

Reproduced with permission from Product Safety & Liability Reporter, 39 PSLR 1007, 09/12/2011. Copyright © 2011 by The Bureau of National Affairs, Inc. (800-372-1033) <http://www.bna.com>

The U.S. Supreme Court's ruling in *Matrixx Initiatives Inc. v. Siracusano* offered the justices an opportunity to speak clearly and authoritatively about the meaning and limits of significance testing, says Professor David H. Kaye in this BNA Insight. "Unfortunately, the Court did not rise to this challenge," the author says. This article examines the ruling, as well as some potentially troublesome dicta on causation and statistical significance in toxic tort and pharmaceutical injury litigation.

Trapped in the *Matrixx*: The U.S. Supreme Court And the Need for Statistical Significance



BY DAVID H. KAYE

One might expect to hear phrases like "Not statistically significant there!" and, "There is no way that anybody would tell you that these ten cases are statistically significant" hurled by a disgruntled professor at an underperforming statistics student. Yet, in

January 2011, they came from the Supreme Court bench during the argument in *Matrixx Initiatives Inc. v. Siracusano*.¹

The case offered the Supreme Court an opportunity to speak clearly and authoritatively about the meaning and limits of significance testing. Unfortunately, the Court did not rise to this challenge. Instead, it failed to explain how a series of physician's case reports about a drug's side-effect could be examined for "statistical significance," and it did not distinguish clearly between causal inference and statistical significance in its many references to biomedical proof of causation.

This article discusses the Court's unanimous opinion and argues for a narrow reading of its dicta about proof of causation. Part I describes the case. Part II explains more precisely than the Court did why plaintiffs alleged that a pharmaceutical company failed to disclose

¹ Transcript of Oral Argument, *Matrixx Initiatives Inc. v. Siracusano*, 131 S. Ct. 1309 (2011) (No. 09-1156), 2011 WL 65028, at *12 & *16 (Kagan, J.).

adverse event reports from physicians should not have to plead a statistically significant number of these case reports. It also clarifies the meaning of “statistical significance” as applied to such anecdotal data. Part III presents the *Matrixx* Court’s dicta on causation and statistical significance and shows that the Court’s remarks do not address the limited value of adverse event reports (AERs) in establishing causation in toxic tort litigation.

I. Speak No Evil: Zicam and Anosmia

An Arizona company, *Matrixx Initiatives Inc.*, developed and sold over-the-counter cold remedies. In 2009, the FDA “concluded that these products may pose a serious risk to consumers who use them” and ordered the company to take corrective action.² The Agency based this conclusion on “more than 130 reports of anosmia (loss of sense of smell, which in some cases can be long-lasting or permanent), associated with use of these products . . . [compared to] few reports of anosmia associated with other widely-used intranasal products for treatment of the common cold . . . [and] evidence in the published scientific literature that various salts of zinc can damage olfactory function in animals and humans.”³

Long before the case reports accumulated, however, a group of investors who had bought *Matrixx* stock in late 2003 and early 2004 filed a securities fraud class action. They alleged that in this early period, *Matrixx* issued reassuring statements that misled them. Specifically, plaintiffs alleged that *Matrixx* did not disclose reports from physicians about consumers who lost their sense of smell after using its homeopathic remedy, *Zicam*. The complaint stated that the company knew of at least 12 cases of anosmia following nasal inhalation,⁴ and that it also knew of animal and human studies showing that zinc sulfate could cause anosmia. Nevertheless, as news reports of product liability suits and the FDA inquiry seeped out, *Matrixx* issued press releases insisting that “the safety and efficacy of zinc gluconate for the treatment of symptoms related to the common cold have been well established in two double-blind, placebo-controlled, randomized clinical trials. In fact, in neither study were there any reports of anosmia related to the use of this compound. The overall incidence of adverse events associated with zinc gluconate was extremely low, with no statistically significant difference between the adverse event rates for the treated and placebo subsets.”⁵

