WHEN THE MATH MATTERS: USE OF P-VALUES IN PHARMACEUTICAL LITIGATION

Robin L. Juni[*]

ABSTRACT

This Article explains the use of p-values as part of statistical analyses to support several types of pharmaceutical litigation, focusing on how lawyers can best present relevant studies to benefit their clients "when the math matters" and indeed critically affects the outcome of the dispute. The Article begins with a conceptual discussion of p-values as utilized in the scientific community and continues by describing the substantial concerns many scientists have voiced about the use of p-values to show much of anything. The Article then explains use by lawyers and judges of p-values in litigation, particularly three aspects of pharmaceutical litigation, where concerns about p-values sometimes are identified, but the discussion is typically relatively superficial, and the broader concerns of the scientific community about p-values are neither generally understood nor discussed. Next, the Article dives deeper into three exemplar cases, showing how p-values were used and perhaps misused in those contexts. The Article concludes by discussing principles of scientific communication that may help lawyers break down the discussion and better explain "the math" in similar cases.

[*] Associate Professor, Fundamentals of Lawyering Program, at the George Washington University Law School. J.D. Harvard Law School; B.A. Hamline University. The author thanks participants at the Health Law Works in Progress panel at the Southeastern Association of Law Schools (SEALS) Conference in August 2022 for valuable ideas and Max Keefe, GW Law Class of 2023, for excellent research assistance.

INTRODUCTION

This Article focuses on the significant role p-values play in statistical analyses used in several types of pharmaceutical litigation, and how lawyers can best understand and use p-values as a tool in litigation. This Article contrasts available litigation approaches,[1] discusses the relevant mathematical methodologies,[2] and shows lawyers how to explain "the math" in similar cases.[3] This Article builds on earlier work from the author and others explaining statistical methodology in other kinds of litigation, "when

1. *See infra* Part II.
2. *See infra* Part III.
3. *See infra* Part IV.

the math matters" and indeed is crucial to understanding and advocacy.[4] Such efforts to educate may usefully start well before law school.[5]

I. USE OF P-VALUES IN PHARMACEUTICAL LITIGATION

A. Definition of P-Values

The calculation of the p-value is part of any effort to engage in statistical significance testing, a method for calculating the significance of a study and whether a data set supports a particular alternative hypothesis against the null

4. *See* Robin L. Juni, *When the Math Matters: Improving Statistical Advocacy in Gerrymandering Litigation*, 100 NEB. L. REV. 727 (2022) (discussing U.S. Supreme Court rejection of multivariable regression analysis that showed gerrymandering as "sociological gobbledygook" and explaining long judicial acceptance of similar methodologies). The *When the Math Matters* series is envisioned as an ongoing conversation about the use of mathematical concepts in legal decision-making. Each article focuses on a different concept and seeks to educate lawyers on how better to utilize these concepts to represent their clients. More specifically, the gerrymandering article discusses a dispute involving an important mathematical idea—the core statistical concept of regression analysis, particularly multivariable regression analysis—then explaining the underlying math involved and connecting it to the legal issues. *Id.* at 728. That article focuses on Gill v. Whitford, 138 S. Ct. 1916 (2018), in which Chief Justice Roberts referred to the multivariable regression analysis that demonstrated gerrymandering as "sociological gobbledygook." *Id.* at 729–30 n.9 (quoting Transcript of Oral Argument at 37–40, Gill v. Whitford, 138 S. Ct. 1916 (2018) (No. 16-1161)). That article explains the statistics behind the multivariable regression analysis, connects it to the issues before the Court, and shows that these concepts are employed in many types of litigation. With the opportunities for redistricting based on new census results, the article concludes, mathematical understanding in the context of gerrymandering litigation is more critical than ever. *See id.* at 761.

5. *See* SHEILA TOBIAS, OVERCOMING MATH ANXIETY 33 (rev. ed. 1993); Caitlin McDermott-Murphy, *The Myth of the 'Math Person,'* HARV. GAZETTE (Nov. 9, 2022), https://news.harvard.edu/gazette/story/2022/11/the-myth-of-the-math-person/. In her article, McDermott-Murphy discusses *Overcoming Math Anxiety*'s text and the responsibilities of mathematics teachers to better reach a broader group of students:

> "There's a genius myth in mathematics," said Brendan Kelly, director of introductory math at Harvard. "There's often this perception that success requires some natural ability, some unteachable qualities, some immutable traits."
>
> When students learn to write stories or play the violin, most don't expect to replicate Toni Morrison or Niccolò Paganini in their first attempts. No one says, "I'm not a writing person." But in math, said Allechar Serrano López, . . . a preceptor in mathematics at Harvard, "It gets decided when they're literally children if they are going to be math people or if they're not math people." And because math is a gateway to almost every other field of science, that early stamp can squeeze students out of the STEM pipeline.
>
> . . . .
>
> "The responsibility really should be mine to create the space where students feel that they can ask questions, share their ideas, and slowly become more confident and overcome their math anxiety," said [Reshma] Menon [another preceptor in Harvard's mathematics department].

Id.

hypothesis.[6] The p-value is calculated as a fraction, expressed as a decimal, and represents the probability of a study producing particular data if the null hypothesis is true.[7] More specifically, the p-value indicates how often, in many repeated trials, one would expect to see a test statistic at least as extreme as the one observed in the data if the null hypothesis was true.[8] The p-value is *not* the probability that the null hypothesis is true and the alternative hypothesis is false,[9] nor the probability that a result is due to

---

6. *See* David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, *in* COMM. ON SCI., TECH., & L. POL'Y & GLOB. AFF., REFERENCE MANUAL ON SCI. EVIDENCE 211, 241 (3d ed. 2011) [hereinafter REFERENCE MANUAL].

7. *See* Paul A. Murtaugh, *In Defense of* P *Values*, 95 ECOLOGICAL SOC'Y AM. 611, 613 (2014); *see also* Lydia Denworth, *The Significant Problem of* P *Values*, 321 SCI. AM. 63, 64 (2019) ("In the past decade the debate over statistical significance has flared up with unusual intensity. One publication called the flimsy foundation of statistical analysis 'science's dirtiest secret.'").

8. Murtaugh, *supra* note 7, at 612.

9. *Id.*; *see also* Sander Greenland et al., *Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations*, 31 EUR. J. EPIDEMIOLOGY 337, 341 (2016) ("A small *P* value simply flags the data as being unusual if all the assumptions used to compute it . . . were correct."). Greenland et al. further explain:

> The difficulty of understanding and assessing underlying assumptions is exacerbated by the fact that the statistical model is usually presented in a highly compressed and abstract form—if presented at all. As a result, many assumptions go unremarked and are often unrecognized by users as well as consumers of statistics. Nonetheless, all statistical methods and interpretations are premised on the model assumptions; that is, on an assumption that the model provides a valid representation of the variation we would expect to see across data sets, faithfully reflecting the circumstances surrounding the study and phenomena occurring within it.
>
> . . . .
>
> Much statistical teaching and practice has developed a strong (and unhealthy) focus on the idea that the main aim of a study should be to test null hypotheses.
>
> . . . .
>
> A more refined goal of statistical analysis is to provide an evaluation of certainty or uncertainty regarding the size of an effect. It is natural to express such certainty in terms of "probabilities" of hypotheses. In conventional statistical methods, however, "probability" refers not to hypotheses, but to quantities that are hypothetical frequencies of data patterns under an assumed statistical model. These methods are thus called *frequentist* methods, and the hypothetical frequencies they predict are called "frequency probabilities." Despite considerable training to the contrary, many statistically educated scientists revert to the habit of misinterpreting these frequency probabilities as hypothesis probabilities. (Even more confusingly, the term "likelihood of a parameter value" is reserved by statisticians to refer to the probability of the observed data *given* the parameter value; it does not refer to a probability of the parameter taking on the given value.)
>
> Nowhere are these problems more rampant than in applications of a hypothetical frequency called the *P* value, also known as the "observed significance level" for the test hypothesis. Statistical "significance tests" based on this concept have been a central part of statistical analyses for centuries. The focus of traditional definitions of *P* values and statistical significance has been on null hypotheses,

chance. These misinterpretations are common and one of many issues that has led the scientific community itself to debate the utility of the p-value.[10]

The typical threshold value for describing a test result as "statistically significant" is a p-value of 0.05, but at times a p-value of 0.01 is recognized as optimal to support a particular result.[11] These values are admittedly

---

> treating all other assumptions used to compute the *P* value as if they were known to be correct. Recognizing that these other assumptions are often questionable if not unwarranted, we will adopt a more general view of the *P* value as a statistical summary of the compatibility between the observed data and what we would predict or expect to see if we knew the entire statistical model (*all* the assumptions used to compute the *P* value) were correct.

Id. at 338–39.

10. *See infra* Parts II.A, III.

11. Adopted at the beginning of the development of statistical science, p-values of 0.05 and 0.01 are recognized as fundamentally arbitrary as against similarly small numbers. *See* GERARD E. DALLAL, THE LITTLE HANDBOOK OF STATISTICAL PRACTICE 243 (2012).

> There are many theories and stories to account for the use of P=0.05 to denote statistical significance. All of them trace the practice back to the influence of R. A. Fisher. In 1914, Karl Pearson published his *Tables for Statisticians & Biometricians*. For each distribution, Pearson gave the value of P for a series of values of the random variable. When Fisher published *Statistical Methods for Research Workers* (SMRW) in 1925, he included tables that gave the value of the random variable for specially selected values of P. SMRW was a major influence through the 1950s. The same approach was taken for Fisher's *Statistical Tables for Biological, Agricultural, and Medical Research*, published in 1938 with Frank Yates. Even today, Fisher's tables are widely reproduced in standard statistics texts.
>
> Fisher's tables were compact. Where Pearson described a distribution in detail, Fisher summarized it in a single line in one of his tables making them more suitable for inclusion in standard reference works. However, Fisher's tables would change the way the information could be used. While Pearson's tables provide probabilities for a wide range of values of a statistic, Fisher's tables only bracket the probabilities between coarse bounds.
>
> The impact of Fisher's tables was profound. Through the 1960s, it was standard practice in many fields to report summaries with one star attached to indicate $P < 0.05$ and two stars to indicate $P < 0.01$[.] Occasionally, three [stars] were used to indicate $P < 0.001$.
>
> . . . .
>
> It was Fisher who suggested giving 0.05 its special status.

Id. at 259. Fisher described the standard normally distributed as follows:

> The value for which P=0.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.

Id. at 259–60 (quoting RONALD A. FISHER, STATISTICAL METHODS FOR RESEARCH WORKERS 44 (F. A. E. Crew ed., 13th ed. 1958)). R. A. Fisher was a professor of genetics at University College London

arbitrary but are nevertheless the most common in the scientific community and have been used in numerous judicial decisions.[12] Many in the scientific community (incorrectly) see the 0.05 threshold as a firm cutoff, with values just below the cutoff being interpreted differently from those just above that level.[13] This approach has its detractors, who find this application to be arbitrary and nonsensical.[14] This latter sentiment likewise has sometimes found support in the courts.[15]

P-values are often difficult to understand,[16] as evidenced by rampant misinterpretations, even by members of the scientific community,[17] much less by lawyers and judges with generally little experience or training in statistics and lawyers who are seeking to advocate for particular legal results.[18]

The *Reference Manual on Scientific Evidence*, designed for use by judges, contains a putatively simple example, describing a jury selection process in which a jury was drawn from a panel of 350 persons, only 102 of which were women.[19] Assuming the juror population available was

---

and at Cambridge University. *See Ronald Aylmer Fisher (1890–1962)*, UNIV. COLL. LONDON, https://www.ucl.ac.uk/biosciences/gee/ucl-centre-computational-biology/ronald-aylmer-fisher-1890-1962 (last visited Dec. 15, 2023):

> [His] contributions to statistics and to evolution/genetics are so massive and ground-breaking that it is hard for scientists in one field to imagine how he did anything substantial in the other. In statistics, most of what is commonly taught in a standard statistics or biostatistics course is due to Fisher, including significance test[ing], analysis of variance, t distribution, F distribution, design of experiments (randomization, Latin squares), variance, sufficiency, Fisher information, estimation theory, maximum likelihood, and so on.

Id.

12. *See* REFERENCE MANUAL, *supra* note 6, at 252 ("These levels of 5% and 1% have become icons of science and the legal process."). As noted, expression of p-values as a decimal is more useful than expression as a percentage. *See* Greenland et al., *supra* note 9, at 339 (discussing inappropriate characterization of p-values as probabilities).

13. *See* Murtaugh, *supra* note 7, at 612.

14. *See* Douglas H. Johnson, *The Insignificance of Statistical Significance Testing*, 63 J. WILDLIFE MGMT. 763, 765 (1999) ("Use of a fixed [P] level, say [P] = 0.05, promotes the seemingly nonsensical distinction between a significant finding if $P = 0.049$, and a nonsignificant finding if $P = 0.051$.").

15. *See, e.g.*, Allergan, Inc. v. Teva Pharms. USA, Inc., No. 2:15-cv-1455-WCB, 2017 WL 4803941, at *25 (E.D. Tex. Oct. 16, 2017) ("If a person of skill assigns meaning to results when the p-value is 0.0499, the Court sees no reason why that person of skill would suddenly assign no meaning to those same results if the p-value were 0.0501.").

16. *See, e.g.*, Christie Aschwanden, *Not Even Scientists Can Easily Explain P-Values*, FIVETHIRTYEIGHT (Nov. 24, 2015), https://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/.

17. *See* Greenland et al., *supra* note 9, at 341 (containing a particularly good and expansive set of common misinterpretations).

18. *Id.* at 339; *see also infra* Part III.

19. REFERENCE MANUAL, *supra* note 6, at 249.

approximately 50% women (the null hypothesis), the expected number of women on the panel would be 175.[20] Given the null hypothesis, the percent chance of there being ten or fewer women on the panel is the p-value. In this example, the p-value is essentially zero, meaning it would be nearly impossible for the panel to have been a random selection from the population: "In short, the jury panel was nothing like a random sample of the community."[21]

The p-value thus is a measure of how surprised we should be at the 102 result observed.[22] The higher the p-value, the less surprised we should feel, with low p-values—conversely—representing a large degree of surprise. With a low p-value, we should feel so sufficiently surprised at the results we observed that we reject the null hypothesis in favor of the alternative hypothesis.

In the above example, the null hypothesis is that sex did not play a role in the selection of the panel, with the alternative hypothesis being that it did play a role; and thus, the panel likely was chosen with prejudice.[23] With only 102 women on the panel versus the expected 175, the p-value associated with a result at least as extreme as this one can be calculated as 1.81×10^{-15}, or essentially zero, though this calculation is beyond the scope of this Article. In other words, if indeed 50% of the population from which the panel was selected were women, one would expect to see in a fair process as few as 102 women on the panel only once in more than a quadrillion identical such processes. With such an improbable event having occurred in the selection process under review, the null hypothesis is rejected in favor of the alternative hypothesis: the jury panel selection was infected by bias.[24]

---

20. *Id.*

21. *Id.* at 250.

22. *See* Nikhil Karve, *P-value—a Measure of Surprise*, MEDIUM (May 21, 2020), https://medium.com/@nikhilkarve007/p-value-a-measure-of-surprise-28fa96ccd07b ("Let's call p-value as a 'measure of surprise.' Higher the p-value, the less surprised you should be. Lower the p-value, the more surprised you should be, because that means your data is not what you expect *assuming 'null-hypothesis' is true*.") (emphasis in original).

23. REFERENCE MANUAL, *supra* note 6, at 250 ("[T]he jury panel was nothing like a random sample from the community.").

24. The *Reference Manual* states that "the null hypothesis says that the panel is like 350 persons drawn at random from a large population that is 50% female." *See* REFERENCE MANUAL, *supra* note 6, at 249. This articulation is not the null hypothesis, which is the statement one is trying to disprove (and the attorney is presumably not trying to disprove from this sample that the proportion of women in the population is 50%). *Id.* at 241. The null hypothesis here is instead that the panel selection process was fair and would result in 50% of the panel (n = 175) being women (corresponding to the 50% estimate for the general population). *Id.* at 249. Instead, only 29% of the panel were women (n = 102) and the question is how surprised we should be with only 102 women on the panel when we expected 175. *Id.* The answer is: so surprised that we should seriously doubt the veracity of the null hypothesis.

