No. 13-180

IN THE

Supreme Court of the United States

W. SCOTT HARKONEN,

Petitioner,

v.

UNITED STATES,

Respondent.

ON PETITION FOR WRIT OF CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE NINTH CIRCUIT

Kenneth J.Rothman, Timothy L. Lash, and Nathan A. Schachtman, Scientists and Academics as *Amici Curiae* in Support of Petitioner

BECKY WALKER JAMES
Counsel of Record
LAW OFFICES OF
BECKY WALKER JAMES
17383 SUNSET BOULEVARD
SUITE A315
PACIFIC PALISADES, CA 90272
(310) 492-5104
becky@walkerjameslaw.com

Counsel for Amici Curiae

LEGAL PRINTERS LLC, Washington DC • 202-747-2400 • legalprinters.com

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Cases

Rules

Other Authorities

BRIEF OF SCIENTISTS AND ACADEMICS AS *AMICI CURIAE* IN SUPPORT OF PETITIONER

INTEREST AND IDENTITY OF *AMICI CURIAE*[1]

Amici are scientists and scholars who teach and write about statistics and epidemiology. All have long-standing interests in statistical evidence as used in science, regulation, and litigation. *Amici* share a concern that scientists make statements of the sort at issue in many contexts, including:

- grant proposals and grant reports to funding agencies,
- submissions of journal manuscripts,
- peer review and editorial comments in publishing articles,
- submissions to agencies about rulemaking,
- statements to the media about scientific studies, and
- expert witness reports and testimony in litigation.

---

[1] Pursuant to Rule 37.6, *amici* affirm that no party's counsel authored this brief in whole or in part, and that no one other than *amici* and their counsel financially contributed to its preparation or submission. Counsel of record for all parties received notice at least 10 days before the due date of *amici*'s intention to file this brief. Petitioners have filed a blanket consent to *amici*'s participation, with this Court. Respondent's communication consenting to the filing of this brief has been lodged with the Clerk's office.

All branches of government depend upon access to scientific data, interpreted and evaluated by capable scientists, without fear of reprisal. The prosecution and resulting conviction in this case threaten to chill scientific speech in many important activities and contexts, to the detriment of public health.

SUMMARY OF THE ARGUMENT

This *amicus* brief is submitted in support of Dr. Harkonen's petition and request for reversal of the judgment below, which was based upon incomplete and inaccurate descriptions of statistical practice, theory, and inference. In particular, we wish to inform the Court that the proper interpretation of p-values and the inferential worth of "post-hoc" analyses are topics of active scientific debate. Furthermore, the government's prosecution incorporates serious misunderstandings of statistical language and principles, and poses an ominous threat to the integrity of scientific discourse and progress, which the scientific community cannot ignore. This Court should grant this petition because the conviction, if allowed to stand, will place scientists in an untenable position. Many scientists will have to conform their writing, public presentations, and grant proposals to the government's statistical orthodoxy, with which they deeply disagree, or risk facing criminal prosecution.

The Ninth Circuit's decision states that "Harkonen's scientific methods were not on trial; the issue was whether he misleadingly presented his analyses in the Press Release." *United States v.*

Harkonen, 2013 WL 782354, *3 (9th Cir. 2013). This distinction is scientifically unclear and legally unsound: the language used to describe study results are inextricably linked to the methods that led to those results. *Cf. General Electric Co. v. Joiner*, 522 U.S. 136, 146 (1997) ("conclusions and methodology are not entirely distinct").

STATEMENT OF FACTS

Amici Curiae adopt the statement of facts in Dr. Harkonen's Petition, and in his Briefs below, and further note the following significant facts. Professor Fleming, chair of the Data Safety Monitoring Board, testified at trial that the clinical trial at issue was "well-conducted" and that he had confidence in its results. ER0541. Dr. Harkonen's press release, which led to his conviction, presented data from primary and secondary endpoints, and results from an earlier trial. The p-value for a secondary survival endpoint was 0.084. In a per-protocol analysis (which compares actual rather than assigned treatment), the survival benefit for patients who met inclusionary/eligibility criteria was 48% greater for those randomized to Actimmune compared with those randomized to placebo, p=0.055. ER2070.[2] The statement for which Dr. Harkonen has been convicted derived from an unplanned subgroup analysis, which noted that Actimmune "demonstrate(d) a significant survival benefit in patients with mild-to-moderate disease…(p = 0.004)." This presentation of results in the press release is typical of similar presentations in

[2] "ER" Refers to Petitioner's Excerpts of Record filed in the Ninth Circuit.

academic and journal press releases, published scientific literature, and presentations at scientific meetings.

The government did not call any independent expert witnesses to testify about the validity or general acceptance of statistical principles. Rather, the government presented its statistical theory through percipient witnesses without the benefit of expert witness jury instructions, which would have told the jury that it was free to disregard the opinions it had heard.

