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

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.

19

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.

20

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.").

21

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).

Discussion Questions:

1. We have talked about the typical desire of an analyst to have the largest feasible "n" to get the most precise results. Why are the incentives of a drug company different from this usual case? Do you have ideas about how to address this challenge?
2. Why do the authors argue that using specific statistical significance thresholds is problematic in this context? In what contexts might using specific significance thresholds of 5% or 10% be more useful? What other types of significance (besides statistical) should be considered?