B. Use of P-Values in Pharmaceutical Cases

Studies on the efficacy of pharmaceutical products fundamentally use p-values to test those products against placebos.[25] Specifically, a new drug will be tested with the null hypothesis assuming the drug is ineffective and there is no difference between the treated and the control (i.e., untreated or placebo) groups. Clinical trials of the drug that then result in low p-values are seen to indicate that the drug is effective, in other words, that the null hypothesis is false.[26]

Such efficacy analyses are mandated by the Kefauver-Harris Drug Amendments to the Federal Food Drug and Cosmetic Act,[27] under which manufacturers of drug products must disclose all potential side effects of their drugs and establish a drug's effectiveness by substantial evidence.[28] Effectiveness is defined as a drug having health benefits superior to those that could be obtained through use of a placebo, as tested in a controlled situation such as a clinical trial.[29]

---

25. THOMAS D. COOK & DAVID L. DEMETS, INTRODUCTION TO STATISTICAL METHODS FOR CLINICAL TRIALS 335 (Thomas D. Cook & David L. DeMets eds., 1st ed. 2008).

26. *Id.*

27. Drug Amendments Act of 1962, Pub. L. No. 87–781, 76 Stat. 780 (codified as amended at 21 U.S.C. § 301–399i).

28. *See* FDA, GUIDANCE FOR INDUSTRY: PROVIDING CLINICAL EVIDENCE OF EFFECTIVENESS FOR HUMAN DRUGS AND BIOLOGICAL PRODUCTS 3 (1998).

29. *See* Jeremy A. Greene & Scott H. Podolsky, *Reform, Regulation, and Pharmaceuticals—The Kefauver-Harris Amendments at 50*, 367 NEW ENG. J. MED. 1481 (2012). Efficacy must be shown to authorize FDA-approved use of a drug, though "off-label" use of pharmaceuticals is common. AGATA BODIE, CONG. RSCH. SERV., R45792, OFF-LABEL USE OF PRESCRIPTION DRUGS (2021).

> When the Food and Drug Administration (FDA) approves a drug for sale in the United States, the approval includes a section entitled "Indications for Use." This section lists the one or more diseases, conditions, or symptoms for which the drug's sponsor (usually the manufacturer) has provided, to FDA's satisfaction, evidence in support of the drug's safety and effectiveness. FDA approval is also based on its review of the drug's dosage, packaging, manufacturing plan, and labeling. Before changing any of those elements, the sponsor must inform, and usually receive permission from, FDA.
>
> In essence, FDA regulates all approval and post-approval aspects of a drug product. But FDA traditionally has not regulated the practice of medicine. Physicians, therefore, may prescribe an FDA-approved drug for indications that FDA has not reviewed for safety and effectiveness. Those uses, furthermore, are not addressed in the labeling information regarding, among other things, dosing, warnings about interactions with other drugs, and possible adverse events.
>
> . . . .
>
> Estimates for how common off-label prescriptions are in the United States are hardly precise. Credible researchers have estimated they make up as little as 12% and as much as 38% of doctor-office prescriptions.

Id.

The process of clinical trials for Investigational New Drug Applicants occurs in three phases. Phase 1 introduces the drug to human subjects.[30] A Phase 1 study is small, typically between 20 and 80 individuals, and focuses on safety: determining potential side effects, dosages, and excretion of the drug from the body.[31]

Phases 2 and 3 are geared toward showing effectiveness, and further evaluating side effects and risks.[32] The sample size in Phase 2 is still relatively small, typically no more than several hundred individuals.[33] Once Phase 2 has gathered enough preliminary data to show the drug's effectiveness, testing moves to Phase 3. Phase 3 is simply a larger trial of drug effectiveness, with a sample size from a few hundred to a few thousand participants that allows the Food and Drug Administration (FDA) to evaluate whether the drug is effective in various populations and in modified doses.[34] Phase 3 gives the FDA a larger set of data in order to conduct a risk-benefit analysis and decide if the drug is effective enough to be put on the market with appropriate labeling.[35]

As noted, FDA approval must be premised on "substantial evidence," defined as:

> Evidence consisting of adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of which it could fairly and responsibly be concluded by such experts that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the labeling or proposed labeling thereof.[36]

Although no precise methodology is articulated either as a matter of statute or regulation, the FDA typically has required at least two such "adequate and well-controlled" studies to meet the substantial evidence

30. 21 C.F.R. § 312.21(a)(1) (2023).
31. *Id.*
32. *Id.* § 312.21(b)–(c).
33. *Id.* § 312.21(b).
34. *The FDA's Drug Review Process: Ensuring Drugs Are Safe and Effective*, FDA, https://www.fda.gov/drugs/information-consumers-and-patients-drugs/fdas-drug-review-process-ensuring-drugs-are-safe-and-effective (last visited Dec. 15, 2023).
35. 21 C.F.R. § 312.21(c) (2022).
36. 21 U.S.C. § 355(d) (2018).

standard.[37] As a key metric used to decide whether a study outcome shows anything at all (meaning the outcome is "statistically significant"), p-values are a crucial part of that approval process and any subsequent litigation.

In the three cases described below, the validity of p-values in relevant pharmaceutical studies—whether a solution treated dry eye (*Allergan v. Teva Pharmaceuticals*),[38] whether an antidepressant led to an increased risk of suicide (*Vanderwerf v. SmithKline Beecham Corporation*),[39] and whether a drug for osteoporosis could also limit the onset of breast cancer (*Zeneca v. Eli Lilly*)[40]—was vital to a judicial decision under three different sets of statutes, regulations, and decisional law.[41] Each case fundamentally turned, however, on the interpretation of p-values, and the losing party faced judicial criticism for its improper extrapolations.[42] This Part of the Article will briefly introduce the issues, while the next Part will further unpack the statistical analyses.[43]

First, in *Allergan v. Teva Pharmaceuticals*,[44] p-values were used to evaluate the "obviousness" factor of a patentability claim to determine whether a patent could stand.[45] Allergan asserted a patent for a cyclosporin formulation to treat dry eye under the brand name Restasis, particularly a

---

37. FDA, GUIDANCE FOR INDUSTRY, *supra* note 28, at 3. The FDA explains:

> The usual requirement for more than one adequate and well-controlled investigation reflects the need for *independent substantiation* of experimental results. A single clinical experimental finding of efficacy, unsupported by other independent evidence, has not usually been considered adequate scientific support for a conclusion of effectiveness.
>
> . . . .
>
> Independent substantiation of experimental results . . . provid[es] consistency across more than one study, thus greatly reducing the possibility that a biased, chance, site-specific, or fraudulent result will lead to an erroneous conclusion that a drug is effective.

Id. at 4–5.

38. *See* Allergan, Inc. v. Teva Pharms. USA, Inc., No. 2:15-cv-1455-WCB, 2017 WL 4803941, at *1 (E.D. Tex. Oct. 16, 2017).

39. *See* Vanderwerf v. SmithKline Beecham Corp., 529 F. Supp. 2d 1294 (D. Kan. 2008).

40. *See* Zeneca Inc. v. Eli Lilly & Co., No. 99 CIV. 1452(JGK), 1999 WL 509471, at *1 (S.D.N.Y. July 19, 1999).

41. *See infra* text accompanying notes 46–67.

42. *See, e.g.*, *Zeneca Inc.*, 1999 WL 509471, at *34 ("Courts in this Circuit and elsewhere routinely enjoin claims of proven therapeutic efficacy on the ground that the underlying tests are irrelevant and/or unreliable to support them.").

43. *See infra* Part II.

44. *See* Allergan, Inc. v. Teva Pharms. USA, Inc., No. 2:15-cv-1455-WCB, 2017 WL 4803941, at *1 (E.D. Tex. Oct. 16, 2017).

45. *See* 35 U.S.C. § 103 (2018) ("A patent for a claimed invention may not be obtained . . . if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious . . . to a person having ordinary skill in the art to which the claimed invention pertains.").

0.05% formulation as compared to a 0.1% formulation for that treatment.[46] At the center of the dispute were a Phase 2 study and a Phase 3 study testing the formulations and their accompanying reports (*Stevenson* and *Sall*, respectively).[47] The Phase 2 study tested the performance of multiple formulations, including the 0.05% and 0.1% formulations.[48] In that study, those two formulations far outperformed the others, leading Stevenson to recommend only testing those two formulations in Phase 3.[49] Following the results of the Phase 3 study, Allergan filed and was granted a patent for, among others, the 0.05% cyclosporin formulation.[50] After receiving the 0.05% formulation patent, Allergan sued Teva Pharmaceuticals for having produced generic alternatives to the 0.05% formulation.[51] In response, Teva challenged the validity of Allergan's patent on the 0.05% formulation, arguing that formulation was obvious and thus a patent should not have been granted.[52] Allergan countered, pointing out that the 0.1% cyclosporin formulation outperformed the 0.05% formulation in the Phase 2 study, but the opposite result occurred in the Phase 3 study.[53] This unexpected result in Phase 3, Allergan argued, was sufficient to show that the patent was non-obvious.[54] As will be further discussed below, "outperformance" was largely measured by Allergan's choice of drug effectiveness as reported by participants—help with eye dryness or eye grittiness, for example—and as measured by p-values associated with the data collected for those specific measures.[55]

Second, in *Vanderwerf v. SmithKline Beecham Corporation*,[56] the plaintiffs' husband and father committed suicide after taking defendant's antidepressant drug, Paxil (paroxetine). The plaintiffs alleged that Paxil caused an increased risk of suicide and causally led to William Vanderwerf's death.[57] To support this contention, the plaintiffs primarily relied on a FDA report, *Clinical Review: Relationship Between Antidepressant Drugs and Suicidality*, which contained extensive analysis of the clinical effects of many

46. *Allergan*, 2017 WL 4803941, at *1.
47. *Id.* at *5–6.
48. *Id.* at *5.
49. *Id.* at *6.
50. *Id.* at *5, *8.
51. *Id.* at *1.
52. *Id.* at *14–15.
53. *Id.* at *8.
54. *Id.*
55. *See infra* Part III.A.
56. Vanderwerf v. SmithKline Beecham Corp., 529 F. Supp. 2d 1294 (D. Kan. 2008).
57. *Id.* at 1306.

antidepressants, including Paxil.[58] The plaintiffs primarily relied on the study's finding of a statistically significant increased risk of "preparatory acts [of suicide] or worse" for adult Paxil patients with all psychiatric disorders compared to placebo ($p = 0.02$).[59] However, this value was related only to one of the secondary endpoints of the study—the primary focus of the study was suicidal ideation—and the court ultimately found the p-value associated with a secondary endpoint unreliable.[60]

---

58. MARC B. STONE & M. LISA JONES, CLINICAL REVIEW: RELATIONSHIP BETWEEN ANTIDEPRESSANT DRUGS AND SUICIDALITY IN ADULTS 8 (2006). *Vanderwerf* touches on the broader debate regarding the scope of warnings that FDA should provide, based on what evidence exists. *Vanderwerf*, 529 F. Supp 2d at 1302. *See* Tamsen Valoir & Shubha Ghosh, *FDA Preemption of Drug and Device Labeling: Who Should Decide What Goes on a Drug Label?*, 21 HEALTH MATRIX: J. L.-MED. 555, 585–86 (2011) (footnotes omitted) (noting that FDA ultimately required a "black box warning" of increased suicide risk in children and adolescents taking antidepressants but terming that decision "less compelling in hindsight" based on "differences in coding between various trials, low numbers of young patients, and the fact that the adverse event—suicide—is one of the same outcomes as untreated depression"). *Id.* (emphasis removed) (footnotes omitted). The authors further explain that:

> [The World Health Organization (WHO)] has concluded that SSRIs reduce the overall risk of suicide. Further, evidence suggests that treatment of childhood depression with these drugs has decreased since the black box warnings, and that suicides have increased at the same time. [They conclude that] [t]his example illustrates the consequences of over-warning and the failure to treat serious medical problems.

Id. at 586–87 (emphasis removed) (footnotes omitted); *see also* David A. Kessler & David C. Vladeck, *A Critical Examination of the FDA's Efforts to Preempt Failure-to-Warn Claims*, 96 GEO. L.J. 461, 466 (2008) (explaining that, at approval, the FDA is in the best position to balance risks and benefits of a drug, but even rare risks emerge once a drug is on the market and FDA processes for gathering data after approval "are 'relatively crude and often ineffective"). A "black box warning" is:

> [G]enerally reserve[d by the FDA] . . . for serious or life-threatening risks that best can be minimized by conveying critical information to the prescribing doctor in a highlighted manner. A decision by FDA to set apart a particular drug with a black box warning has serious implications for the licensed practitioner, the pharmacist, the patient, the pharmaceutical manufacturer, and the distributor. Nevertheless, FDA has not articulated specifically the scope of studies it relies on or the special circumstances in which the agency would impose this special warning.

Judith E. Beach et al., *Black Box Warnings in Prescription Drug Labeling: Results of a Survey of 206 Drugs*, 53 FOOD & DRUG L.J. 403, 403 (1998).

59. *Vanderwerf*, 529 F. Supp. 2d at 1308.

60. *Id.* at 1308–09. FDA, MULTIPLE ENDPOINTS IN CLINICAL TRIALS: GUIDANCE FOR INDUSTRY 6 (2022). The FDA explains:

> Endpoints in adequate and well-controlled drug trials are usually grouped hierarchically, often according to their clinical importance, but also taking into consideration the expected frequency of the endpoint events and anticipated drug effects. The critical determination for grouping endpoints is whether they are intended to establish effectiveness to support approval or intended to demonstrate additional meaningful effects. Endpoints critical to establish effectiveness for approval are often designated as primary endpoints. Secondary endpoints can provide useful description to support the primary endpoint(s) and/or demonstrate additional clinically important effects. The third category in the hierarchy includes

Third, in *Zeneca v. Eli Lilly*,[61] Zeneca sued Eli Lilly under the federal Lanham Act and a New York statute prohibiting unfair competition and deceptive trade practices. Zeneca manufactured and sold Nolvadex, a FDA-approved drug for reducing the incidence of breast cancer in women at high risk of developing the disease.[62] Eli Lilly manufactured a competing drug, Evista, which the company marketed as if it had been proven to reduce the risk of breast cancer, despite its approval by the FDA only for the prevention of osteoporosis.[63] The issue in *Zeneca* thus turned on the evidence that Eli Lilly had to support these marketing claims, based primarily on the *Multiple Outcomes of Raloxifene Evaluation* (MORE) study, which was one of ten studies conducted by Eli Lilly to gather the data necessary for the FDA to approve Evista for the prevention of postmenopausal osteoporosis.[64] Eli Lilly sought to base its marketing claims for breast cancer effects on the MORE study—and the MORE study did show a statistically significant increase in reduction of invasive breast cancer by Evista compared to placebo ($p < 0.001$).[65] However, because the MORE study had been designed to evaluate treatment of osteoporosis (not breast cancer), there were few cases of breast cancer among the subjects of the study, and that small sample size undercut the probity of the p-value identified.

II. MEANING OF P-VALUES AND UNDERSTANDING "THE MATH"

A. P-Values in the Scientific Community

Although p-values are commonly taught and utilized by scientists in a variety of fields, there is continued discourse as to whether p-values adequately perform their key function as a measure for evaluating whether studies can be relied upon to support relevant conclusions.[66] This discourse has gone so far as to lead the American Statistical Association (ASA) to release a statement urging work to "steer research into a 'post $p < 0.05$

> all other endpoints, which are referred to as exploratory. Exploratory endpoints can include endpoints for research purposes or for new hypotheses generation. Each category in the hierarchy can contain a single endpoint or a family of endpoints.