ARGUMENT

In *Daubert v. Merrell Dow Pharmaceuticals, Inc.,* 509 U.S. 579 (1993),[3] this Court held that general acceptance, as first described in *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923), was not the sole criterion of admissibility of expert witness opinion testimony in federal courts. This Court rejected general acceptance both as a matter of statutory interpretation of Federal Rule of Evidence 702, and because the general acceptance criterion was too crude and inaccurate to guide decisions on the admissibility of expert witness opinions.

This case presents an even more important issue than presented in *Daubert*, namely, whether purportedly generally accepted strategies for making

[3] Professor Rothman, one of the *amici* authors of this brief, was also an author of an *amicus* brief submitted to this Court in *Daubert*.

causal inferences can be used to brand supposedly non-compliant speech as fraudulent, thus imposing a standard of statistical orthodoxy for causal inference upon all scientific discourse. The *Frye* general acceptance rule was poorly conceived as a rule of evidence for scientific inference. Supposed general acceptance is even less well suited for imposing criminal sanctions on scientific speech, especially in this case, in which the government called only percipient witnesses, who did not establish that their personal opinions reflected reliable general consensus.

I. THE CONVICTION, WHICH RESTS ON THE CLAIM THAT THE PRESS RELEASE IS OBJECTIVELY FALSE, IS BASED UPON INCORRECT ASSUMPTIONS AND MIS-UNDERSTANDINGS OF STATISTICAL AND SCIENTIFIC CONCEPTS AND PRACTICE, AND WILL CHILL IMPORTANT SCIENTIFIC SPEECH.

The Court should grant review because the conviction, if allowed to stand, will compel teachers and scientists to conform their expression of how they interpret data from clinical studies to views with which they disagree, and which many scientists reasonably believe are fundamentally flawed.

A. The Government's Rigid Dichotomy between Successful and Failed Clinical Trials Has No Basis in Scientific or Statistical Practice

Fleming and others offered a rigid, false dichotomy that clinical trials either achieve statistical significance on their pre-specified endpoints, or they have "failed." This view is not generally accepted in the scientific community, which focuses on research as an exercise in measurement and seeks to learn as much as possible from the results of every study. Current and contemporaneous guidelines for publication of clinical trials encourage labeling and presentation of the result pertaining to the pre-specified primary outcome,[4] as found in InterMune's press release. These guidelines further allow for presentation of secondary and subgroup analyses, even those not pre-planned. Guidelines recommend that the latter should be so labeled, but this recommendation derives from research showing that publications of trial results often do not separate pre-planned analyses from those suggested by the data. While such unlabeled presentations of results suggested by the data may not follow orthodox reporting recipes (*see* Section I(D)(3), *infra*), unplanned analyses in subgroups were and are clearly part of the scientific literature presenting clinical trial results, whether in journal articles, grant

---

[4] *See, e.g.,* David Moher, et al., *The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials*, 134 Annals Internal Med. 657 (2001); Kenneth Schulz, et al., *CONSORT 2010 Statement: Updated Guidelines*, 152 Annals Internal Med. 726 (2010).

applications, press releases, or other presentations. Were they not, then guidelines would not be necessary.

The p-values associated with specific findings are only one basis for interpreting experimental results. Fleming's view, based upon dichotomizing p-values into only two categories, ignores the continuity of p-values within the range 0 to 1, and ignores the widely held rejection of classifying complex biomedical studies into binary categories of success or failure by comparing their p-values with a standard of 0.05. Fleming's interpretation would lead one to conclude that a treatment trial fails if $p = 0.050001$, but succeeds if $p = 0.049999$. Rather than sort research results into arbitrarily defined categories of treatment successes and treatment failures based upon p-values, the more productive and sophisticated view would instead focus on the estimated magnitude of treatment effect observed in the trial.

Motivated in part by the potential for the inferential errors induced by dogmatic statistical significance testing, one leading epidemiology journal implemented a near-universal ban of statistical significance tests within its pages. Janet Lang, et al., *That Confounded P-Value*, 9 Epidemiology 7 (1998). Furthermore, the International Committee of Medical Journal Editors, a consortium of editors from top-tier biomedical journals, state in their guidelines that, in reporting study results, authors should "[a]void relying solely on statistical hypothesis testing, such as P-values, which fail to convey important

information about effect size."[5] Fleming's views are especially erroneous because the preplanned secondary survival endpoint achieved clinically significant results of 40% mortality reduction, with p = 0.084. As noted by Professor Goodman in his trial court declarations, many scientists would have found the difference between p = 0.05 and 0.084 irrelevant, and would have focused instead on the apparent beneficial treatment effect for this group of patients on Actimmune. ER2560-61. The small difference between the arbitrary acceptable Type 1 error rate of 0.05, and the p-value of 0.084, shrank even further on the per-protocol survival analysis, p = 0.055. The p-value for this important survival endpoint, in a biologically plausible subgroup of those with less severe disease, was 0.004. It is this latter result, in conjunction with the word "demonstrate," that has been the basis for the fraud conviction.