The federal district court dismissed this complaint, but the Court of Appeals for the Ninth Circuit reinstated it. The Supreme Court granted certiorari to consider “[w]hether a plaintiff can state a claim under § 10(b) of the Securities Exchange Act and SEC Rule 10b-5 based on a pharmaceutical company’s nondisclosure of adverse event reports even though the reports are not al-

leged to be statistically significant.”⁶ The issue, in other words, was not whether *Zicam* causes anosmia. It was whether the reports of anosmia in *Zicam* consumers “would . . . be considered significant to the trading decision of a reasonable investor.”⁷

The answer is clear. Justice Sotomayer’s opinion for the Court explained that a reasonable investor might want to know of such reports if they (along with other information) are sufficiently extensive and disturbing that they could prompt the FDA to take some action or might lead to costly lawsuits. That is enough to trigger a duty to disclose in order to prevent other company statements from being misleading.⁸

II. Beneath the Simple Answer: Why Allegations of Statistical Significance Are Not a Viable Pleading Rule

But how much potentially disturbing information is enough to necessitate disclosure? *Matrixx* argued forcefully for a bright-line rule of statistical significance (at the 0.05 level).⁹ A statistic is “significant” when its value falls outside a range that would be expected on the basis of a probability model of the process that generated the data. The model includes an error term that is supposed to capture all the random influences on the data. Observing a number of anosmia case reports that rarely would occur if there were no association whatever between *Zicam* and loss of the sense of smell suggests that the hypothesis of no association is implausible. If all other features of the statistical model are correct, one then can reject the “null hypothesis” and conclude that the observed degree of an association is not just happenstance. At that stage, further causal analysis is required to ascertain what might be causing the observed association.

For historical reasons, the most common “significance level” in biomedical and social science research is 0.05. This level ensures that inferring that something other than randomness is at work when, in fact, randomness is all there is to it occurs no more often than 1 time in 20 (in the long run). In other words, if Nature is perverse and arranges things so that a scientist never encounters a true association, always demanding significance at the 0.05 level protects the scientist from getting it wrong by declaring a false association more than 5% of the time (on average).

Matrixx wanted companies manufacturing biomedical products to have the same level of protection against investors’ complaints of fraud. Unless the plaintiffs alleged that the undisclosed case reports were numerous enough to have occurred with a probability (a “p-value”) of less than 0.05 given some model of the world that presumes absolutely no association between the adverse events and the company’s products, a court would have to toss out the complaint.

⁶ Petition for Writ of Certiorari at *i, *Matrixx Initiatives Inc. v. Siracusano*, 131 S. Ct. 1309 (2011) (No. 09-1156), 2010 WL 1063936.

⁷ 131 S. Ct. at 1320 (internal quotation marks and citation omitted).

⁸ *Id.* at 1321.

⁹ On the meaning of statistical significance, see *David H. Kaye et al., The New Wigmore: A Treatise on Evidence: Expert Evidence* (2d ed. 2011).

² Joint Appendix, 2010 WL 3337908, at *268a.

³ *Id.* at 270a.

⁴ *Matrixx*, 131 S. Ct. at 1315; Brief for Petitioner at 16, *Matrixx Initiatives Inc. v. Siracusano*, 131 S. Ct. 1309 (2011) (No. 09-1156), 2010 WL 3334501 (referring to between 12 and 23 reports).

⁵ Joint Appendix, *supra* note 2, at *78a & *82a.

Yet, neither the company nor the Court was clear about how a p -value for case reports could be computed. One would have to consider, not just the number of case reports, or the proportion of these reports out of all purchasers of Zicam (the figure that Matrixx's briefs emphasized), but the number expected under a model for the probability of anosmia in a world in which anosmia has no association with the use of Zicam. Matrixx's lawyers grandly suggested "consider[ing] the background rate in a relevant population of the reported event"¹⁰ or, more specifically, "among people with the common cold who do not use Zicam."¹¹ The FDA's 2009 letter suggested looking to the proportion of users of other cold remedies who experience anosmia to produce an expected value for the Zicam users. If one posits that the users of the products in each group are like a random sample of some larger populations, then a p -value could be computed.¹² I do not know whether the parties tried to do so, but the plaintiffs chose not to amend their complaint to allege a statistically significant difference in the incidence of anosmia among Zicam users and any other group of people, and the case worked its way to the Supreme Court.