Id.

61. Zeneca Inc. v. Eli Lilly & Co., No. 99 CIV. 1452(JGK), 1999 WL 509471, at *1 (S.D.N.Y July 19, 1999).

62. *Id.*

63. *Id.*

64. *Id.* at *5.

65. *Id.* at *25.

66. *See, e.g.*, Murtaugh, *supra* note 7, at 613; Lewis G. Halsey et al., *The Fickle* P *Value Generates Irreproducible Results*, 12 NATURE METHODS 179, 179 (2015); Stefan Wellek, *A Critical Evaluation of the Current "*P*-Value Controversy"*, 59 BIOMETRICAL J. 854, 864 (2017).

era.'"[67] At the same time, the ASA's President, Jessica Utts, an *emerita* professor of statistics at the University of California, Irvine,[68] pointed out that research with statistically significant outcomes is more likely to get published, potentially incentivizing researchers to utilize inappropriate research approaches. Such practices can include "p-hacking,"[69] "data dredging,"[70] or "cherry-picking,"[71] in all of which researchers strive to

67. Press Release, Am. Stat. Ass'n, Am. Stat. Ass'n Releases Statement on Stat. Significance and *P*-Values (Mar. 7, 2016), https://www.amstat.org/asa/files/pdfs/p-valuestatement.pdf. The statement's six principles are:

> 1. *P-values can indicate how incompatible the data are with a specified statistical model.*
> 2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
> 3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
> 4. *Proper inference requires full reporting and transparency.*
> 5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
> 6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

Id. (quoting Lakshmi Narayana Yaddanapudi, *The American Statistical Association Statement on* P-*Values Explained*, 32 J. ANAESTHESIOLOGY CLINICAL PHARMACOLOGY 421, 421–23 (2016)) (alteration in the original).

68. *Home Page for Professor Jessica Utts*, UNIV. OF CAL., IRVINE, https://www.ics.uci.edu/~jutts/ (last visited Dec. 15, 2023).

69. Some differentiate *p-hacking* from data *dredging* and others do not. For an example differentiating the concepts, see Chittaranjan Andrade, *HARKing, Cherry-Picking,* P-*Hacking, Fishing Expeditions, and Data Dredging and Mining as Questionable Research Practices*, 82 J. CLINICAL PSYCHIATRY e1, e2 (2021) ("The difference between *P*-hacking and data dredging is that whereas *P*-hacking usually refers to the dragging of statistical significance out of data related to one or more hypotheses of interest, data dredging is the extensive search for significant relationships in a dataset without necessarily having specific hypotheses in mind."). Other terms referring to similarly questionable or detrimental research practices include HARKing (Hypothesizing After Results are Known), cherry-picking, "fishing expeditions," and data mining. *Id.* at e1–e2.

70. With the advent of Big Data, the term "data dredging" is a common form of data mining used in questionable research analyses. *See* Rahul Awati, *Data Dredging (Data Fishing)*, TECHTARGET, https://www.techtarget.com/searchdatamanagement/definition/data-dredging (last visited Dec. 15, 2023) (defining data dredging as a type of data mining practice); *see also Data-Dredging Bias*, CATALOGUE OF BIAS, https://catalogofbias.org/biases/data-dredging-bias/ (last visited Dec. 15, 2023) (discussing data dredging bias).

71. Cherry-picking is generally recognized as a results-driven analysis in which a researcher selects specific results to emphasize (cherry-picked) that agree with the a priori viewpoints of the study investigator by either, for example, focusing on the statistically significant results (and ignoring the non-significant result) or by focusing discussion on specific results that promote a viewpoint and neglecting others. *See, e.g.*, *In re* Lipitor (Atorvastatin Calcium) Mktg., Sales Pracs. & Prods. Liab. Litig. (No. II) MDL 2502, 892 F.3d 624, 634 (4th Cir. 2018) ("Result-driven analysis, or cherry-picking, undermines principles of the scientific method and is a quintessential example of applying methodologies (valid or otherwise) in an unreliable fashion."); *see also* EEOC v. Freeman, 778 F.3d 463, 469–70 (4th Cir. 2015) (Agee, J., concurring) (citing examples of expert testimony exclusion based on "cherry-picked" data and explaining that "'[c]herry-picking' data is essentially the converse of omitting it: just as omitting data

achieve small p-values as a primary goal of their research or to inappropriately emphasize certain a priori viewpoints, rather than seeking broader, underlying truths which more fully represent the gamut of evidence.[72] Utts concluded:

> The contents of the ASA statement and the reasoning behind it are not new—statisticians and other scientists have been writing on the topic for decades . . . . But this is the first time that the community of statisticians, as represented by the ASA Board of Directors, has issued a statement to address these issues.[73]

Like scientists, lawyers have often been quick to adopt threshold meanings for p-values in seeking to demonstrate legally meaningful outcomes and should be as wary as scientists in performing such a superficial and potentially incomplete or misleading analysis.

Each p-value is critically tied to the study from which it has been calculated.[74] Even if a study's power[75] is high—with a generally accepted 80% or 90% probability that an effect for a given sample size will be identified in that study through a p-value that crosses the desired threshold of significance—the p-value is not necessarily stable.[76] For example, if a study

might distort the result by overlooking unfavorable data, cherry-picking data produces a misleadingly favorable result by looking only to 'good' outcomes."); *see also In re* Bextra & Celebrex Mktg. Sales Pracs. & Prods. Liab. Litig., 524 F. Supp. 2d 1166, 1176–78 (N.D. Cal. 2007) (ruling expert opinion inadmissible because the expert "reache[d] his opinion by first identifying his conclusion—causation at 200 mg/d—and then cherry-picking observational studies that support his conclusion and rejecting or ignoring the great weight of the evidence that contradicts his conclusion").

72. *In re Lipitor*, 892 F.3d at 634.

73. *See* Am. Stat. Ass'n, *supra* note 67.

74. *See* Halsey et al., *supra* note 66, at 180.

75. *See* REFERENCE MANUAL, *supra* note 6, at 254 ("Power is the chance that a statistical test will declare an effect when there is an effect to be declared."); see also Pritha Bhandari, *Statistical Power and Why It Matters | A Simple Introduction*, SCRIBBR (June 22, 2023), https://www.scribbr.com/statistics/statistical-power/, where Bhandari explains:

> A power analysis is made up of four main components. If you know or have estimates for any three of these, you can calculate the fourth component.
>
> - **Statistical power:** the likelihood that a test will detect an effect of a certain size if there is one, usually set at 80% or higher.
> - **Sample size:** the minimum number of observations needed to observe an effect of a certain size with a given power level.
> - **Significance level (alpha) [p-value]:** the maximum risk of rejecting a true null hypothesis that you are willing to take, usually set at 5%.
> - **Expected effect size:** a standardized way of expressing the magnitude of the expected result of your study, usually based on similar studies or a pilot study.

Id. (alteration in original).

76. *See* Greenland et al., *supra* note 9, at 342–43.

at a certain sample size and a power even of 90% produced a statistically significant result at p = 0.03, a study seeking to replicate those results would not likely result in a p-value that is close to the original p = 0.03 or even be statistically significant, but rather fall somewhere in an extremely large range of 0.0–0.6.[77] In fact, there would only be a 56.1% chance that the p-value would be less than 0.05 and thus replicate the finding of the first study.[78]

The p-value is also frequently misinterpreted.[79] Even in the scientific community, working scientists have engaged in faulty analysis based on p-values, undercutting the scientific literature in many fields.[80] An editorial co-

77. *Id.* at 343; *see also* Geoff Cumming, *Replication and* p *Intervals:* p *Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better*, 3 PERSPS. ON PSYCH. SCI. 286, 286 (2008). ("If my experiment results in $p = 0.05$, for example, what p in an exact replication—with a new sample of participants—likely to give? Surprisingly, the answer is 'Pretty much anything.'").

78. *See* Greenland et al., *supra* note 9, at 343. Confidence intervals can perhaps more readily than p-values be used to determine replicability of statistical significance. *See* Cumming, *supra* note 77, at 286–87. Cumming claims:

> Confidence intervals (*CIs*), by contrast [to p-values], give useful information about replication. There is an 83% [probability] that a replication gives a mean that falls within the 95% CI from the initial experiment. . . . Any 95% CI can thus be regarded as an 83% prediction interval for a replication mean. The superior information that CIs give about replication is a good reason for researchers to use CIs rather than *p* values wherever possible.

Id. at 286 (citation omitted).

79. *See, e.g.*, Wellek, *supra* note 66, at 859; Greenland et al., *supra* note 9, at 340–42.

80. *See* Greenland et al., *supra* note 9, at 346. The authors explain:

> Less often stated is the even more crucial assumption that the analyses themselves were not guided toward finding nonsignificance or significance (analysis bias), and that the analysis results were not reported based on their nonsignificance or significance (reporting bias and publication bias). Selective reporting renders false even the limited ideal meanings of statistical significance, *P* values, and confidence intervals. Because author decisions to report and editorial decisions to publish results often depend on whether the *P* value is above or below 0.05, selective reporting has been identified as a major problem in large segments of the scientific literature.

Id.; *see also* Christine Parry, *Taking the P: Why the Founder of* P*-values Would Be Turning in His Grave*, PHARM. J. (Jan. 7, 2021), https://pharmaceutical-journal.com/article/feature/taking-the-p-why-the-founder-of-values-would-be-turning-in-his-grave. While interviewing Janet Peacock, an epidemiologist and biomedical data scientist at Dartmouth College, she states:

> [H]ow *P*-values are used today has gradually morphed further and further away from Fisher's original intention [because Fisher saw p-values as probabilities.] [Other scientists building on his work] "introduced the idea that *P*-values could be used to make decisions through a cut-off value. While that's useful, that's where the potential difficulties start to creep in."

Id. Parry then continues by explaining that:

> This issue, referred to as the dichotomi[z]ation of the *P*-value, is a core argument against statistical significance—using $P \leq 0.05$ as a cut-off to sift clinical research into binary categories, with statistically significant results considered meaningful and the rest discarded.

authored by the Executive Director of the ASA acknowledged this phenomenon and its effect on the scientific community.[81] Misinterpretations of the p-value become even more concerning in harder-to-control study areas such as medicine,[82] leading some researchers to believe that the harms of

---

Id. Parry further elaborates that this phenomenon is particularly concerning in the context of medical studies, specifically:

> [A] study can have a large sample size that lowers the *P*-value to statistically significant levels, while the treatment effect remains clinically negligible. Choosing to repeat a test to increase sample size, and therefore artificially reduce the *P*-value, is just one example of how *P*-values can be manipulated—part of the popularly termed "*P*-hacking" toolkit.

Id. Parry further describes:

> *P*-hacking is an umbrella term for actions that tweak the *P*-value towards statistical significance—intentionally or not. *P*-hacking techniques often take advantage of these compositional, influencing factors; for instance, in addition to altering the sample size, *P*-values can be manipulated by emphasi[z]ing treatment effect size, such as [by] removing outlying results. One of the more insidious forms of *P*-hacking is the issue of multiplicity; because *P*-values are a frequency probability, if enough *P*-values are calculated, then one is bound to be statistically significant. The risk of returning a false positive in this way can hit 40% if even 10 *P*-values are calculated.

Id. Though not all researchers would agree with her characterization of experiment repetition as p-hacking, the multiplicity issues Parry identifies are methodological concerns and can easily arise when numerous primary and secondary endpoints are reported, as is common in pharmaceutical studies. *Id.* Suitable corrections can be done *post-hoc*, the most common of which are the Benjamini-Hochberg correction and the more stringent Bonferroni adjustment. *See* Yoav Benjamini & Yosef Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, 57 J. ROYAL STAT. SOC'Y SERIES B (METHODOLOGICAL) 289, 291 (1995).

81. *See* Ronald L. Wasserstein & Nicole A. Lazar, Editorial, *The ASA Statement on* P-*Values: Context, Process, and Purpose*, 70 AM. STATISTICIAN 129, 131 (2016).

> [T]he scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the *p*-value. The issues touched on here affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law.

Id.

82. Many major medical journals (through the International Committee of Medical Journal Editors, or ICMJE) seek to control potentially questionable research practices, including manipulation of p-values, by requiring that studies be "pre-registered" during study planning, such that proposals, study designs, planned statistical methods, and other background information are made available to the public for review prior to study initiation. *See Clinical Trials*, INT'L COMM. OF MED. J. EDS., https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html (last visited Dec. 15, 2023). The ICMJE claims:

> Briefly, the ICMJE requires, and recommends that all medical journal editors require, registration of clinical trials in a public trials registry at or before the time of first patient enrollment as a condition of consideration for publication. Editors requesting inclusion of their journal on the ICMJE website list of publications that follow ICMJE guidance should recognize that the listing implies enforcement by the journal of ICMJE's trial registration policy.
>
> . . . .

statistical testing in these areas outweigh the benefits.[83] These misconceptions show a general trend: based on pressure to publish or otherwise, researchers want p-values to be more probative than they actually are.[84] And many misconceptions with respect to p-values likewise are founded on a belief that the measure is far more convincing than it actually is.[85] In all three of the cases analyzed in this Article, lawyers made crucial mistakes with p-values, emphasizing the need for the legal profession to understand potential weaknesses in p-value analysis when advocating for clients.[86]

B. P-Values in the Legal Community

Fundamental legal guidance on the use of p-values and statistics is found in the *Reference Manual on Scientific Evidence*, designed for use by judges and widely relied upon by lawyers involved in cases with statistical implications.[87] The *Reference Manual* guides courts to evaluate the magnitude and circumstances of p-values and highlights the importance of tying p-values to strong studies, emphasizing that a highly significant difference might have no practical impact if, for example, the study uses a small sample size.[88] The *Reference Manual* fails, however, to discuss many of the detailed criticisms of p-values that have arisen in the scientific community[89] and that may confront judges and lawyers in individual cases.

> The purpose of clinical trial registration is to prevent selective publication and selective reporting of research outcomes, to prevent unnecessary duplication of research effort, to help patients and the public know what trials are planned or ongoing into which they might want to enroll, and to help give ethics review boards considering approval of new studies a view of similar work and data relevant to the research they are considering.

Id. One example of such a registration website is managed by the National Institutes of Health. *See Clinical Trials.gov*, NAT'L. LIBR. OF MED., https://clinicaltrials.gov/ (last visited Dec. 15, 2023).

83. *See* Greenland et al., *supra* note 9, at 346.

84. *See* Wasserstein & Lazar, *supra* note 81, at 131 ("Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither.").

85. *See* Greenland et al., *supra* note 9, at 340–42.

86. *See infra* Part IV.

87. *See* REFERENCE MANUAL, *supra* note 6, at 249–57.

88. *Id.* at 253 (discussing the relationship between the p-value and sample sizes, explaining that "[a] 'significant' effect can be small. Conversely, an effect that is 'not significant' can be large. By inquiring into the magnitude of an effect, courts can avoid being misled by *p*-values"). The *Reference Manual* largely focuses on sample size as a proxy for power; while important (and a small sample size in a study will indeed undercut its power), a power analysis requires discussion of other variables, including the effect size of interest, the sample size, and the selected p-value. *See supra* note 75 and accompanying text.

89. *See infra* Part II.A.