In presenting the clinical trial results and characterizing what conclusions might be drawn, Dr. Harkonen was well within mainstream scientific practice to consider not only the post-hoc subgroup analysis of the current study, but other pertinent information, including prior studies and the design of the Phase III trial itself. Such information included the results of the previous published Austrian clinical trial, *see* Rolf Ziesche, et al., *A Preliminary Study of Long-Term Treatment with Interferon Gamma-1b and Low-Dose Prednisolone in Patients with*

---

[5] Int'l Committee of Med. J. Eds., Uniform Requirements for Manuscripts Submitted to Biomedical Journals (April 2010), http://www.icmje.org/
manuscript_1prepare.html, last visited July 24, 2013.

Idiopathic Pulmonary Fibrosis, 341 New Eng. J. Med. 1264 (1999), the long-term follow up of the patients in that initial trial, which supported therapeutic benefits (with very low p-values), the exclusion of patients with severe IPF from the Phase III trial, as well as clinical experience in treating IPF with Actimmune, and the large body of research on mechanisms by which Actimmune inhibits lung fibrosis. ER2001-02; 2820. Researchers reasonably examine the totality of evidence in drawing conclusions about drug efficacy or harm.

B. The Government's Theory of Fraud Requires an Untenable Interpretation of "Demonstrate" in the Press Release

The government's principal basis for claiming that Dr. Harkonen engaged in fraudulent speech was his use of the word "demonstrate" to interpret subgroup results of Intermune's Phase III clinical trial. ER1906, 1907. There is no technical, generally accepted use of this verb in scientific discourse. The government's contention that the use of "demonstrate" was false and misleading ignores the variability of acceptable scientific usage.

The government has argued that, by using this word, Dr. Harkonen claimed to have proven efficacy conclusively. This position is linguistically and practically untenable. Of the several distinct meanings of "demonstrate," the one clearly not at issue here is that which describes mathematical or geometrical proofs: "to prove beyond the possibility of doubt." *The Oxford English Dictionary* (2013) (entry

4, for "demonstrate"), available at http://www.oed.com/. This meaning, one among several, is clearly not the typical meaning used by biomedical researchers, because in the empirical sciences, as opposed to mathematics and logic, proof beyond the possibility of doubt is impossible. "To demonstrate" has other accepted meanings, including the more modest, less mathematical, meaning of "to show." *Id.*

The mathematical, certain sense of "demonstrate" could not reasonably have been attributed to Dr. Harkonen. Empirical scientists understand that proof of theories, in the mathematical sense, is not possible. No reasonable person would have thus understood the Press Release to be claiming absolute, conclusive, or definitive proof of efficacy. "[T]here are no certainties in science." *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 590 (1993). Biomedical research uses statistical methodologies in part because there is uncertainty and variability in patient samples and their responses to treatment. Unlike mathematical papers, biomedical papers do not conclude with "*Quod erat demonstrandum*."

Even large, well-conducted clinical trials do not satisfy the government's imputed meaning of "demonstrate," which supposedly connotes conclusive proof of causality. Reasonable scientists understand that such clinical trials do not provide conclusive proof. The government's position profoundly misunderstands the limited role of statistical inference based on p-values compared with an

acceptable Type 1 error rate (typically 5%). In the paradigm of statistical significance testing, the acceptable Type 1 error rate is defined as the acceptable probability that studies will produce statistically significant differences favoring the new treatment when the treatment, in reality, produces no benefit. On average, over the long-term of conducting many repeated studies, assuming that there is no treatment effect, and no bias in data collection or analysis, then one expects that 5% of studies will produce p-values below 0.05. For any particular result, such as the result at issue, one can never know whether that result is an example of Type 1 error or a true finding of treatment benefit. Thus, researchers know that "demonstrate," when used in conjunction with a p-value compared with an acceptable Type 1 error rate, does not connote conclusive proof of causality. Sander Greenland, *Randomization, Statistics, and Causal Inference*, 1 Epidemiology 421 (1990). Indeed, clinical trials of the same size, and of the same treatment, can be discordant with respect to statistical significance, sometimes producing a p-value <0.05 and sometimes not, thus showing that p-values cannot guarantee reliability. Jacques Lelorier, et al., *Discrepancies Between Meta-Analyses and Subsequent Large Randomized, Controlled Trials*, 337 New Eng. J. Med. 536 (1997). As teachers and scientists, who write about science, we are at a loss for how to teach students, or to communicate clearly with peers, when we know that the government can prosecute scientists by imputing definitions of a key word in our vocabulary, and when few, if any, scientists understand the word to have the governmentally approved meaning.