The Court was correct to reject a rule requiring pleading $p < 0.05$ in all Rule 10b-5 actions. A rigid 0.05 rule would have been somewhat arbitrary. It is hard to justify the particular threshold of 0.05 as opposed to, say, 0.04 or 0.06. Yet, many legal rules are no less arbitrary than this particular statistical convention. A more fundamental objection is that using any such cutoff at the pleading stage would not have achieved the purpose of a rule for screening out meritless cases based on the pleadings. Demanding statistical significance would have helped employ statisticians willing to look for models and data to achieve the desired threshold, but it would not accurately have filtered out the cases for which no reasonable investor would care about the "insignificant" number of case reports. Combined with other evidence about a drug's safety, even a number for which p exceeds 0.05 or the like could justify investor concern about the drug's future in the marketplace.

Interestingly, Matrixx did not question this possibility. It conceded that "a claim can be pled absent statistically significant evidence, but that's . . . because doctors and researchers will conclude that there may be causation under . . . the Bradford-Hill [or similar] criteria. But nothing like that is pled here . . ." ¹³ Matrixx's actual argument—that a number of adverse event reports (AERs) falling within the range that would be expected to arise by chance alone when the Zicam has no association with anosmia and when AERs are the only

evidence of risk that plaintiffs allege—is not so implausible as the one the Court dispatched.¹⁴

Ultimately, however, the proposed rule is undesirable even in the subset of cases in which AERs are all that plaintiffs can plead. In practice, significance testing is not the purely mechanical, objective process that some courts think it is.¹⁵ Should the incidence of anosmia among Zicam users (to the extent this can be inferred from AERs) be compared to the incidence in the entire population? The population of all cold sufferers? The population of cold sufferers who use over-the-counter cold remedies? Cold sufferers who use nasal inhalants? A useful rule of complaint drafting must avoid inquiries into the soundness of expert judgments about the population, the test statistic, and the model. But if the court cannot look behind the words of the complaint, the rule will simply encourage strategic behavior by inventive statisticians. Indeed, if the courts allow Bayesian analyses to substitute for classical hypothesis testing,¹⁶ as they probably should,¹⁷ the realm for creative statistical work at the complaint stage becomes even greater.

The *Matrixx* opinion does not mention the problem of searching for significance. Guided by the Federal Judicial Center's *Reference Guide to Epidemiology*, it talks fuzzily about "a study" being "statistically significant."¹⁸ It takes the cheerful view that "a study that is statistically significant has results that are unlikely to be the result of random error."¹⁹ Two economists, writing what they called the "Brief of Amici Curiae Statistics Experts" in the case, went even farther in transposing conditional probabilities and confusing mere association with a treatment effect. They complained that "[t]he 5 percent significance rule" is too demanding when the costs of Type II errors are large because the rule "insists on 19 to 1 odds that the measured effect is real."²⁰ Advice like this from "friends of the court" does

¹⁴ See Paul Meier et al., *What Happened in Hazelwood: Statistics, Employment Discrimination and the 80% Rule*, 1984 *Am. B. Found Res. J.* 139, 152 ("If a difference does not attain the 5% level of significance, it does not deserve to be given weight as evidence of a disparity. It is a 'feather.'").

¹⁵ See, e.g., D.H. Kaye, *Is Proof of Statistical Significance Relevant?*, 61 *Wash. L. Rev.* 1333 (1986).

¹⁶ On the nature of Bayesian inference and its use in court, see Kaye et al., *supra* note 9.

¹⁷ See, e.g., Sander Greenland & Charles Poole, *Problems in Common Interpretations of Statistics in Scientific Articles, Expert Reports, and Testimony*, 51 *Jurimetrics J.* 113 (2011).

¹⁸ *Id.* at 1319 n.6. Loose statements about "samples" being significant are criticized in David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, in *Reference Manual on Scientific Evidence* (3d ed. 2011).