Moreover, even when caution is specifically articulated—as with respect to small sample sizes—some details are missing. For example, the *Reference Manual* warns that underlying assumptions about small samples are hard to validate, p-values may be difficult to calculate for hypotheses of interest, and that small samples are generally unreliable.[90] Despite this examination, the *Reference Manual* nonetheless fails both to appreciate the nuances of defining when a sample is "small"[91][SN1] and to discuss the tendency for small (underpowered) studies to inflate or exaggerate the size of any observed effects.[92]

---

90. REFERENCE MANUAL, *supra* note 6, at 255.

91. No universally applicable small or large sample size can be said to be inadequate or adequate for a given question. Some researchers will incorrectly argue from basic statistical textbooks that an n = 30 is a *large* sample size. *Small Sample Estimation of a Population Mean*, SAYLORDOTORG, https://saylordotorg.github.io/text_introductory-statistics/s11-02-small-sample-estimation-of-a-p.html (last visited Dec. 15, 2023) ("A sample is considered small when n < 30."). This fallacy is based on the "Student's *t*-distribution" tables where it is said that after n = 30, the t-statistics are not materially different from (large sample) z-statistics. *The T-Distribution*, STATISTICS ONLINE https://online.stat.psu.edu/stat500/book/export/html/521 (last visited Dec. 15, 2023) (emphasis in original). While true, this is often irrelevant to the task of interest. Generally, a sample size can be said to be sufficient a priori if the desired power to detect an effect size of interest at a given probability is achieved, and these power calculations are (or should be) mathematically calculated prior to the trial or experiment and require various assumptions regarding means or other effect sizes, sample variability, and response rates. *See, e.g.*, Chittaranjan Andrade, *Sample Size and Its Importance in Research*, 42 INDIAN J. PSYCH. MED. 102, 102–03 (2020). Andrade explains that sample size often is driven by study practicality—"[m]any investigators increase the sample size by 10%, or by whatever proportion they can justify, to compensate for expected dropout, incomplete records, biological specimens that do not meet laboratory requirements for testing, and other study-related problems,"—so that a "sample size necessary to be 80% certain of identifying a statistically significant outcome[,] should the hypothesis be true for the population, with P for statistical significance set at 0.05" is reached, because to utilize a larger sample size that would increase power (to 90%, for example) and lower the significant threshold (to 0.01, for example), may render studies "more expensive and more difficult to conduct." *Id.* at 103.

92. For example, many researchers understand that a small or underpowered study may not have the ability to find an effect that is truly there; fewer are aware that such small/underpowered studies will tend to systematically inflate or magnify any statistically significant effects that *are* observed. *See* John P. A. Ioannidis, *Why Most Discovered True Associations Are Inflated*, 19 EPIDEMIOLOGY 640, 640 (2008) ("Newly discovered true (non-null) associations often have inflated effects compared with the true effect sizes."); Andrew Gelman & John Carlin, *Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors*, 6 PERSPS. ON PSYCH. SCI. 641, 641 (2014) ("In this article, we show that when researchers use small samples and noisy measurements to study small effects . . . a significant result is often surprisingly likely to be in the wrong direction and to greatly overestimate an effect."). For simulation methods designed to quantitatively estimate the degree of potential effect size inflation, see David J. Miller et al., *Emagnification: A Tool for Estimating Effect-Size Magnification and Performing Design Calculations in Epidemiological Studies*, 20 STATA J., 548, 548 (2020) (explaining that "[a]rtificial effect-size magnification (ESM) may occur in underpowered studies, where effects are reported only because they or their associated *p*-values have passed some threshold."). For associated conference presentation, see DAVID J. MILLER ET AL., U.S. EPA, EMAGNIFICATION: A TOOL FOR ESTIMATING EFFECT SIZE MAGNIFICATION AND PERFORMING DESIGN CALCULATIONS IN EPIDEMIOLOGICAL STUDIES (2019), https://www.stata.com/meeting/nordic-and-baltic19/slides/nordic19_miller.pdf.

A thorough understanding of p-values is critical in many legal contexts. Indeed, outside the pharmaceutical context that is the focus of this Article, p-values have been equally key to court decisions that illustrated crucial issues, like improperly selected juries[93] and Title VII disparate treatment claims.[94] For example, courts have recognized that p-values are not a "magic number." While discussing jury representation of African Americans, a federal district court noted that although the p-value of 0.092 offered by the defense was outside the traditional standard of 0.05, the absence of reaching the touchstone number did not mean there was no cause for caution and inquiry.[95] Likewise, in assessing a Title VII claim, the Tenth Circuit rejected a statistically significant p-value of 0.01231 as evidence of sex-based discrimination: not because of the number itself, but because the study of the defendant's employment records did not use a proper methodology.[96]

Evidentiary considerations may frustratingly limit the ability of lawyers to argue for p-values other than 0.05 or lower, because that number is "generally accepted by the scientific community."[97] Similar state standards for admission of expert testimony thus may prevent introduction of evidence simply because it is not deemed statistically significant.[98] However,

93. *See* United States v. Fell, No. 5:01-cr-12-01, 2018 U.S. Dist. LEXIS 228624, at*6, *8 (D. Vt. Aug. 6, 2018).

94. *See* Frappied v. Affinity Gaming Black Hawk, 966 F.3d 1038, 1052–53 (10th Cir. 2020).

95. *See Fell*, 2018 U.S. Dist. LEXIS 228624, at *6, *11.

96. *See Frappied*, 966 F.3d at 1052–53 ("Although the p-value is probative of whether Affinity discriminated against older women, because the plaintiffs did not compare older women to only older men in calculating it, the p-value does not itself give rise to a plausible inference of discrimination because of sex.").

97. Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 584, 593–94 (1993).

98. *See* General Electric Co. v. Joiner, 522 U.S. 136, 145–46 (1997) (finding no abuse of discretion to exclude expert testimony not deemed statistically significant because "[a] court may conclude that there is simply too great an analytical gap between the data and the opinion proffered"). Commentators have recognized, however, that not all courts may apply a sufficiently searching analysis in such cases:

> It has been relatively clear since . . . *Joiner* . . . that the proponent of expert testimony must make a showing of validity as applied as well as foundational validity. Unfortunately, in the past, many courts have glossed over that distinction. In their review of the proponent's validation studies, they have tended to focus on such considerations as the size of the study and the overall accuracy rate reported by the researchers. Those quantitative factors are highly relevant to a judge's decision on the issue of foundational validity, but they shed little light on the question of validity as applied.
>
> When the question is validity as applied, the courts must scrutinize both the validity studies and the test in the instant case more closely. . . . [A] hypothesis about the validity of a methodology is a conditional proposition: when certain factors or conditions are specified, what is the likely outcome of the use of the methodology? The judge must identify those factors and then inquire whether the same conditions obtained in the test are also found in the pending case. A validation

consistent with the scientific commentary questioning primary focus on a particular level of p-value, 0.05, some courts have recognized that a precise p-value does not by itself connote reliability.[99] Understanding these issues, lawyers must evaluate their reliance on statistical evidence and prepare relevant experts to explain the context for their conclusions without singular focus on the p-value, including study design, the quality of the data, and the underlying scientific mechanisms being tested. Understanding the causative inferences to be drawn from evaluation of p-values is particularly critical, because that is where a case may be won or lost, and no specific p-value can save the analysis of an expert who does not ensure that all conclusions are founded on relevant facts.[100]

These discussions have been highlighted in cases involving pharmaceuticals, where some of the most hard-fought statistical issues arise. For example, the Fourth Circuit, while discussing whether a statistician's testimony was admissible, stated that statistically significant p-values are not necessarily determinative of a case's outcome.[101] Likewise, taking a broad view of the evidence, the Ninth Circuit in addressing a violation of the Securities Exchange Act held that asserting an improper methodology in a drug study was insufficient to show a false or misleading statement.[102] Finally, in *Allergan v. Teva*, which will be discussed further in Part III below, the U.S. District Court for the Eastern District of Texas explained in dicta

> study supports an inference of reliability, satisfying Rule 702 only under the conditions of the study. By happenstance, an extrapolation beyond the conditions of the study may indeed be correct, but without more, the study furnishes no empirical support or justification for the extrapolation.

Edward J. Imwinkelried, *The Admissibility of Scientific Evidence: Exploring the Significance of the Distinction Between Foundational Validity and Validity as Applied*, 70 SYRACUSE L. REV. 817, 846–47 (2020) (footnotes omitted).

99. *See* Milward v. Acuity Specialty Prods. Grp., Inc., 639 F.3d 11, 24 (1st Cir. 2011) (finding error to "treat[] the lack of statistical significance as a crucial flaw"); Matrixx Initiatives, Inc. v. Siracusano, 563 U.S. 27, 40–41 (2011) ("[C]ourts frequently permit expert testimony on causation based on evidence other than statistical significance.") (citations omitted); Kennedy v. Collagen Corp., 161 F.3d 1226, 1229 (9th Cir. 1998); Henricksen v. Conocopillips Co., 605 F. Supp. 2d 1142, 1177 (E.D. Wash. 2009) ("[T]he absence of statistical support of causation is not necessarily fatal to a plaintiffs' [sic] case.").

100. *See In re* Brand Names Prescription Drugs Antitrust Litig., 186 F.3d 781, 788 (7th Cir. 1999) (upholding exclusion of expert testimony by a Nobel Prize-winning economist because its underlying factual assumptions were not supported in the record).

101. *See In re* Lipitor (Atorvastatin Calcium) Mktg., Sales Practices & Prods. Liab. Litig. (No. II) MDL 2502, 892 F.3d 624, 641 (4th Cir. 2018) ("Just as statistically significant evidence won't result in automatic admission, the absence of a p-value that is smaller than .05 (or some other threshold) isn't necessarily fatal to a case.").

102. *See In re* Rigel Pharms., Inc., Sec. Litig., 697 F.3d 869, 878 (9th Cir. 2012) ("[T]he district courts that have addressed this issue support our conclusion that merely alleging that defendants should have used different statistical methodology in their drug trials is not sufficient to allege falsity.").

that drawing a hard line for statistical significance was a flawed approach, and a p-value of slightly more than 0.05 should be considered informative.[103]

In the cases discussed in Part III, courts recognized the weaknesses in arguments based on the use of p-values. In addition, many of the concerns of the scientific community regarding p-values and their use in studies were reflected in the court opinions.[104] The three cases thus provide useful exemplars for counsel seeking to maximize the chances for success in making—or refuting—similar arguments in future cases.

III. USE OF EXPERTS AND CONNECTING STATISTICAL ANALYSIS TO THE LEGAL ISSUES

A. Allergan v. Teva

Recall from Part I that Allergan's patent for its 0.05% Cyclosporin A (CsA) formulation for treating moderate-to-severe dry eye disease was at issue.[105] Teva argued that Allergan's patent was invalid because extrapolation to the 0.05% formulation was obvious.[106] The "obviousness" condition for patentability requires that an invention not be obviously based on the prior art existing at the time the patent was filed.[107] Obviousness here turned on whether a "person of ordinary skill" in the pharmaceutical field would have found the 0.05% formulation performance in the Phase 3 trials

103. *See infra* Part III.A and accompanying text.

104. *See infra* Part III.A–C.

105. *See supra* Part I.B.

106. Allergan, Inc. v. Teva Pharms. USA, Inc., No. 2:15-cv-1455-WCB, 2017 WL 4803941, at *18 (E.D. Tex. Oct. 16, 2017).

107. *Id.* at *17. As the *Allergan* court explained:

> When an applicant seeks to overcome a *prima facie* case of obviousness by showing improved performance within a range that is within or overlaps with a range disclosed in the prior art, the applicant must "show that the [claimed] range is *critical*, generally by showing that the claimed range achieves unexpected results relative to the prior art range." "[O]ne way for a patent applicant to rebut a *prima facie* case of obviousness is to make a showing of 'unexpected results,' *i.e.*, to show that the claimed invention exhibits some superior property or advantage that a person of ordinary skill in the relevant art would have found surprising or unexpected. The basic principle behind this rule is straightforward—that which would have been surprising to one of ordinary skill in a particular art would not have been obvious."

Id. at *19 (first quoting *In re* Geisler, 116 F.3d 1465, 1469–70 (Fed. Cir. 1997); and then quoting *In re* Soni, 54 F.3d 746, 750 (Fed. Cir. 1995) (citation omitted)).

unexpected—using the "surprise" concept explained with respect to p-value interpretation[108]—after its performance in the Phase 2 trials.[109]

The Phase 2 study of Allergan's formulations was analyzed in Efficacy and Safety of Cyclosporin A Opthalmic Emulsion in the Treatment of Moderate-to-Severe Dry Eye Disease (*Stevenson Report*).[110] The *Stevenson Report* was intended to evaluate the safety and efficacy of CsA 0.05%, 0.1%, 0.2%, and 0.4% formulations.[111] The table below details the study's overall size as well as the sample size for each formulation.[112]

Table 1. Patient Disposition

	Moderate-to-Severe Dry Eye Disease (n = 90)				Intent-to-Treat Population (Total Enrollment) (n = 162)			
	Completed		Discontinued		Completed		Discontinued	
Treatment Group	n	%	n	%	n	%	n	%
Vehicle	16	100.0	0	0.0	30	90.9	3	9.1
CsA 0.05%	17	100.0	0	0.0	30	96.8	1	3.2
CsA 0.1%	18	94.7	1	5.3	30	93.8	2	6.3
CsA 0.2%	20	100.0	0	0.0	32	94.1	2	5.9
CsA 0.4%	17	94.4	1	5.6	28	87.5	4	12.5
Total	88	97.8	2	2.2	150	92.6	12	7.4

CsA = cyclosporin A.

Figure 1. Data from *Stevenson Report* (2000).[113]

The *Stevenson Report* used eight different outcome measures: rose bengal staining, superficial punctate keratitis, Schirmer tear test, symptoms of ocular discomfort, tear film debris, tear breakup time, frequency and amount of formulation used, and Ocular Surface Disease Index (OSDI).[114] The null hypothesis expected no differences between each of the formulations versus the placebo baseline, with the alternative hypothesis

108. *Id.* at *40. *See supra* notes 19–23 and accompanying text (discussing p-value surprise and rejection of the null hypothesis, after finding only 102 women in a jury panel).

109. *See* Allergan, Inc. v. Teva Pharms. USA, Inc., No. 2:15-cv-1455-WCB, WL 2017 4803941, at *18 (E.D. Tex. Oct. 16, 2017).

110. *See* Dara Stevenson et al., *Efficacy and Safety of Cyclosporin A Ophthalmic Emulsion in the Treatment of Moderate-to-Severe Dry Eye Disease: A Dose-Ranging, Randomized Trial*, 107 AM. ACAD. OPHTHALMOLOGY 967 (2000).

111. *Allergan*, 2017 WL 4803941 at *5.

112. *See* Stevenson et al., *supra* note 110, at 969–70 (discussing the subgroup analysis of the initial treatment population that revealed a group of subjects already suffering from moderate-to-severe dry eye disease).

113. *Id.* at 970 tbl.1.

114. *Id.* at 968.

expecting a beneficial difference.[115] The *Allergan* decision focused primarily on the relationship between the 0.05% and 0.1% CsA formulations because that was the focus of the patent dispute. For all of the major outcome measures, a p-value of 0.05 was considered statistically significant.[116]

In the *Stevenson Report*, the 0.1% formulation outperformed the 0.05% formulation on four measures (rose bengal staining, superficial punctate keratitis, Schirmer tear test, and OSDI score), while the 0.05% formulation outperformed the 0.1% formulation on two measures (sandy or gritty feeling, ocular dryness).[117] Discussing the results, the *Stevenson Report* noted that, despite there being no clear dose-response relationship between the different formulations, the 0.1% formulation produced the most consistent improvement in objective and subjective end points, while the 0.05% formulation produced the most consistent improvement in patient symptoms.[118] Based on this data, the report suggested that further studies should focus on these two formulations.[119]

The Phase 3 study was summarized in Two Multicenter, Randomized Studies of the Efficacy and Safety of Cyclosporine Ophthalmic Emulsion in Moderate to Severe Dry Eye Disease (*Sall Report*).[120] The *Sall Report* followed the recommendation of the *Stevenson Report* and focused on only the 0.05% and 0.1% formulations.[121] The study had nine outcome measures: corneal and interpalpebral conjunctival staining, Schirmer tear test, tear breakup time, OSDI score, the facial expression subjective rating scale, symptoms of dry eye, investigator's evaluation of global response to treatment, treatment success, and formulation usage.[122] Results were reported only if the comparison among all three groups (0.1% and 0.05% formulations and the pharmacologically inactive "vehicle" for the formulations) were significant at $p = 0.05$ and the pairwise comparison between a formulation

115. *Id.* at 969.