C. Scientists Frequently Claim to Have Demonstrated Causation From Data and Statistical Analyses Similar, or Inferior, to Those Reported in the Press Release

In the natural sciences, "to demonstrate" does not imply mathematical certainty, or even the level of certainty that some purists wish to retain for the strongest inferences from the most rigorous studies. Scientists often claim, in publications, to have demonstrated health effects, both safety and efficacy, using post-hoc analyses of subgroups from clinical trials. *See, e.g.*, Jerome Goldstein, et al., *Treatment of Severe, Disabling Migraine Attacks in an Over-the-Counter Population of Migraine Sufferers*, 19 Cephalgia 684, 689 (1999) (reporting post-hoc sub-group analysis based upon multiple clinical trials; "results of this post-hoc analysis *demonstrate* that AAC is effective. . . .") (emphasis added)). Placing such authors at risk for prosecution would have a chilling effect on scientific discourse. Even if scientists could quickly learn to avoid the term "demonstrate," they would then have to worry about what combination of words and statistics conveyed an equivalent viewpoint, only to find themselves targeted for governmental prosecution.

Scientists use "demonstrate" frequently in published papers to describe study results that have much less scientific probative value than the clinical trials at issue. Frequently, these papers report observational studies with claims to have demonstrated health effects, but for which there is no consensus that the proclaimed effects are real.

Observational studies, lacking randomized assignment to exposure and prospective ascertainment of outcomes, have greater potential for bias and confounding than most clinical trials. Federal Judicial Center, *Reference Manual on Scientific Evidence* 555 (3d ed. 2011) (describing randomized clinical trials as the gold standard) [hereinafter "*Reference Manual*"].

For example, in a published observational study, government scientists published post-hoc subgroup results based upon a non-standard age stratification, in looking for occupational risk factors for Parkinson's disease. Robert Park, et al., *Potential Occupational Risks for Neurodegenerative Diseases*, 48 Am. J. Indus. Med. 63 (2005). For welders, these authors found a statistically significant *decreased* risk overall, but the authors also published their post-hoc subgroup analysis that compared welder death under and over age 65, and concluded that their post-hoc subgroup supported a claim of increased risk among welders. *Id.* at 73. These authors conducted *hundreds* of tests and subgroup analyses without discussing multiplicity or modifying their reported p-values, or qualifying their conclusions. This approach is not unusual, and indeed is encouraged because of the expense of collecting valuable data and the large opportunity cost resulting from a failure to analyze such valuable data thoroughly.

The government's prosecution criminalizes legitimate diction choices. Scientists reasonably take different views about when they have sufficient data to claim "demonstration" as opposed to a "suggestion"

of benefit. To the extent that the linguistic distinction between "demonstration" and "suggestion" has any meaning, scientists must be free to assert their views of how to characterize their inferences. The way to resolve disagreements over causal inferences is with scientific debate and further research, not with criminal prosecutions.

In pretrial briefs, the government argued that "suggested" would have been a better word choice for the Press Release. ER2497. In opposing post-trial motions, the government adopted a statement from one of its witnesses that the post-hoc subgroup could "show," but not "prove" a survival benefit. U.S. Opp'n to Def's Post-Trial Mots. at 7 (Doc. 256) (quoting Crager: "There was, however, a trend in the survival data that appeared to show a benefit."). A Columbia University guide to "Writing a Scientific Research Article" recommends that writers use "show" as a "shorter word" to replace "demonstrate" without a change in meaning. Columbia Univ. Biology Dep't, *Writing a Scientific Research Article*, http://www.columbia.edu/cu/biology/ug/research/paper.html, last visited July 24, 2013.

Scientific practice cannot be corralled by semantic legerdemain. There is no consistent, meaningful difference between "demonstrate" and "show" in scientific practice. The government's resort to prosecutions to legislate scientists' speech, and their characterization of inferences and conclusions, chills scientific discourse. Scientists, whether in

industry or government, need the freedom to characterize the conclusions they draw from data.

D. The Government Misrepresents Key Statistical Concepts

1. Misrepresentations of the Meaning of P-Values

A *p*-value is the probability of observing data "as extreme as, or more extreme than, the actual data—given that the null hypothesis is true." *Reference Manual* at 250. This probability is also sometimes called "the attained level of significance" because of its role in statistical significance testing.

"Significance" is a technical statistical term; it does not have the ordinary connotations of "important" or meaningful. Statistically significant results may be clinically unimportant, because the treatment effect is small, and results with statistically non-significant p-values can arise from studies of treatments with large, clinically important effects. *Id.* at 252.