¹⁹ *Id.*

²⁰ Indeed, the economists maintained that "a p -value for a particular test [of] 9 percent [should not] be considered 'insignificant' in practical, human, or economic terms [because] 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." Brief of Amici Curiae Statistics Experts Professors Deirdre N. McCloskey and Stephen T. Ziliak in Support of Respondents at 18, *Matrixx Initiatives, Inc. v. Siracusano*, 131 S. Ct. 1309 (2011) (No. 09-1156), 2010 WL 4657930. It is trivial to construct examples contradicting this interpretation. A bag contains 100 coins. One of them is a trick coin with tails on both sides; the other 99 are biased coins that have a 0.3 chance of coming up tails and a 0.7 chance of coming up heads. I pick one of these coins at random and flip

¹⁰ Brief for Petitioner, *supra* note 4, at 13.

¹¹ *Id.* at 14.

¹² See Joseph L. Gastwirth, *Statistical Considerations Supporting the Supreme Court's Decision in Matrixx Initiatives v. Siracusano* (Aug. 12, 2011) (unpublished manuscript).

¹³ Transcript of Oral Argument, *supra* note 1, at *5-6. See Austin Bradford Hill, *The Environment and Disease: Association or Causation?*, 58 *Proceedings Royal Soc'y Med.* 295 (1965) (listing "nine different viewpoints from . . . which we should study association before we cry causation," rejecting the proposition "that we can usefully lay down some hard-and-fast rules of evidence that must be obeyed before we can accept cause and effect," and emphasizing that "formal tests of significance [are useful only to] remind us of the effects that the play of chance can create, and [to] instruct us in the likely magnitude of those effects.").

not make the Court's task any easier. Fortunately, many publications explain why the odds that the null hypothesis is true (or false) is a meaningless statement in the context of classical hypothesis testing and why, in a Bayesian framework, it is not computed from a p -value.²¹

III. From Association to Causation: Sharpening the Distinction

That reasonable investors might want to know about small numbers of adverse events reports in some contexts does not mean that even large numbers of AERs provide valid statistical proof of causation. Thus, Justice Sotomayor wisely cautioned that "we do not attempt to define here what constitutes reliable evidence of causation."²² Yet, in the same breath she wrote about proof of causation in the toxic tort context. This Part explains why courts should treat the dicta on causation with care.

Case reports are anecdotal evidence—a series of little stories of event X (e.g., use of the remedy) followed by event Y (e.g., anosmia). At best, such anecdotes can establish an association between X and Y. That is where statistical significance comes in. Ideally, it justifies an inference that something other than random chance produced the observed association. This "something" might not be Zicam at all, but some other factor associated with Zicam use. That is why clinical trials (or, at a minimum, further analysis of potentially confounding variables in observational studies) are important. They

it twice to obtain two tails. On the basis of only the sample data of two tails, you must decide which type of coin I picked. The p -value with respect to the "null hypothesis" that the coin is the heads-tails one is the probability of seeing two tails in the two tosses: $p = 0.3 \times 0.3 = 0.09$. Should you reject the null hypothesis and conclude that I flipped the unique tails-tails coin? Are the odds for this alternative hypothesis 10:1, as the brief of the statistical experts asserts?

Of course not. Just consider repeating this game over and over. Ninety-nine percent of the time, you would expect me to pick a heads-tails coin. In 9% of those cases, you expect me to get tails-tails on the two tosses (9% 99% = 8.91%). The other way to get tails-tails on the tosses is to pick the tails-tails coin. You expect this to happen in about 1% of the time. Thus, the odds of the tails-tails coin given the data on the outcome of the tosses are 1:8.91, which is about 1:9. Despite the allegedly significant (in "practical, human, or economic" terms) p -value of 0.09, the alternative hypothesis remains quite improbable.

The lesson of this example is not that a statistic with a p -value of 0.09 always can be safely ignored. It is that the p -value, by itself, cannot be converted into a probability that the alternative hypothesis is true ("that the adverse effect is 'real'"). Knowing that the two tails arise only 9% of the time when the head-tails coin is the cause does not imply that 9% is the probability that a heads-tails coin is the cause or that 91% is the probability that the tails-tails coin is the "real" cause.