116. *Id.*

117. *Id.* at 973–74.

118. *Id.*

119. *Id.* at 974.

120. *See* Kenneth Sall et. al., *Two Multicenter, Randomized Studies of the Efficacy and Safety of Cyclosporine Ophthalmic Emulsion in Moderate to Severe Dry Eye Disease*, 107 AM. ACAD. OF OPHTHALMOLOGY 631 (2000) [hereinafter Sall Report].

121. *Id.*

122. Sall Report, *supra* note 120, at 633.

and the vehicle was significant at $p = 0.05$.[123] The table below details the study's size as well as the sample size for each formulation.[124]

Ophthalmology *Volume 107, Number 4, April 2000*

Table 1. Patient Disposition

	CsA 0.05%	CsA 0.1%	Vehicle
Enrolled	293	292	292
Completed	235 (80.2%)	218 (74.7%)	218 (74.7%)
Discontinued			
Lack of efficacy	1 (0.3%)	3 (1.0%)	3 (1.0%)
Adverse events	19 (6.5%)	29 (9.9%)	13 (4.5%)
Lost to follow-up	4 (1.4%)	3 (1.0%)	11 (3.8%)
Personal reasons	9 (3.1%)	14 (4.8%)	9 (3.1%)
Protocol or enrollment violations*	19 (6.5%)	21 (7.2%)	30 (10.3%)
Other†	6 (2.0%)	4 (1.4%)	8 (2.7%)

CsA = cyclosporin A.

* Protocol or enrollment violations included improper entry, non-compliance, and use of prohibited medications.

† Other included pregnancy, relocation, and removal from the study by the sponsor.

Figure 2. Data from *Sall Report* (2000).[125]

The study's method for measuring corneal staining is explained in depth in the *Sall Report*'s Outcome Measures section.[126] Improvement in corneal staining was significantly greater in both formulation groups than in the vehicle at four months ($p \leq 0.044$), in the 0.05% formulation group at six months ($p = 0.008$) and "a trend ([p] = 0.062) toward a significantly greater improvement in the CsA 0.1% group than the vehicle group" at six months ($p = 0.062$).[127]

123. *Id.* at 634. The *Sall Report* further explained:

> Specifically, a pair wise comparison between either cyclosporine group and vehicle groups is considered statistically significant if and only if (1) the overall comparison among the three groups is significant at the 0.05 level, and (2) the pair wise comparison between cyclosporine and vehicle is significant at the 0.05 level.

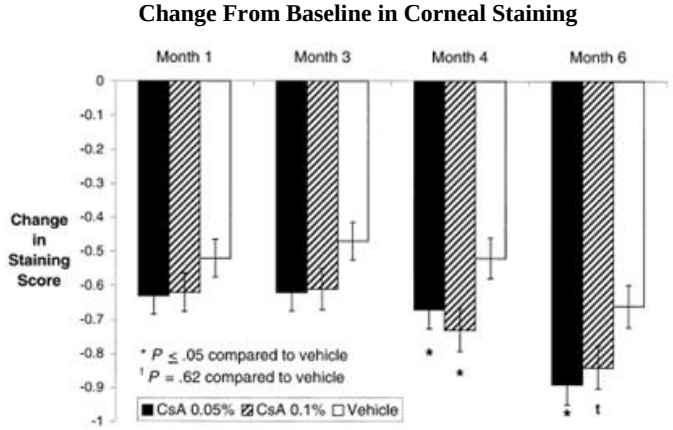
Id.

124. *Id.*

125. *Id.* tbl.1.

126. *Id.* at 633.

127. *Id.* at 635.



Change from baseline in corneal staining. Mean value ± standard error. Graded on a scale from 0 to 5.

Figure 3. Data from *Sall Report* (2000).[128]

The Schirmer tear test is used to measure whether one's tear glands produce enough tears to keep the eye moist.[129] The test is performed by placing filter paper on the eye for five minutes and then measuring the length of the paper that is moist after five minutes.[130] At three months, the 0.05% formulation performed significantly better than the vehicle ($p = 0.009$), and at six months both formulations had stronger results than the baseline.[131]

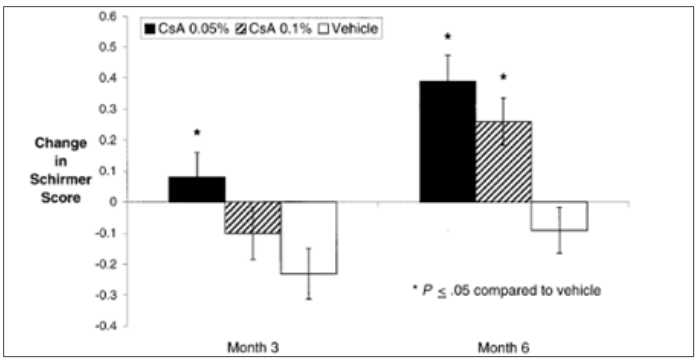
128. *Id.* fig.1.

129. *Schirmer's Test*, JOHNS HOPKINS SJÖGREN CTR., https://www.hopkinssjogrens.org/disease-information/diagnosis-sjogrens-syndrome/schirmers-test (last visited Dec. 15, 2023) (discussing Schirmer tear test and process).

130. *Id.*

131. Sall Report, *supra* note 120, at 635.

Change From Baseline in Categorized Schirmer Values Measured With Anesthesia

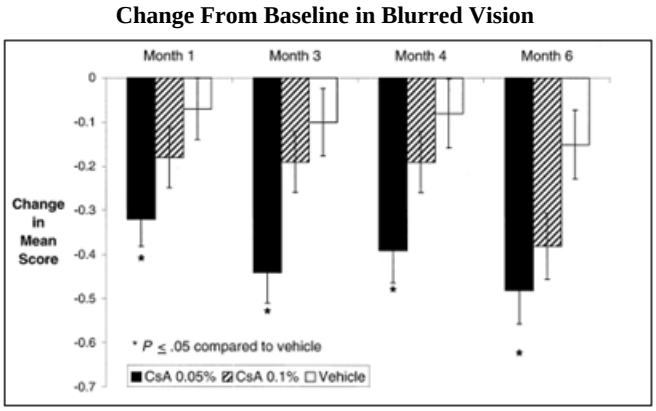


Change from baseline in categorized Schirmer values (measured with anesthesia). Mean value ± standard error. Categorized Schirmer values were graded on a 5-point scale as follows: 1 (< 3 mm/5 min), 2 (3–6 mm/5 min), 3 (7–10 mm/5 min), 4 (11–14 mm/5 min), and 5 (>14 mm/5 min) using the worse eye.

Figure 4. Data from *Sall Report* (2000).[132]

Among the various symptoms of dry eye disease, blurred vision saw the most significant differing measurements.[133] Both formulations saw statistically significant decreases at follow-up visits ($p \leq 0.012$), with the 0.05% formulation exhibiting the greatest improvement over the vehicle.[134] For all other variables under symptoms (dryness, sandy or gritty feeling, itching, photophobia, burning and stinging, and pain), "there were no statistically significant among-group differences."[135]

132. *Id.* fig.2.
133. *Id.* at 636.
134. *Id.*
135. *Id.*



Change from baseline in blurred vision. Mean value ± standard error. Graded on a scale from 0 to 4.

Figure 5. Data from *Sall Report* (2000).[136]

Summarizing results, the *Sall Report* indicated that both formulations led to improvements in two objective signs of dry eye compared to the vehicle, while treatment with the 0.05% formulation performed better than the vehicle in three subjective parameters.[137] The remainder of the *Sall Report* spoke generally about cyclosporin being a potential treatment for dry eye disease and refrained from stating if either formulation performed better.[138]

In the subsequent litigation, both the parties and the court relied on the *Stevenson Report* and the *Sall Report* for the obviousness analysis. The court first turned to the *Stevenson Report* and the results of the Phase 2 study, finding that:

> [T]he Court rejects Allergan's contention that a person of skill would look at the Schirmer scores and corneal staining results in Phase 2 and conclude that the 0.1% cyclosporin formulation was more effective than the 0.05% cyclosporin formulation. For one thing, Allergan has not explained why those two endpoints, in particular, are important in determining effectiveness. Even if Schirmer scores and corneal staining were shown to be the two most relevant

136. *Id.* fig.3.
137. *Id.* at 637 (discussing the overall results of the study).
138. *Id.* at 638.

> endpoints, the p-values of 0.834 and 0.112 reveal that the difference observed between the 0.1% and 0.05% formulations for the Schirmer scores may be the result of random chance more than 83% of the time, and on corneal staining, the result of random chance more than 11% of the time. Even for corneal staining, the p-value is more than twice as large as the value that clinicians usually regard as representing a real difference between two means.[139]

The court then addressed the analysis of Dr. Calman, Teva Pharmaceuticals' expert, who looked at all 58 measured categories for the cyclosporin formulations and found that only two of those categories showed a statistically significant difference favoring the 0.01% formulation over the 0.05% formulation, the OSDI score at week 12 and the temporal rose bengal conjunctival staining at week 2.[140] Evaluating that analysis, the court concluded that it did "not find that those two individual points of statistical significance, out of all of the tested categories and time points, [were] sufficient to demonstrate a real difference in effectiveness between the 0.05% and 0.1% cyclosporin formulations."[141] Finally, the court found that Allergan made a crucial mistake by focusing on its Phase 2 trial results on the 0.1% formulation compared to the baseline rather than comparing the two formulations.[142] Accordingly, the court determined that a person "skilled in the art" would not consider the 0.1% formulation to have outperformed the 0.05% formulation in the Phase 2 trials,[143] which thus undercut Allergan's patentability argument.

The court then turned to the results of the Phase 3 trials. Allergan again attempted to cherry pick categories with statistically significant p-values that

---

139. Allergan, Inc. v. Teva Pharms. USA, Inc., No. 2:15-cv-1455-WCB, 2017 WL 4803941, at *25 (E.D. Tex. Oct. 16, 2017) (noting "the Court will not draw a rigid, artificial line at a p-value of 0.05: Results accompanied by a p-value of slightly more than 0.05 would also be likely to inform the reasonable expectations of one of skill in the art"); *see also supra* Part I. P-values do not represent the probability that a result is or is not due to chance, though the court did use that information to reach the correct judgment that the data did not support Allergan's argument.

140. *Id.* at *26.

141. *Id.* (stating further that "[s]ignificantly, during the prosecution of the Restasis patents, Allergan did not rely on either of those two categories as proof of unexpected results").

142. The court wrote:

> [T]he relevant comparisons are those between the 0.05% and 0.1% formulations, which appear in the last column. The only statistically significant difference in that key category appears in one endpoint, the OSDI score, a subjective measurement upon which Allergan did not rely to support its claim of unexpected results before the PTO.

Id. at *27.

143. *Id.*

supported its case to no avail.[144] The court instead focused on the Phase 3 data as a whole.[145] The court noted that, of 80 total data points, only four

144. Discussing Allergan's heavy dependence on the Schirmer tear test scores rather than viewing the results of the study as a whole, the court stated:

> Allergan points out that, for the categorized Schirmer tear test scores with anesthesia, the p-value for the pair-wise comparison between the 0.05% and 0.1% cyclosporin formulations at month 3 was 0.076. While not statistically significant, that p-value approached statistical significance. Even granting some flexibility to the conventional line of "statistical significance" at 0.05, however, the Court is not persuaded that the single cherry-picked data point of categorized Schirmer scores with anesthesia at month 3 demonstrates a real difference in efficacy between the two cyclosporin formulations, for several reasons.
>
> First, at the end of the treatment period (month 6), the categorized Schirmer tear test scores with anesthesia for the 0.05% and 0.1% cyclosporin formulations were both statistically significant as compared to the vehicle ($p < 0.01$ for both pair-wise comparisons to the vehicle). In addition, there was no statistically significant difference between the two cyclosporin formulations at the end of the treatment period; rather, the pair-wise comparison produced a p-value greater than 0.25.
>
> Second, for categorized Schirmer tear test scores without anesthesia, the 0.1% cyclosporin formulation performed better than the 0.05% cyclosporin formulation at all time points, although there was no statistically significant difference between those two formulations. That is the same result as seen in Phase 2 at all of the testing times. ([T]he 0.1% cyclosporin formulation did better than the 0.05% cyclosporin formulation in Phase 2 for categorized Schirmer tear test scores without anesthesia, although pair-wise comparisons of the 0.05% and 0.1% cyclosporin formulations are not available or are not statistically significant). Allergan has not adequately explained why the Court should compare categorized Schirmer scores without anesthesia in the Phase 2 study, to categorized Schirmer scores with anesthesia in the Phase 3 study, when the point is to understand whether the results in Phase 3 were surprising in light of the results in Phase 2. Schirmer tests with anesthesia were not performed in Phase 2, but both types of tests were performed in the Phase 3 studies. Allergan could have conducted, but chose not to conduct, a direct comparison of the categorized Schirmer scores without anesthesia. That direct comparison shows similar results in both phases: The 0.1% cyclosporin formulation performed better than the 0.05% cyclosporin formulation at all time points in both [P]hase 2, and [P]hase 3.
>
> Third, the underlying Phase 3 raw Schirmer scores with anesthesia at month 3 tell a different story than the derived Phase 3 categorized Schirmer scores with anesthesia at month 3.

Id. at *34 (citations omitted); *See also id.* at *8:

> The defendants contend that there was no statistically significant difference between the 0.05% cyclosporin formulation and the 0.1% formulation in the Phase 2 study. The defendants also point out that while both of the cyclosporin formulations did better than the castor-oil-only vehicle formulation in the Phase 3 studies, there was no overall statistically significant difference between the 0.05% cyclosporin formulation and the 0.1% formulation in those trials. Accordingly, the defendants contend, there has been no showing that the 0.05% formulation performed in a way that was unexpected, so as to render the 0.05% formulation patentable . . . .

Id. at *8.

145. *Id.* at *34.

favored the 0.05% formulation with statistical significance.[146] In addition, 71 of the 80 data points showed no statistically significant difference between the two formulations.[147] Concluding its analysis of the *Stevenson Report* and the *Sall Report*, the court noted that there was a lack of evidence showing any real difference between the formulations and the fact that a "person of skill [in the art]" would reach this conclusion when looking at both trials.[148]

More specifically, the court pointed out:

> Stevenson also reported that "[t]here was no clear dose-response relationship" shown in the Phase 2 study between the tested cyclosporin formulations, i.e., the increase in cyclosporin did not result in an increase in clinical efficacy. In a typical dose-response relationship, an increase in the dose of the active ingredient results in an increase in the therapeutic effect of the drug. Therapeutic efficacy may continue to increase until it reaches a plateau, at which point further increases in the amount of the active ingredient no longer result in an increase in therapeutic effect. If there is no dose-response relationship between lower and higher amounts of a drug (such as after the efficacy plateau is reached), then there is no reason to use greater amounts of the drug in an effort to achieve greater therapeutic efficacy. Thus, there would be no motivation to move from a 0.05% cyclosporin formulation to a 0.1% cyclosporin formulation if the higher concentration provided no greater therapeutic effect. Because Stevenson noted the lack of a dose-response relationship between the tested formulations, a person of ordinary skill would not understand Stevenson's paper to suggest using the 0.1% cyclosporin formulation over the 0.05% cyclosporin formulation.[149]

Allergan had further sought to mine appropriate p-values that would demonstrate its desired legal result from Phase 2 data, an attempt the court likewise found lacking:

> It was the Phase 3 studies, not the Phase 2 study, that were intended to determine whether the approved drug should

146. *Id.* at *33.

147. *Id.* at *33 (analyzing the data as a whole and determining that the overwhelming majority of the data supports the notion that the formulations performed similarly and that the 0.05% formulation did not outperform the 0.1% formulation).

148. *Id.* at *36.

149. *Id.* at *21 (alteration in original) (citations omitted).