The government's principal brief in the Ninth Circuit misstated the concept that is at the heart of this prosecution:

> Generally, the significance of primary endpoint results is primarily expressed through the p-value, which is a number between 1 and 0. ER 43; SER 437. The lower the p-value, the greater the

> probability that the result reflected by the data is meaningful, and not due to chance. ER 43; SER 437-39. For example, a p-value of 0.05 indicates that the data obtained in the trial would occur by chance less than 5% of the time. ER 43; SER 438-39.

Appellee/Cross-Appellant's Brief at 12. The p-value, however, is not the probability that the observed result has occurred as a result of chance. As noted above, the p-value is calculated assuming that the null hypothesis is true, that is, assuming that chance accounts for the results. Logically, one cannot assume something to be true as part of a calculation, and then use that calculation to measure whether the incorporated assumption is true. For this reason, and because the p-value also depends on study size, it does not measure the probability that the data are meaningful. Results with small p-values measured in large samples can pertain to inconsequential differences between treatment groups, and large and important differences between treatment groups can have moderately sized p-values when, for example, the sample size is small.

The government's errors in defining and interpreting p-values are unfortunately common, and have, in fact, been widely decried by public health scientists.[6] Similar misstatements occur elsewhere in the government's appellate brief. *See id.* at 17 n.11.

---

[6] *E.g.,* Timothy Lash, et al., *Re Promoting Healthy Skepticism in the News: Helping Journalists Get It Right*, 102 J. Nat'l Cancer Inst. 829 (2010); *see also* Editorial Signatories, Petition

The irony that a government, intent upon prosecuting a scientist for drawing an allegedly improper statistical inference, would erroneously define core conceptual issues is deeply disturbing. The government's brief invites an error so common that it has a name: the "transposition fallacy." *Reference Manual* at 250-51 & n.99. The p-value does not give the probability that the null hypothesis is correct (*i.e.,* that the data would occur by chance), because it cannot measure the "correctness" of a hypothesis assumed to be true in the course of its calculation; nor does the p-value provide a measure of the probability that an observed result is correct or meaningful.

The government's errors raise serious concerns about criminalizing allegedly fraudulent statistical statements or inferences. If the government cannot correctly define basic statistical concepts, then it has no business prosecuting others on allegations of committing similar offenses.

2. Misrepresentations of the Importance of P-Values

The government's rigid dichotomization of clinical trials as successes or failures derives from an equally rigid application of statistical hypothesis testing. Hypothesis testing as sometimes practiced is a binary decision process in which the null hypothesis (usually of no association between treatment and

---

Supporting Peer Review of Materials for Journalists (over 40 scientists joining letter), https://sites.google.com/site/editorialsignatories/, last visited August 30, 2013.

outcome) is rejected only if the study results yield a p-value below 0.05 (or some other pre-specified allowable Type 1 error rate). When the p-value is above this pre-specified level, statistical hypothesis testing dictates that the null hypothesis is not rejected. That does not mean, however, that the study results establish that the treatment provides no benefit. There is no equivalence between failing to reject and accepting the null hypothesis as true; this important distinction has been reiterated frequently in the statistical inference literature.

Statistical hypothesis testing as a rigid decision procedure, based upon p-values less than 0.05, is commonly practiced but widely derided. Indeed, the idea of statistical testing as driven by a rigid, pre-selected level of acceptable Type 1 error rate was rejected by the very statistician who developed computations of the p-value. *See* Sir Ronald Fisher, *Statistical Methods and Scientific Inference* 42 (Hafner 1956) (ridiculing rigid hypothesis testing as "absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.").

The district court's acceptance of p-values as "magic numbers" that determine when data are "reliable," with the Ninth Circuit's affirmance, takes a highly nuanced, essentially contested approach to statistical inference, and turns it into a rule of law. The consequence will be to allow the government to ensnare scientists and regulators who vehemently

object to this method of statistical inference in a linguistic and potentially literal prison, curtailing healthy debate and free interchange at the intersection of science, law, and policy.

3. Multiple Testing Does Not Undermine the Meaning of P-Values

As discussed above, scientists often describe post-hoc subgroup findings as demonstrated effects. Although some scientists would disagree with this reporting, the practice is common, and the very idea that pre-specified hypotheses are inherently more reliable than post-hoc hypotheses is the subject of ongoing debate. *See* Timothy Lash & Jan Vandenbroucke, *Should Preregistration of Epidemiologic Study Protocols Become Compulsory?*, 23 Epidemiology 184 (2012). One survey that compared grant applications with subsequently published papers reported that subgroup analyses were pre-specified in only a minority of cases; in a substantial majority (77%), the subgroup analyses in published papers were not characterized as either pre-specified or post-hoc. Chantal Boonacker, et al., *A Comparison of Subgroup Analyses in Grant Applications and Publications*, 174 Am. J. Epidemiology 291, 291 (2011). By comparing grant applications with the published papers, Boonacker was able to identify most subgroup analyses as post-hoc. Furthermore, this survey found that authors of published papers rarely reported justifications for subgroup analyses or other analyses that would be classified as post-hoc. *Id.*