²¹ See, e.g., Kaye et al., *supra* note 9. Unlike classical statisticians, who must remain silent about the probabilities of hypotheses, Bayesian statisticians regard probabilities for hypotheses as analogous to the "prior" probabilities for picking a type of coin in the example of the preceding note. They can compute how sample data changes a prior probability. In the coin example, the prior odds for a tails-tails coin grew from 1:99 to a little under 1:9. The sample data supported the alternative hypothesis of a tail-tail coin, but not by an amount sufficient to give the posterior odds of 10:1 claimed for it in the amicus brief.

²² *Id.* at 1319.

can help eliminate other factors as explanations for an observed difference. Thus, a statistically significant number of adverse events does not establish causation, but it can trigger further study or regulatory action.²³ Causal inference of the sort required under *Daubert v. Merrell Dow Pharmaceuticals*,²⁴ on the other hand, requires suitably designed and interpreted studies, whether they are inspired by AERs or anything else.²⁵

One might not know this from the dicta in *Matrixx*. Justice Sotomayor wrote that:

A lack of statistically significant data does not mean that medical experts have no reliable basis for inferring a causal link between a drug and adverse events. As *Matrixx* itself concedes, medical experts rely on other evidence to establish an inference of causation. . . . [C]ourts frequently permit expert testimony on causation based on evidence other than statistical significance. . . . It suffices to note that, as these courts have recognized, medical professionals and researchers do not limit the data they consider to the results of randomized clinical trials or to statistically significant evidence.²⁶

Indeed, the Justice lists "a temporal relationship"²⁷ in a single patient as one indication of "a reliable causal link."²⁸ Furthermore, the opinion seems to suggest that even when the number of AERs cannot reasonably be attributed to anything but chance, medical researchers or regulators could treat them as proving causation, at least when combined with other information. As support, the Justice writes that "ethical considerations may prohibit researchers from conducting randomized clinical trials to confirm a suspected causal link for the purpose of obtaining statistically significant data."²⁹ And she notes that "[s]tatistically significant data are not always available."³⁰

It would be a mistake to read too much into these remarks about the multiple types of information that underlie judgments of causation in different arenas. After all, they occur in the context of deciding whether the facts known to *Matrixx* "revealed a plausible causal relationship between Zicam Cold Remedy and anosmia" insofar as "[c]onsumers likely would have viewed the risk associated with Zicam (possible loss of smell) as substantially outweighing the benefit of using the product (alleviating cold symptoms), particularly in light of the existence of many alternative products on the market."³¹ Whether AERs, which are haphazardly collected data, can be used to prove causation in a toxic tort case

²³ In the absence of a remotely significant association, AERs themselves supply no reason to undertake further studies that might clarify the speculative causal links. In simple terms, there is no reason to consider rival explanations for an association if there is no association to explain.

²⁴ 509 U.S. 579 (1993).

²⁵ *Matrixx*'s brief contains an allegedly "representative sample of 50 federal and state opinions" expressing skepticism under *Daubert* of AERs as evidence of causation. Brief for Petitioner, *supra* note 4, at 24.

²⁶ *Matrixx*, 131 S. Ct. at 1319-20 (citations and internal quotation marks omitted). In this paragraph, the Court tried to soften the impact of its comments, stating that in the lower court cases on causation in toxic tort litigation, "[w]e need not consider whether the expert testimony was properly admitted" *Id.* at 1319.

²⁷ *Id.* at 1322.

²⁸ *Id.*

²⁹ *Id.* at 1319.

³⁰ *Id.*

³¹ *Id.* at 1323.

is a very different question from whether they merely “suggest [] a significant risk to the commercial viability of Matrixx’s leading product.”³²

To be sure, other types of studies and knowledge of biological mechanisms are relevant in causal analysis for both regulatory decision and tort verdicts. For example, there are situations when it is reasonable to infer causation from observational data. The relationship between smoking and lung cancer is one. But the fact that “medical professionals and researchers do not

limit the data they consider to the results of randomized clinical trials or to statistically significant evidence”³³ obviously does not mean that anyone reliably can infer causation in the absence of good experimental or observational data. AERs rarely, if ever, constitute such data.³⁴ Investors may care about them, and they can stimulate further research that sometimes pans out, but courts should not accept them as proof of “a reliable causal link.”

³² *Id.*

³³ *Id.* at 1320.

³⁴ *See supra* note 25.