> contain a 0.05% cyclosporin emulsion or a 0.1% cyclosporin emulsion. Allergan's flawed effort to convert the Phase 2 study into an assessment of the relative efficacy of the 0.05% and 0.1% cyclosporin formulations lies at the heart of the problem with its "unexpected results" analysis.
>
> The Phase 2 study was small and was not designed to reveal statistically significant differences between the various tested formulations.[150]

The court thus rejected efforts to cherry pick p-values on particular parameters from the overall context of both studies, particularly when small sample sizes were present.[151] The court concluded:

> In sum, there is a dearth of evidence showing any real difference between the efficacy of the 0.05% and 0.1% cyclosporin formulations in Phase 2, as presented in Stevenson, and in Phase 3, as presented in Sall. A person of skill reviewing those papers would come to the conclusion that neither formulation was more effective than the other in Phase 2. That person of skill would reach the same conclusion for Phase 3.[152]

Allergan thus failed in its claim that the 0.1% formulation was not obvious, and accordingly lost its patent for that formulation.[153]

B. Vanderwerf v. SmithKline

Recall from Part I that the dispute in *Vanderwerf* centered on whether SmithKline's prescription drug Paxil led to the suicide of plaintiffs' husband and father, William Vanderwerf.[154] The central study in the case was a Food

---

150. *Id.* at *23.

151. *Id.* at *33.

> The Court does not find that the few data points [in the Phase 3 study] reflecting statistical significance demonstrate a real difference in effectiveness between the 0.05% and 0.1% cyclosporin formulations. More specifically, the Court does not find that the four individual data points (at most) that showed statistical significance in favor of the 0.05% cyclosporin formulation indicate a real difference in effectiveness favoring the 0.05% over the 0.1% cyclosporin formulation, as Allergan contends.

Id.

152. *Id.* at *36.

153. *Id.* at *65.

154. *See supra* Part I.B and accompanying footnotes.

and Drug Administration (FDA) report entitled *Clinical Review: Relationship Between Antidepressant Drugs and Suicidality in Adults*.[155] Paxil was one of eleven antidepressant drugs measured for suicidality against the placebo in that study.[156] The FDA requested eight manufacturers of these drugs to submit data from over 400 clinical trials of over 100,000 subjects.[157] The manufacturers were instructed to search adverse events for keywords, including "attempt," "gun," and "jump," and classify those events into one of the categories in the table below.[158]

Table 3: Coding of suicide-related events within the suicidality datasets

Event	Coding
Completed suicide	1
Suicide attempt	2
Preparatory acts toward imminent suicidal behavior	3
Suicidal ideation	4
Self-injurious behavior, intent unknown	5
Not enough information (Fatal)	6
Not enough information (Non-Fatal)	9

Figure 6. Data from Stone & Jones, *Clinical Review: Relationship Between Antidepressant Drugs and Suicidality in Adults* (2006).[159]

Outcomes 1–4 in the table above fell under "suicidal ideation," the report's primary outcome. Outcomes 1–3 fell under "preparatory actions or worse," the report's secondary outcome.[160] Below, Figure 7 shows the sample distribution among the antidepressant drugs, with paroxetine (Paxil), the drug at issue in the *Vanderwerf* litigation, highlighted.

155. *See* STONE & JONES, *supra* note 58.

156. *Id.* at 8 (listing Wellbutrin, Celexa, Cymbalta, Lexapro, Prozac, Symbyax, Luvox, Remeron, Serzone, Paxil, Zoloft, and Effexor as the tested drugs).

157. *Id.* at 11.

158. *Id.* at 11, 13.

159. *Id.* at 13 tbl.3.

160. *Id.* at 13.

Table 7: Numbers of subjects by drug, drug class and treatment assignment

Drug	Primary	Active Control	Placebo
SSRI			
Citalopram	1,928	733	1,371
Escitalopram	2,567	563	2,604
Fluoxetine	9,070	2,418	7,645
Fluvoxamine	2,187	0	1,828
Paroxetine	8,728	1,223	7,005
Sertraline	5,821	1,129	5,589
SNRI			
Duloxetine	6,361	0	4,172
Venlafaxine	5,693	129	4,054
Other Modern Antidepressants			
Bupropion	6,018	0	3,887
Mirtazapine	1,268	0	726
Nefazodone	3,319	0	2,173
Tricyclic Antidepressants			
Amitriptyline	0	625	627
Clomipramine	0	632	617
Desipramine	0	315	298
Dothiepin	0	106	95
Imipramine	0	2,345	2,304
Other Antidepressants			
Mianserin	0	28	28
Trazodone	0	121	125
All Drugs	**52,960**	**10,367**	**35,904**

The median number of subjects per trial assigned to the primary drug was 109.5 while the median number of placebo subjects was 89. When a trial contained an active control arm the median number of subjects assigned to the active control was 88.5. A summary of demographic information is given in Table 8.

Figure 7. Data from Stone & Jones, *Clinical Review: Relationship Between Antidepressant Drugs and Suicidality in Adults* (2006).[161]

The null hypothesis for the study was that there was no difference in the primary outcome (suicidal ideation or worse) between these drugs and placebo, with the alternative hypothesis that there was a difference.[162] In simpler terms, a finding of statistical significance would lend credence to the assertion that the drugs led to increased risk of suicide. Abbreviated tables reporting each drug's results on the primary and secondary outcomes are included below.[163] Note paroxetine's statistically significant results in the secondary outcome (preparatory actions or worse), but not for the primary outcome (suicidal ideation or worse).

161. *Id.* at 18 tbl.7.
162. *Id.* at 14.
163. *Id.* at 24 tbl.15, 26 tbl.16.

Table 15: Suicidality Risk for Active Drug relative to Placebo – Ideation or Worse – Adults with Psychiatric Disorders – By Drug and Drug Class

Drug Class Drug	Odds Ratio	95% Confidence Interval	p value
All Drugs	**0.83**	**0.69 - 1.00**	**0.05**
SSRI	**0.86**	**0.69 - 1.06**	**0.16**
Citalopram	2.11	0.90 - 4.94	0.08
Escitalopram	2.44	0.90 - 6.63	0.08
Fluoxetine	0.71	0.52 - 0.99	0.04
Fluvoxamine	1.25	0.66 - 2.39	0.49
Paroxetine	0.93	0.62 - 1.42	0.75
Sertraline	0.51	0.29 - 0.91	0.02

Figure 8. Data from Stone & Jones, *Clinical Review: Relationship Between Antidepressant Drugs and Suicidality in Adults* (2006).[164]

Table 16: Suicidal Behavior Risk for Active Drug relative to Placebo – Preparation or Worse – Adults with Psychiatric Disorders – By Drug and Drug Class

Drug Class Drug	Odds Ratio	95% Confidence Interval	p value
All Drugs	**1.10**	**0.77 - 1.56**	**0.60**
SSRI	**1.23**	**0.82 - 1.85**	**0.31**
Citalopram	1.97	0.56 – 7.00	0.29
Escitalopram	5.67	0.94 – 34.2	0.06
Fluoxetine	1.08	0.52 – 2.23	0.83
Fluvoxamine	1.31	0.51 – 3.38	0.58
Paroxetine	2.76	1.16 – 6.60	0.02
Sertraline	0.25	0.07 – 0.90	0.03

Figure 9. Data from Stone & Jones, *Clinical Review: Relationship Between Antidepressant Drugs and Suicidality in Adults* (2006).[165]

164. *Id.* at 24 tbl.15.
165. *Id.* at 26 tbl.16.

The FDA determined that the null hypothesis was supported by this data, the hypothesis being that the drugs did not cause an increased risk of suicide.[166] The FDA further considered any association between antidepressant treatment and an increased risk of suicidality paradoxical.[167] Among other reasons, even seemingly significant p-values—like that for Paxil—had to be discounted due to the large number of comparisons being made in the studies[168] because "if enough comparisons are made, random error almost guarantees that some will yield significant findings, even when no real effect" is present.[169] This phenomenon was explained further in the court's opinion, where the plaintiffs sought a different result:

> Plaintiffs also argue that the FDA has found a statistically significant increase in risk for suicidal behavior in adult Paxil patients with all psychiatric disorders compared to placebo. Again, plaintiffs have taken a single FDA finding out of context. As noted above, as to suicidal behavior defined as suicidal ideation or worse (the study's primary endpoint), the FDA did not find an increased risk for Paxil patients compared to placebo. The primary analysis therefore showed no association between Paxil and suicidal thinking (or worse) in adults. As to suicidal behavior defined as preparatory acts or worse (the study's secondary endpoint), the FDA found a statistically significant increased risk for adult Paxil patients with all psychiatric disorders versus placebo (relative risk of 2.76; 95% confidence interval of 1.16 to 6.60; P value of 0.02). The FDA, however, cautioned that "[a]lthough the values for some individual drugs are statistically significant at 0.05 level, the significance of those findings must be discounted for the large number of comparisons being made." In addition, the FDA specifically rejected any association between suicidality or suicidal behavior in adults age 25 or older. Even now, the FDA rejects the notion of a causal link between Paxil and suicide or suicidal behavior in adults beyond the age of 24. As with the testimony of Dr. Kraus, the FDA limited statistical finding on Paxil in its

166. *Id.* at 23.
167. *Id.*
168. *Id.*
169. Vanderwerf v. SmithKline Beecham Corp., 529 F. Supp. 2d 1294, 1308 (D. Kan. 2008).

> 2006 Clinical Review does not create a genuine issue of material fact on the issue of general causation.[170]

The *Vanderwerf* plaintiffs additionally sought to establish causation through a data point allegedly representing "a statistically significant increase in the frequency of suicidal behavior (including preparing for suicide, suicide attempts and completed suicides) in patients treated with Paxil compared to placebo."[171] The *Vanderwerf* court rejected that attempt, finding that data point inappropriately combined information from study groups of 18–30 year-olds and 25–64 year-olds, and merely:

> [D]iscloses a "possible" risk in adult patients, states that the risk is likely limited to younger adults between the ages of 18 and 30, and emphasizes that it is difficult to conclude a causal relationship because of (1) the small incidence and absolute number of events, (2) the retrospective nature of the metaanalysis [sic] and (3) the fact that the risk of suicidal behavior is a symptom of the underlying psychiatric illnesses. . . . Therefore, even giving plaintiffs the benefit of all favorable inferences, GSK [the manufacturer of Paxil] has at most admitted that Paxil *may* increase the risk of suicidal behavior and suicide in adult patients between the ages of 18 and 30. Dr. Kraus' [sic] testimony, viewed in its entirety, does not create a genuine issue of material fact whether Paxil causes suicidal behavior and/or suicide in adult patients beyond the age of 30, let alone constitute an "admission" that Paxil does so. Even if Paxil increases the risk of suicide below age 30 and decreases the risk of suicide above age 65, plaintiffs have presented no evidence which would assist a jury in determining whether Paxil more likely than not caused suicide in 36-year-old individuals in general or in Mr. Vanderwerf, in particular. On this record, even given Dr. Kraus' [sic] testimony, such a conclusion would be sheer speculation.[172]

Based on this analysis, the court found for the defendants on summary judgment.[173] The court stated that despite the seemingly relevant p-values identified in the FDA report, these values were too isolated to possess any

170. *Id.* at 1308–09 (alteration in original) (footnote omitted) (citations omitted).
171. *Id.* at 1307.
172. *Id.* at 1307–08.
173. *Id.* at 1309.

probative value.[174] Thus, no genuine issue of material fact on general causation (whether Paxil could cause suicidality in anyone) could be found.[175]

C. Zeneca v. Eli Lilly

Recall from Part I that Zeneca sued Eli Lilly for unfair competition and deceptive trade practices for marketing Evista (raloxifene) as a viable treatment for reducing the incidence of breast cancer.[176] Zeneca brought this claim because its own drug, Nolvadex (tamoxifen), was clinically proven to do the same and was suffering from Eli Lilly's marketing of Evista.[177]

The primary study examined in the case was the *Multiple Outcomes of Raloxifene Evaluation* (MORE).[178] The MORE trial was conducted from 1994–1998 in clinical centers primarily in the United States and Europe.[179] The trial was designed to determine whether three years of treatment with Evista reduced the risk of bone fracture in postmenopausal women with osteoporosis.[180] Whether Evista was effective in reducing the risk of breast cancer was a secondary endpoint of the trial.[181]

Of the 5,129 women given raloxifene, 13 cases of breast cancer were reported, with 27 cases reported among the 2,576 women given the placebo.[182] The primary explanation provided for this circumstance was the inversely proportional relationship between osteoporosis and breast cancer.[183] As the report explained, estrogen was believed to play a central role in the pathogenesis of breast cancer while decreased bone density (a major cause of osteoporosis) can serve as a marker of lower lifetime exposure to estrogen.[184] The graph below illustrates the small percentage of the study's sample that included patients with breast cancer.[185]

---

174. *Id.*

175. *Id.*

176. *See* Zeneca Inc. v. Eli Lilly & Co., No. 99 CIV. 1452(JGK), 1999 WL 509471, at *1 (S.D.N.Y. July 19, 1999).

177. *Id.* at *5–6 (discussing the immense success of Nolvadex in the Breast Cancer Prevention Trial (BCPT)).

178. *Id.* at *7.

179. *Id.*

180. *Id.*

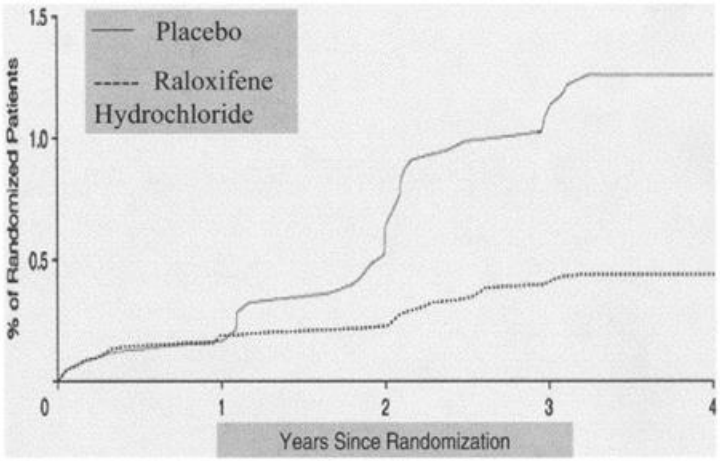
181. *See* Maura N. Dickler & Larry Norton, *The MORE Trial: Multiple Outcomes for Raloxifene Evaluation*, 949 ANNALS N.Y. ACAD. SCIS. 134, 136 (2001) (stating "[t]he trial was also designed to prospectively evaluate whether raloxifene reduces the risk of breast cancer").

182. Steven R. Cummings et al., *The Effect of Raloxifene on Risk of Breast Cancer in Postmenopausal Women Results from the MORE Randomized Trial*, 281 JAMA 2189, 2192 (1999).

183. *Id.* at 2194–95.

184. Dickler & Norton, *supra* note 181, at 135–36.

185. *Id.* at 137.



Cumulative incidence of breast cancer in study participants, represented as a percentage of all patients randomized to either group ($P < 0.001$). Reproduced with permission from Ref. 4.

Figure 10. Data from Dickler & Norton, *The MORE Trial: Multiple Outcomes for Raloxifene Evaluation* (2001).[186]

Evista showed a statistically significant improvement ($p < 0.001$) in those 13 patients compared to the placebo in reducing the risk of breast cancer. Despite the small sample size, the MORE report found this data to be reliable.[187] This determination was expressly limited both by the small sample size and by the fact that the sample only included women with osteoporosis.[188]

The results of the MORE trial led to another study of breast cancer prevention focusing on Evista and Nolvadex, called the *Study of Tamoxifen and Raloxifene* (STAR).[189] The STAR trial was intended to produce results in 2006, so with the litigation beginning in 1998, Eli Lilly could only rely on the MORE report.[190]

186. *Id.* fig.1.