The practice of presenting unplanned subgroup analyses, whether optimal or not, is quite common in the scientific literature. A survey of publication practice in the *New England Journal of Medicine* reported similar findings. Rui Wang, et al., *Statistics in Medicine—Reporting of Subgroup Analyses in Clinical Trials*, 357 New Eng. J. Med. 2189 (2007). In general, these authors were unable to determine the total number of subgroup analyses performed; and in the majority (68%) of trials discussed, Wang could not determine whether the subgroup analyses were pre-specified. *Id.* at 2912. Although Wang proposed guidelines for identifying subgroup analyses as pre-specified or post-hoc, she emphasized that the proposals were not "rules" that could be rigidly prescribed. *Id.* at 2194.

4. There Is No Consensus on Whether, When, or How Adjustments Should Be Made to P-Values from Subgroups

The government's position in this case is based upon testimony that represents an extreme, rigid view within statistics: multiple post-hoc testing renders subgroup findings "meaningless." This extreme view prohibits learning from data that does not meet a pre-specified p-value, or that came from an unanticipated concentration of harm or benefit in a subgroup. If this approach were widely adhered to, many serendipitous discoveries might never have surfaced.

At trial, Fleming described the Bonferroni correction for multiple testing, which involves lowering the pre-specified Type I error rate by dividing it by the number of independent tests. ER0551. The Bonferroni correction is only one of many approaches to address multiplicity in clinical trials and can be unduly restrictive. In a 2008 textbook co-authored by two of the *amici*, it is labeled as an "awful response" to the multiplicity problem. Kenneth Rothman, Sander Greenland, Timothy Lash, *Modern Epidemiology* 236 (3d ed. 2008). The study's Statistical Analysis Plan did not provide for any multiple testing adjustments for secondary endpoints or subgroups. ER2281-94. This absence is typical, and is presumably why the government conceded that the data, including the p-values for the subgroup survival endpoint, are accurate. Although Fleming and others testified that the p-values as presented were "meaningless," no one calculated whether the subgroup's p-value of 0.004 would have remained statistically significant with adjustments for multiple testing. Fleming's testimony was thus inappropriately dismissive of Dr. Harkonen's view that the trial results for the survival subgroup supported causal language.

There is no consensus whether, when, or how to adjust p-values or Type I error rates for multiple testing. *See, e.g.,* Kenneth Rothman, *No Adjustments Are Needed for Multiple Comparisons*, 1 Epidemiology 43, 43 (1990) ("policy of not making adjustment for multiple comparisons is preferable"); Steven Goodman, *Multiple Comparisons, Explained*, 147 Am. J. Epidemiology 807 (1998). Although in

some circumstances adjustments for multiple comparisons may be appropriate, the issue is not settled among scientists, and government control of the topic under the guise of fraud prosecutions would petrify the scientific debate well before a consensus has been reached.

5. Press Releases Are Understood in the Scientific Community to Advance Conclusions Less Rigorously Than Published Articles

Not only did the government's case ignore the complexities of causal and statistical inference, the prosecution's insistence upon a statistical orthodoxy as the touchstone for truth and falsity ignored the social context of Dr. Harkonen's press release. The Press Release did not purport to be a thorough presentation of the data, and it alerted the reader to forthcoming presentations on investor conference calls, and at upcoming scientific conferences.

It is widely acknowledged that all scientists, whether seeking publicity or funding, tend to report findings of their studies less technically and less rigorously in press releases than in submitting manuscripts to scientific journals. *Developments in the Law: Confronting the New Challenges of Scientific Evidence* 108 Harv. L. Rev. 1481, 1553 & n.135 (1995) (citing Lawrence Altman, M.D., *The Doctor's World; Promises of Miracles: News Releases Go Where Journals Fear to Tread*, N.Y. Times, Jan. 10, 1995, at C3.)

Empirical surveys show that academic medical centers often oversimplify and exaggerate findings in press releases. In one recent study of 200 press releases randomly selected, a substantial number omitted important quantitative information. In the press releases on human research, 23% failed to report study size, and 34% did not quantify the findings. Steven Woloshin, et al., *Press Releases by Academic Medical Centers: Not So Academic?*, 150 Annals Internal Med. 613 (2009). These investigators (three of whom were with the Department of Veteran Affairs) similarly found that a high percentage (29% of all releases) exaggerated the importance of the research. *Id.* at 615.

Lack of rigor in press releases is not limited to academic and industry press releases. Consider the press release recently issued by the National Institutes of Health (NIH) in connection with a NIH-funded clinical trial on age-related macular degeneration (AMD). NIH Press Release, NIH Study Provides Clarity on Supplements for Protection against Blinding Eye Disease (May 5, 2013), available at: http://www.nih.gov/news/ health/may2013/nei-05.htm, last visited July 23, 2013.

