 *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.").

in which adverse effects could be unearthed in a clinical trial. In a clinical experiment, which is often preferred when performing statistical analysis in epidemiology, test and control groups are well defined and carefully monitored. This makes statistical testing more straightforward, and potentially more powerful relative to an analysis of data obtained from the field.[34]

Because not all adverse events are reported, and because those that are reported may trickle in over time, a *statistical* study of AERs using field data is pre-disposed to understate the true incidence of adverse events. In addition, assessing the number of persons taking the medication within a specific period of time may also be difficult. Consequently, performing a statistical significance test on AER data, particularly before a large amount of that data has been compiled, may be an exercise in futility. That is, with downward bias in the incidence of adverse events, and with a potentially inaccurate measure of the number of users of the drug, a significance test is unlikely to render accurate results.

---

[34] *See* Linda Baily, Leon Gordis, & Michael Green, Reference Guide on Epidemiology, *in* Federal Judicial Center, Reference Manual on Scientific Evidence 343 (2d ed. 2000).

CONCLUSION

The judgment of the Ninth Circuit should be affirmed.

Respectfully submitted,

EDWARD LABATON
Counsel of Record
IRA A. SCHOCHET
CHRISTOPHER J. MCDONALD
LABATON SUCHAROW LLP
140 Broadway
New York, NY 10005
(212) 907-0700
ELABATON@LABATON.COM

November 12, 2010