187. *Id.* at 136 (stating that "[i]t can therefore be stated with confidence that, in cancer-free women with osteoporosis, the use of raloxifene reduces the incidence of breast cancer, at least over the short term").

188. Small sample sizes can skew data by making limited occurrences seem more common than actually is the case, and even common occurrences may not show up at all in a study with a small sample size. *See supra* Part III.B.

189. Dickler & Norton, *supra* note 181, at 138.

190. The STAR report was finalized in 2010. *See The Study of Tamoxifen and Raloxifene (STAR): Questions and Answers*, NAT'L. CANCER INST., https://www.cancer.gov/types/breast/research/star-trial-

The FDA did not agree with the MORE report's conclusion that Evista could be marketed for reducing the risk of breast cancer.[191] Specifically, the FDA was concerned because reduction of breast cancer was only a secondary endpoint of the MORE study,[192] and thus only 40 women in the study developed breast cancer.

After conducting a preliminary injunction hearing, the court found the Lanham Act claims were likely to succeed on the merits:

> Based on all the evidence adduced, Zeneca and Barr will likely succeed in proving that the MORE trial is "not sufficiently reliable to permit one to conclude with reasonable certainty that [it] established the proposition for which [it was] cited"—namely, that Evista has been proven to reduce the risk of breast cancer.
>
> The FDA, as well as numerous other experts in the field of clinical oncology, have reviewed the breast cancer data from the MORE trial and reached the nearly unanimous conclusion that it does not prove that Evista reduces the incidence of breast cancer. The reasons for the unanimity of these organizations are described at length in the Findings of Fact. Most notably, the MORE protocol was not designed to determine whether Evista could be efficacious in reducing the risk of breast cancer. Accordingly, women were not selected for enrollment and once enrolled were not randomized between the raloxifene and placebo arms based on their degree of breast cancer risk. The protocol also did not require annual mammograms or breast physical exams, among other diagnostic deficiencies. Because of these and other critical flaws, the risk factors may have been imbalanced, the incidence of breast cancer may have been underdiagnosed and the results may yet turn out—as the CORE extension goes forward—to be a "false positive."

---

results-qa# (last updated Apr. 19, 2010). In that report, Eli Lilly's Evista was found to be 76% as effective as Zeneca's Nolvadex in reducing the risk of breast cancer over nearly seven years of measurement. *Id.* Based on these results, the National Cancer Institute did not recommend use of either drug to reduce the risk of breast cancer. *Id.* ("Even if a woman is at increased risk of breast cancer, raloxifene or tamoxifen therapy may not be right for her.").

191. *See* Zeneca Inc. v. Eli Lilly & Co., No. 99 CIV. 1452(JGK), 1999 WL 509471, at *18 (S.D.N.Y. July 19, 1999) ("[T]he FDA has repeatedly determined and communicated to Eli Lilly in Meeting Minutes that the MORE study does not and cannot prove that Evista reduces the risk of breast cancer.").

192. *Id.* (discussing the MORE trial's design limitations with respect to analyzing Evista's efficacy for reducing the risk of breast cancer).

> Given the small number of invasive breast cancers that were diagnosed, a small number of additional invasive breast cancers in the raloxifene arm would have seriously compromised the results of the study.[193]

The court further evaluated Eli Lilly's contention that the MORE study should be found probative with respect to breast cancer, explaining why its small sample size doomed any reliable result:

> Zeneca offered the expert testimony of Dr. Mark Scott, a biostatistician with extensive experience in the design and analysis of clinical drug trials. Dr. Scott responded to Eli Lilly's argument that the many flaws in the MORE trial may be overlooked because the breast cancer results in MORE were statistically significant at a level of p = .000005. A p value [sic] of less than .05 typically is required for a finding of statistical significance in a clinical trial.
>
> Dr. Scott explained that because breast cancer risk reduction was not the primary endpoint of the MORE trial, and there was no pre-specified statistical plan for analyzing breast cancer data, it is inappropriate to use a p value [sic] of .05 as a benchmark to assess the statistical significance of the MORE breast cancer data. Rather, to ensure that the results in question were not due to chance, Dr. Scott opined that the appropriate p value [sic] should be adjusted to take into account the fact that breast cancer risk reduction was a secondary endpoint and just one of hundreds of statistical tests performed in the MORE trial.
>
> Dr. Scott made that adjustment, using the well-established formula, acknowledged by Eli Lilly's witnesses, of dividing the p value [sic] of .05 by the number of tests conducted. According to the lead MORE investigator, Eli Lilly's expert Dr. Steven Cummings, 400 safety tests alone were conducted in MORE. Using that number, which did not even take into account the non-safety statistical analyses performed on the MORE data, Dr. Scott concluded

193. *Id.* at *33 (alteration in original) (citation omitted).

> that the appropriate p value [sic] to determine statistical significance was .000125.
>
> Although the MORE trial's breast cancer results still achieved statistical significance using that figure, Dr. Scott's testimony illustrated the significance of the fact that the MORE trial had relatively few cases of breast cancer. As previously noted, in MORE there were 40 total cases of invasive breast cancer; this compares with 264 cases in the BCPT. Dr. Scott explained that, given the low number of overall breast cancer cases reported to date in the MORE trial, the addition of only five more cases on the raloxifene arm of the study—without a corresponding increase on the placebo arm—would render the result on which Eli Lilly now relies statistically insignificant. Dr. Scott testified to a number of hypothetical scenarios under which those five additional cancers on the raloxifene arm could occur.[194]

With respect to the FDA, the court explained:

> The fact that the FDA has not approved raloxifene for breast cancer risk reduction does not conclusively demonstrate that the defendant's claim that raloxifene has been proven to reduce the risk of breast cancer is literally false under the Lanham Act because "a Lanham Act plaintiff must prove that the defendant's efficacy claims are literally false, not simply that they fail to meet current federal licensing standards." However, as a recognized expert in evaluating data from clinical trials, the FDA's conclusion as reflected in the Evista label and various FDA documents that "[t]he effectiveness of raloxifene in reducing the risk of breast cancer has not yet been established" is persuasive evidence that Eli Lilly's claims to the contrary are untrue.[195]

Eli Lilly's marketing of Evista thus was found to constitute unfair competition, and Zeneca's claims under the Lanham Act and state law were upheld.[196]

Each of these cases demonstrates challenges in connecting statistical analysis to the legal claims at issue. *Allergan* illustrates the danger of

194. *Id.* at *25–26 (footnote omitted) (citations omitted).
195. *Id.* at *34 (alteration in original) (citation omitted).
196. *Id.* at *43.

interpreting relevant expert reports too selectively and then incorrectly assuming that a few favorable p-values would carry the day.[197] The court rejected "cherry picking" of the data in this manner, particularly in the context of small sample sizes in the Phase 2 study reported by Stevenson.[198]

Similarly, *Vanderwerf* demonstrates an effort to select particular p-values favoring one side of the litigation, and despite the FDA's data-dredging-like study, *Vanderwerf* sought a specific outcome rather than waiting to see what the dredging might uncover.[199] As that court explained: "[R]epeated testing complicates interpretation of significance levels; if enough comparisons are made, random error almost guarantees that some will yield significant findings, even when [there is] no real effect."[200] As already explained, multiplicity issues like those present in *Vanderwerf* can be addressed through use of statistical adjustment methodologies.[201]

Finally, the *Zeneca* case illustrates the difficulty in bending identified statistical significance to support an argument when only a small sample size supports that position. Eli Lilly first presented a legitimate argument based on p-values, arguing that the flaws of the MORE trial could be ignored because the breast cancer benefits were statistically significant at $p = 0.000005$ and—using a proper multiple comparison adjustment—even at $p = 0.000125$.[202] But focusing on the small number of cases of breast cancer present in a study not designed to address that outcome, the court nonetheless rejected the MORE data as "not the stuff of proof" for purposes of the unfair competition claims.[203]

197. *Id.* at *23–24, *31–42.

198. *See supra* notes 140–42 and accompanying text (explaining the *Allergan* court's rejection of the single OSDI result and related analysis).

199. *See supra* notes 69–71 and accompanying text (explaining data dredging, cherry picking, and p-hacking).

200. Vanderwerf v. SmithKline Beecham Corp., 529 F. Supp. 2d 1294, 1308 (D. Kan. 2008) (citing Paul C. Giannelli et al., *Reference Guide on Forensic Identification Expertise*, *in* REFERENCE MANUAL 55, 127–28 (3d ed. 2001)).

201. *See supra* note 80 and accompanying text (explaining multiplicity adjustment through the Benjamini-Hochberg and/or Bonferroni correction protocols).

202. *See* Yoav Benjamini & Yosef Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, 57 J. ROYAL STAT. SOC'Y SERIES B (METHODOLOGICAL) 289, 291 (1995); *see also* Zeneca, Inc. v. Eli Lilly & Co., No. 99 CIV. 1452(JGK), 1999 WL 509471, at *25–26 (S.D.N.Y. July 19, 1999).

203. *Zeneca*, *Inc.*, 1999 WL 509471, at *23 (quoting Transcript of Dr. Mark Scott at 1167, Zeneca, Inc. v. Eli Lilly & Co., No. 99 CIV. 1452(JGK), 1999 WL 509471, at 1 (S.D.N.Y. July 19, 1999)); *see also supra* notes 60–64 and accompanying text (explaining court concerns about small sample size, particularly with respect to a secondary endpoint of the study).

IV. COMMUNICATING WITH THE DECISION MAKER

In each of these cases, the losing parties misinterpreted p-values and the methods by which they were produced. When the math mattered so much, clearer understanding by the lawyers as to what the p-values and underlying data meant, and how that data related to the legal claims, may have resulted in better advocacy and a more favorable result for the client.[204]

Bridging the gap between the legal and scientific communities, in pharmaceutical cases and otherwise, is not straightforward. Harold Green argues the gap between the two communities is not just knowledge based, but also approach based: lawyers focus on optimal outcomes for their clients, regardless of whether the outcome is "correct," while scientists are more likely to be seen—at least absent use of inappropriate techniques like those discussed in this Article—to be attempting to focus on progress and truth.[205] Green explores this distinction in the litigation context through *Wells v. Ortho Pharmaceutical Corporation*.[206] There, the U.S. District Court for the Northern District of Georgia awarded the plaintiffs $5.1 million in damages for birth defects related to use of a spermicide.[207] The judge stressed that the court's duty was to make a legal decision, not a medical one, by weighing the evidence presented by the parties.[208] The court's ultimate ruling for the plaintiff led to outcry in the scientific community, as Green explained:

> The usually cautious *New York Times* dealt with the case in an emotional editorial entitled *Federal Judges vs. Science*. The editorial referred to an assertion by the *New England Journal of Medicine* that the courts would no longer be bound by reasonable standards of scientific proof and went on to spell out the reasons why it regarded the descision [sic] in *Wells* as grossly erroneous. According to the *Times*, there was "no serious difference among experts" as to the safety of the product because "after reviewing some 20 epidemiological studies, an expert committee advised the Food and Drug Administration in 1983 that the preponderance of available evidence 'indicates no

204. *See supra* Parts III.A–C.

205. *See* Harold P. Green, *The Law-Science Interface in Public Policy Decisionmaking*, 51 OHIO. ST. L.J. 375, 387–88 (1990) ("[T]he scientist's vision of progress is also closely related to [the] truth. . . . The law, on the other hand, is more concerned with the optimal resolution of disputes than it is with achieving 'correct' decisions that accord with objective truth.").

206. *See* Wells v. Ortho Pharm. Corp., 615 F. Supp. 262, 292 (N.D. Ga. 1985).

207. *See* Green, *supra* note 205, at 388–89.

208. *Id.* at 389 (citing *Wells*, 615 F. Supp. at 266).

> association' between spermicides and birth defects." The editorial complained that Judge Shoob, trying the case without a jury, rejected the written evidence of the scientific literature and focused on the oral testimony presented, paying "close attention to each expert's demeanor and tone." He chose to believe the plaintiff's main witnesses, three pharmacologists and an expert in birth defects, "none of whom had any expertise in epidemiology," which the *Times* characterized as "the science of determining the causes of disease." According to the *Times*, "science's finest achievement is finding methods to raise objective evidence above the merely anecdotal," but Judge Shoob was not moved by the preponderance of the scientific evidence, nor was the Court of Appeals, which "espoused the fiction that there had been a battle of experts, even though no scientist would consider pharmacologists expert[s] in a matter of epidemiology." According to the *Times*, the Eleventh Circuit rejected scientific standards of evidence when it upheld Judge Shoob's decision because there was "sufficient evidence of causation in the legal sense in this particular case, and that . . . finding is not clearly erroneous." The *Times* labelled Judge Shoob's and the Court of Appeals' position an "intellectual embarrassment" that could have profound practical consequences in driving spermicides off the market and further narrowing contraceptive choice for women.[209]

Lawyers seeking to better present difficult mathematical issues to judges and juries should delve into their communication processes, breaking down both the math and the law to best use both in their advocacy.

A. Recognize the Differing Roles of Science and Law

Writing for the Journal of the American Medical Association, Daniel Merenstein details a medical malpractice case that shows the disconnect between the legal and scientific community.[210] In 1999, Merenstein had a patient with whom he discussed the risks and benefits of screening for prostate cancer (PSA screening).[211] The patient was never

209. *Id.* at 390 (footnotes omitted).

210. *See* Daniel Merenstein, *Winners and Losers*, 291 JAMA 15, 15 (2004) (describing patient treatment where medical science supported one approach, but a legal verdict contrary to that science nonetheless was awarded).

211. *Id.*

screened by Merenstein, but went to another doctor.[212] The new doctor did not discuss the risks and benefits but instead ordered the screening, finding Gleason 8 cancer, "a horrible cancer . . . that is very difficult to treat in any stage . . . ."[213] The patient subsequently brought a medical malpractice suit against Merenstein and his practice for not ordering the PSA screening.[214] Merenstein's defense was that the consensus in the medical community was for a physician and a patient together to make the call on whether to do a PSA screening due to its questionable benefits and risks.[215] While Merenstein might have had the weight of his profession behind him, the plaintiff's lawyers stuck to the law to get their optimal outcome, arguing that Merenstein did not practice the standard of care required by Virginia (which they argued was different from the opinion of the broader medical community) and that Merenstein's evidence-based medicine was a "cost-saving method."[216] Merenstein stresses that his focus was with progress and the truth, while the plaintiff's lawyers argued for their client, regardless of the recommendations of the medical community.[217] Unfortunately for Merenstein, his method lost—a jury found him liable for one million dollars.[218]

Green provides context for this sort of outcome, explaining that:

> It is essential to bear in mind . . . that that there is a substantial difference between the skills and data necessary to make determinations concerning safety and those required to make determinations concerning the cause of a particular malady.
>
> . . . .The [legal] system [has] the duty to decide, in accordance with established legal procedures and on the basis of all the evidence presented by the parties, whether use of the spermicide [or any other action at issue] caused injury in the particular case before the court. Whereas science can duck issues of this kind by asserting that the evidence is inconclusive, a court does not have this luxury. When a lawsuit is filed, the case must be decided in a binary

212. *Id.*

213. *Id.*

214. *Id.*

215. *Id.* ("Our experts explained that because of the questionable benefit vs associated risks of PSA screening, a shared decision by the physician and the patient was recommended by all of the national health associations. The science was clearly in our favor.").

216. *Id.* at 15–16.

217. *Id.*

218. *Id.* at 16.