The clinical trial studied a modified dietary supplement in common use to prevent or delay AMD. The NIH's press release states that the study "provides clarity on supplements," and announced a "finding" of "some benefits" when looking at just two of the subgroups. The press release does not use the words "post hoc" or "ad hoc" in connection with the

subgroup analysis used to support the "finding" of benefit.

The clinical trial results were published the same day in a journal article that labeled the subgroup findings as post hoc subgroup findings. Emily Chew, et al., *Lutein + Zeaxanthin and Omega-3 Fatty Acids for Age-Related Macular Degeneration*, 309 J. Am. Med. Ass'n 2005 (2013). The published paper also reported that the pre-specified endpoints of the clinical trial did not show statistically significant differences between therapies and placebo. None of the p-values for any of the post-hoc subgroup analysis was adjusted for multiple comparisons. NIH webpages with Questions and Answers for the public and the media both fail to report the post-hoc nature of the subgroup findings. *See* For the Public: What the Age-Related Eye Disease Studies Mean for You (May 2013), http://www.nei.nih.gov/areds2/PatientFAQ.asp, last visited July 23, 2013; For the Media: Questions and Answers about AREDS2 (May 2013), http://www.nei.nih.gov/areds2/MediaQandA.asp, last visited July 23, 2013. By the standards imposed upon Dr. Harkonen in this case through Dr. Fleming's testimony, and contrary to the NIH's public representations, the NIH trial had "failed," and no inferences could be drawn with respect to any endpoint because the primary endpoint did not yield a statistically significant result.

II. FDA REGULATIONS AND PRACTICE DO NOT SUPPORT THE CLAIM THAT DR. HARKONEN MISREPRESENTED THE EFFICACY OF ACTIMMUNE

The petition correctly states that the FDA or its advisory committees have acted in ways inconsistent with Fleming's self-proclaimed statistical orthodoxy. See Pet. 33 & n.7; Pet. App. 99a-100a. The Press Release at issue here never represented that the demonstrated benefit was assessed under FDA regulations, which may or may not reflect how scientists communicate outside its regulatory purview.

No rule or regulation of the FDA requires statistical significance to mean $p < 0.05$; indeed FDA regulations do not prescribe any particular statistical analysis. Although the FDA has stated that it is "unlikely" to accept conclusions based upon exploratory subgroup analysis, the agency is not prohibited from doing so, and its guidelines expressly "do not bind the public." U.S. Food & Drug Admin., *Guidance for Industry: E9 Statistical Principles for Clinical Trials* 1, 34 (1998).

III. THE GOVERNMENT'S ARGUMENTS AND OPINIONS ADVANCED IN *MATRIXX* ARE INCONSISTENT WITH ITS THEORY OF THIS PROSECUTION

In *Matrixx Initiatives Inc. v. Siracusano*, 131 S.Ct. 1309 (2011), a securities fraud class action, the defendant moved to dismiss the complaint because

plaintiffs had failed to plead "statistically significant" evidence showing that defendant's drug caused adverse effects. Plaintiffs alleged that the failure to disclose evidence of harm was fraudulent, given the company's bullish sales projections. The district court dismissed the complaint; the Ninth Circuit reversed; and this Court affirmed. *Id.* In affirming, the Court drew heavily from the Solicitor General's brief, noting that "the premise that statistical significance is the only reliable indication of causation … is flawed." *Id.*; Brief for the United States as *Amicus Curiae* Supporting Respondents, 2010 WL 4624148. The government's position in *Matrixx* was clear: "*Statistical significance is a limited and non-exclusive tool for inferring causation*." *Id.* at *13. The brief, in broad generalities, disclaimed the necessity and importance of statistical significance: "data showing a statistically significant association are not essential to establish a link between use of a drug and an adverse effect." *Id.* at *12. The government declared that the lack of statistical significance "does not refute an inference of causation." *Id.* at *14. The government's *Matrixx* brief argued against statistical significance for causality of *both* safety and efficacy outcomes: "[t]he same principle applies to studies suggesting that a particular drug is efficacious." *Id.* at *15 n.2.

The district court below denied motions based upon the *Matrixx* brief. The court stated that *Matrixx* involved safety rather than efficacy outcomes, that *Matrixx* involved civil securities fraud, and that the *Matrixx* brief was not newly discovered evidence. ER0019-0035. None of these bases is persuasive; and

none can hide the inconsistency between the government's prosecution of Dr. Harkonen and its *Matrixx* argument. There is no basis for distinguishing the degree of evidence needed for causal claims of efficacy or harm, and the government disavowed the distinction. The differences between civil and criminal fraud argues in favor of applying the government's arguments from *Matrixx* even more vigorously in this criminal case. The district court's point that the *amicus* brief was argument, not evidence, suggests that contentions about whether causality may be inferred are opinions, and not factual evidence that could be false or misrepresented. Indeed, interpreting statistical tests and drawing statistical inferences are considered arguments by statisticians. *See, e.g.,* David Cox & Nancy Reid, *The Theory of the Design of Experiments passim* (2000).

The *Matrixx* brief argued for drawing causal inferences in a manner that flatly contradicted the prosecution's arguments in this case. The government should not be permitted to argue the non-necessity of statistical significance when it advocates for popular plaintiffs, but argue the contrary when prosecuting unpopular scientists.

CONCLUSION

In the Press Release presentation of results from this clinical trial, there was no evidence that anyone had fabricated or falsified data, or that anyone had manipulated statistical analyses to report a result that was objectively false. The government

based its allegations of fraud upon its disagreement with how Dr. Harkonen interpreted data in a Press Release. Although some scientists might disagree with the language of the Press Release, others might defend it. The language is consistent with scientific practice, is not fraudulent, and should be viewed in the social context of Press Releases. It was, in fact, in form and content, consistent with scientific discourse one can find in many clinical journals and many Press Releases regarding the results of human-subject research. The potential for criminal fraud prosecution arising from language commonly found in scientific discourse would have a chilling effect on scientists conducting and reporting research, submitting grant proposals, communicating about regulations, and providing expert witness testimony.

Amici Curiae respectfully urge this Court to grant certiorari, and reverse the judgment of conviction, in this case.

Dated: September 4, 2013

BECKY WALKER JAMES
Counsel of Record
LAW OFFICES OF
BECKY WALKER JAMES
17383 SUNSET BOULEVARD,
SUITE A315
PACIFIC PALISADES, CA 90272
(310) 492-5104
becky@walkerjameslaw.com
Counsel for Amici Curiae

APPENDIX A

Amici participate here in their personal capacities, and provide their titles and university affiliations only for identification.

Timothy Lash D.Sc., M.P.H. holds appointments as Professor of Epidemiology in the Rollins School of Public Health at Emory University and Honorary Professor of Cancer Epidemiology at Aarhus University in Denmark. His research focuses on predictors of cancer recurrence, including biomarkers that predict the effectiveness of cancer drugs and the effect of drugs directed at other medical conditions but which also affect cancer outcomes. He is coauthor of a leading textbook of epidemiologic methods, *Modern Epidemiology* (3rd ed. 2008), which contains several chapters on statistical and causal inference. He is also coauthor of a textbook, *Applying Quantitative Bias Analysis to Epidemiologic Data*, which describes methods for quantifying uncertainties in human subjects research. He has published several commentaries regarding the proper definition and interpretation of p-values, and the proper inference from unplanned secondary analyses. These topics are central to the topics at issue in this case, which is the basis for his interest in submitting this brief.

Kenneth J. Rothman, Dr.P.H., is a Distinguished Fellow at the Research Triangle Institute, and a Professor of Epidemiology at Boston University. His research interests in epidemiology have spanned a wide range of health problems, including cancer,

cardiovascular disease, neurologic disease, birth defects, injuries, and adverse effects of pharmaceutical treatments, and he has authored or co-authored hundreds of peer-reviewed scientific papers. His main career focus, however, has been teaching and contributing to the development of the concepts and methods of epidemiologic research. He has written two epidemiologic textbooks: *Modern Epidemiology*, first published in 1986 and now in its third edition, is a comprehensive and widely used advanced text of epidemiologic methods; and *Epidemiology – An Introduction* is a popular introductory text published by Oxford University Press, now in its second edition. His long-standing interest in how scientific and statistical evidence are received and interpreted within the judicial system led to his contributing as a reviewer to the Federal Judicial Center's Reference Manual on Epidemiology, a section of the Reference Manual for Scientific Evidence, which cites his work in numerous places, and to authoring an *amicus* brief, 1992 WL 12006438, Brief *Amici Curiae* of Professors Kenneth Rothman, *et al.* in *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993). In addition to the above activities, he is the founding editor of the journal *Epidemiology*, and has served in many editorial posts, including the Editorial Board of the *New England Journal of Medicine* and the International Advisory Board of *The Lancet*.

Nathan A. Schachtman, J.D., is a lawyer in private practice, and a Lecturer in Law, in the Columbia Law School, where he teaches a course on probability and statistics in the law. He is an elected member of the

American Law Institute and a fellow of the American Bar Foundation. For close to thirty years, he has tried cases and argued appeals involving statistical and scientific evidence, including some of the leading cases involving the epidemiology of silicone-gel breast implants and other medical devices, over-the-counter and prescription medications, and various occupational exposures. Mr. Schachtman has lectured and published widely on scientific and statistical evidence, medico-legal causation, and expert witness issues.