> manner: liability or no liability. Moreover, in the Anglo-American legal system, responsibility for managing the case rests with the judge, who probably has no scientific competence; and responsibility for actually deciding the issue of causation rests with a jury of lay persons.[219]

Recognizing this dichotomy, it is perfectly understandable for a lawyer to focus—like the lawyers for Dr. Merenstein's patient—on a legal process outcome that deflected attention from the "pure" science and resulted in a win for that plaintiff. Commentary on tobacco litigation provides another example. Micah Berman and Annice Kim note that the discovery of smoking being linked to lung cancer and the subsequent response by attorneys and policymakers, led to greatly reduced smoking rates in the 20th century.[220] Building on Harold Green's analysis of the disconnect between the legal and scientific communities, Berman and Kim argue that scientists have little incentive to see their research lead to new policies, while attorneys want to use science to advance their policies without the patience to wait for new evidence.[221] Their discussion focuses on *R.J. Reynolds v. FDA*, where the D.C. Circuit struck down the Food and Drug Administration's (FDA) proposal for graphic health warnings to be placed on cigarette packaging and advertisements on First Amendment grounds.[222] Berman and Kim thus argue that, despite considerable evidence supporting the proposal, failure to anticipate answers to legal and doctrinal questions the court was likely to ask led to defeat.[223]

B. Hone the Analysis

Recognizing its inherent challenges, science communication in any context has been analogized to mountain-climbing[224] and defined as "the use of appropriate skills, media, activities, and dialogue to produce" awareness, enjoyment, interest, opinions, or understanding of a scientific topic.[225] Commenters further argue that effective science communication can create any of these responses for the individuals involved, including students, the

219. Green, *supra* note 205, at 391–92.

220. *See* Micah L. Berman & Annice E. Kim, *Bridging the Gap Between Science and Law: The Example of Tobacco Regulatory Science*, 43 J.L. MED. & ETHICS 95, 95 (2015).

221. *Id.*

222. *Id.* at 96.

223. *Id.*

224. *See* T. W. Burns et al., *Science Communication: A Contemporary Definition*, 12 PUB. UNDERSTANDING OF SCI. 183, 193 (2003).

225. *Id.* at 191.

public, members of the government, and the science practitioners starting the communication.[226]

Though developing understanding through science communication requires substantial effort,[227] Green's assertion that the different approaches used by the legal and scientific communities create a nearly unbridgeable gap is unnecessarily bleak.[228] Certainly there is some credence to the argument that lawyers focus more on optimal outcomes for the client while scientists are more likely to focus on progress and "correct" answers. However, if a client's optimal outcome is consistent with the truth, which the legal system is purported to achieve, then there is a smaller gap between the two communities than Green theorizes.

Because p-values are complicated, as evidenced by confusion on the subject even within the scientific community, lawyers seeking to explain how that analysis contributes to their case are best served to simplify. At bottom, perhaps just four things about p-values need to be understood to make a difference in cases like those that have been explored in this Article. First, a p-value is the probability of observing a result equal to or more extreme than the observed result, given that the null hypothesis is true and assuming all error is random.[229] Second, the typical threshold value of a p-value to be potentially worthy of consideration is 0.05, but that is not a "magic number," and courts have found probative p-values that do not meet the 0.05 standard.[230] Third, picking isolated p-values out of a larger data set is likely to be scrutinized, and understanding the context for all of the data is key.[231] Fourth, small sample sizes—particularly when those small samples are not related to the primary question being studied—are often suspect, but the meaning of "small" can be debated and the various effects caused by small sample sizes are not well understood even in the scientific community.[232] Supplementing these key points with deeper dives into additional relevant information and then coming back to connect these threads to the groundwork already laid is most likely to lead to a successful communication process. Further elements of that communication process follow.

226. *Id.* at 196.
227. *Id.* at 194.
228. *See* Green *supra* note 205, at 378–80.
229. *See supra* notes 6–11, 17–20, 25 and accompanying text.
230. *See supra* notes 11–15 and accompanying text.
231. *See supra* Parts II–III and accompanying text.
232. *See supra* notes 88–90 and accompanying text.

C. Proceed in Small Steps

Legal and mathematical/scientific concepts share the burden of complexity.[233] Everything to be communicated is intellectually difficult to understand. In these circumstances, simplicity of communication is critical, and can be broken down into four critical steps.

First, prepare the audience. An advocate seeking to communicate about difficult concepts like p-values should anticipate issues likely to drive the decision maker and be ready to address them at the outset. Michael Alley provides a relevant example from nuclear weapons policy:

> Contrast the failed one-on-one presentation of Niels Bohr with Winston Churchill in 1944 with the surprisingly successful one-on-one presentation of Edward Teller with President Reagan in 1982. In Bohr's meeting with Churchill, his purpose was to have Churchill realize the potential nuclear weapons race that Bohr anticipated would follow the Second World War. However, Churchill, already defensive about his decision to relinquish intellectual rights to nuclear weapons, ended the meeting after only twenty minutes and asked Bohr to leave. The purpose of Teller's meeting with Reagan was to persuade him to change the United States nuclear weapons policy of mutually assured destruction to a policy of a strategic defense initiative [colloquially termed "Star Wars" at the time]. Given the resistance in the military to such a change and doubts by other scientists such as Hans Bethe as to the potential of the initiative, such a goal seemed out of reach. However, the receptiveness of Reagan and some of his advisors to an alternative to mutually assured destruction proved to be an ally for Teller. The result of that meeting and a later meeting between Teller and one of Reagan's advisors led to the dramatic shift in nuclear weapons policy in March 1983.[234]

A lawyer who recognizes questions and biases that a judge may have developed from prior cases and is ready to address them at the outset—as Teller did, but Bohr and smoking advocates in the D.C. Circuit did not—is more likely to be successful.

---

233. *See, e.g.*, *infra* notes 238–47 and accompanying text.

234. MICHAEL ALLEY, THE CRAFT OF SCIENTIFIC PRESENTATIONS: CRITICAL STEPS TO SUCCEED AND CRITICAL ERRORS TO AVOID 77 (2003).

More specifically, as noted above, failure to anticipate doctrinal questions likely to be raised led to a disconnect between the reasoning processes of science and of law, and ultimate defeat for anti-smoking advocates in *R.J. Reynolds v. FDA*.[235] Specifically, scientists criticized the court for "fail[ing] to comprehend the difficulty of establishing causation in real-world settings, where the influence of graphic warnings cannot possibly be disentangled from the impact of other tobacco control policies and the general decline in tobacco use," when the D.C. Circuit sought "evidence that the graphic health warnings 'directly caused' smoking rates to fall."[236] Had FDA lawyers better anticipated those causation questions, a stronger regulatory record and subsequent judicial advocacy might have resulted.

Second, begin with familiar concepts and connect new ideas to them. This strategy is founded in psycholinguistics, where research has shown that known information (the "given") is typically expressed before previously unknown information (the "new"), in what has been termed a "contract" between the reader and the writer.[237] A similar "shallow-deep" approach has been suggested by scientific commentators to convey complex topics to audiences of mixed expertise:

> [B]egin at a shallow depth that orients everyone in the room to the subject. That orientation includes showing (not just telling) the importance of the subject. Then for each division of the presentation's middle, before diving into the new topic, you begin in the shallows where everyone in the room can follow you. During the deeper dives, many members of the nontechnical and general technical audience will not be able to stay with you, but you should bring them back into the presentation with the beginning of the next topic. At the presentation's end, you should come back to the shallows and then examine the results in a way that everyone

235. *See* R.J. Reynolds Tobacco Co. v. FDA, 696 F.3d 1205 (D.C. Cir. 2012).

236. *See* Berman & Kim, *supra* note 220, at 96.

237. TREVOR A. HARLEY, THE PSYCHOLOGY OF LANGUAGE: FROM DATA TO THEORY 378 (3d ed. 2008). Harley explains:

> One of the most important factors that determines comprehensibility and coherence is the order in which new information is presented relative to what we know already. . . . It takes less time to understand a new sentence when it explicitly contains some of the same ideas as an earlier sentence than when the relation between the content of the sentences has to be inferred.

Id. This concept directly ties to the concept of "bridging" taught in many legal writing classrooms. *See, e.g.*, BRYAN A. GARNER, LEGAL WRITING IN PLAIN ENGLISH: A TEXT WITH EXERCISES 67 (1st ed. 2001) (discussing, in particular, "echo links," which are "words or phrases in which a previously mentioned idea reverberates").

> understands. With this strategy, while the nontechnical and general technical audiences may not have followed all of the theoretical derivations or the analysis of the experimental results in the middle, everyone would have learned the main points of the presentation.[238]

Bryan Garner provides a simple example of bridging in his legal writing text, highlighting repeated use of "pragmatic" and "pragmatism": "[T]he succession of theories on a given topic need not produce a linear growth in scientific knowledge. Science in the pragmatic view is a social enterprise. Th[is] spirit of pragmatism is not limited to [a] handful of philosophers . . . ."[239] A broader example of "orienting" language on scientific principles can be found in *Allergan*.[240] Before delving into detailed merits issues, the court provided an explanation of how p-values are used to show statistical significance before discussing the specific p-values at issue, and whether those values show effectiveness—and thus, in the context of the litigation, patentability—of the particular dry-eye drug formulation.[241]

Third, balance precision with clarity. Precision is important in connecting mathematical and scientific concepts to legal issues, but clarity—understanding from the judge or jury rendering a decision—is even more critical:

> Perhaps the most common way that speakers lose audiences in presentations is that they drown their audiences with details. When you effectively present your work, you do not present everything about your work. Rather, you select those details that allow the audience to understand the work, and you leave out details that the audience does not desire or need. Effectively presenting your work also means that you sort details so that the audience is not faced with a long laundry list that has to be catalogued and synthesized on the spot. Finally, effectively presenting your work means that you provide a hierarchy of details so that the audience knows

238. ALLEY, *supra* note 234, at 35–36.

239. GARNER, *supra* note 237, at 69 (quoting RICHARD A. POSNER, THE PROBLEMS OF JURISPRUDENCE 464 (1990)).

240. Allergan, Inc. v. Teva Pharm. USA, Inc., No. 2:15-cv-1455-WCB, 2017 WL 480394, at *1 (E.D. Tex. Oct. 16, 2017).

241. *Id.* at *24–26.

> which details to hang onto and which details to let go in case they are overwhelmed.[242]

One example of effective storytelling in this vein is provided by *Gallaway v. Empire Fire and Marine Insurance*,[243] in which the court specifically praised an expert for the detailed factual groundwork laid in support of an air dispersion model, including meteorological records and emergency response reports. After laying that groundwork, the expert then had made appropriate inputs to the model regarding—among other things—the amount of chemical spilled, the location of the chemical source, the duration of the spill, and the position of the plaintiffs relative to that source.[244] Because the plaintiffs did not present any evidence to counter the defendants' model, the court found that the plaintiffs "could not have been exposed to harmful levels of [chemicals] that would have caused their alleged chronic symptoms" and granted summary judgment to the defendants.[245]

Fourth, consider using visuals to convey key pieces of information. A large body of research has demonstrated that visual cues help people better to retrieve and to remember information, using the substantial resources of the sensory cortex to process images, rather than words.[246] Pictorial images, in particular, "typically convey the same information that text might, but . . . anchor the information through what educational psychologists call dual-coding of text and images."[247] Use of visuals can be particularly

---

242. ALLEY, *supra* note 234, at 88; *see also* MICHAEL ALLEY, THE CRAFT OF SCIENTIFIC WRITING 20 (4th ed. 2018) ("If I had only one piece of stylistic advice to whisper into the ear of every scientist and engineer, that advice would be 'to avoid needless complexity.'").

243. Gallaway v. Empire Fire & Marine Ins., No. 03-113, 2007 WL 1199502, at *2 (W.D. La. Apr. 20), *aff'd*, 255 F. App'x 892, 893 (5th Cir. 2007).

244. *Id.* at *3.

245. *Id.* at *1–2. *See also* ANTONIN SCALIA & BRYAN GARNER, MAKING YOUR CASE: THE ART OF PERSUADING JUDGES 111–12 (2008).

> Nothing clarifies [the] meaning [of abstract concepts] as well as examples. One can describe the interpretive canon *noscitur a sociis* as the concept that a word is given meaning by the words with which it is associated. But the reader probably won't really grasp what you're talking about until you give an example . . . : "pins, staples, rivets, nails, and spikes." In that context, "pins" couldn't refer to lapel ornaments, "staples" couldn't refer to standard foodstuffs, "nails" couldn't refer to fingernails, and "spikes" couldn't refer to hairstyles.

Id. See supra notes 82–85 and accompanying text (explaining lawyers may usefully apply some of the standards beginning to be used in the scientific community to enhance reliability of statistical evidence through additional emphasis on the context surrounding the conclusions of a particular study without singular focus on the p-value, including study design, the quality of the data, and the underlying scientific mechanisms being tested).

246. *See* Steve Johansen & Ruth Anne Robbins, *Art-iculating the Analysis: Systemizing the Decision to Use Visuals as Legal Reasoning*, 20 J. LEGAL WRITING INST. 57, 60 n.6 (2015).

247. *Id.* at 67.

important in disputes centered on math and science, because visuals help simplify complex concepts. For example, the graphs used in this Article illustrate key study findings that could not as readily be described in words.

A wealth of literature has been developed to advise lawyers on how best to utilize graphics to persuade.[248] Recognizing the power of visuals, commentators have urged lawyers and law schools to address these issues just as rigorously as textual argument:

> To competently analyze a visual argument from a professional and ethical standpoint, attorneys also need a thorough grounding in principles of visual rhetoric, an emerging discipline that draws upon cultural studies, psychology, classical rhetoric, and media studies. Visual rhetoric asks how visual arguments are constructed and how images persuade. Because our common law system is still text-based and reliant upon inductive and deductive reasoning, visual rhetoric can help lawyers translate visual arguments into logical text and vice versa. This translation skill is necessary to construct a visual argument in the legal context and to spot weaknesses and fallacies in an opposing counsel's visual argument.[249]

Likewise, Elizabeth Porter has observed that "longstanding barriers to visual persuasion in written advocacy are finally coming down,"[250] representing a "sharp departure from the accepted tone of written legal argument."[251] Porter adds that visual images in legal documents are:

> At the same time, . . . seductively natural, because we are bombarded daily with visual messages, and because in the

248. *See e.g.*, Kerri L. Ruttenberg, IMAGES WITH IMPACT: DESIGN AND USE OF WINNING TRIAL VISUALS 3 (2017); JOHN SAMMONS & LARS DANIEL, DIGITAL FORENSICS TRIAL GRAPHICS: TEACHING THE JURY THROUGH EFFECTIVE USE OF VISUALS 2 (2017). However, these visual presentations are not easy. Edward Tufte, a noted scholar of data visualization, provides a well-known example of poor data presentation choices about O-ring performance under cold conditions that directly led to the space shuttle *Challenger* disaster. *See* EDWARD R. TUFTE, VISUAL EXPLANATIONS: IMAGES AND QUANTITIES, EVIDENCE AND NARRATIVE 40 (2d ed. 1997).

249. Lucille A. Jewel, *Through a Glass Darkly: Using Brain Science and Visual Rhetoric to Gain a Professional Perspective on Visual Advocacy*, 19 S. CAL. INTERDISC. L.J. 237, 239–40 (2010) (footnotes omitted). *See also* Lauren A. Newell, *Redefining Attention (and Revamping the Legal Profession?) for the Digital Generation*, 15 NEV. L.J. 754, 806 (2015) (urging education on new technologies to prepare students for future practice); William Wesley Patton, *Opening Students' Eyes: Visual Learning Theory in the Socratic Classroom*, 15 L. & PSYCH. REV. 1 (1991).

250. Elizabeth G. Porter, *Taking Images Seriously*, 114 COLUM. L. REV. 1687, 1718 (2014).

251. *Id.* at 1723 (footnote omitted).

> past decade other forms of writing—even "serious" scholarly or journalistic writing—have embraced multimedia exposition. Multimedia legal argument is thus novel and, simultaneously, utterly pedestrian—an intriguing combination that may lead courts, and scholars, to underestimate its potential impact on the structure and substance of legal decisionmaking.[252]

Such impacts are even more important where interpretation of mathematical concepts like p-values are critical to the decisional result.

CONCLUSION

P-values are complex, especially in the context of multifaceted pharmaceutical cases. However, p-values are just another form of evidence, and one in which understanding a few crucial points of interpretation may well make a difference to the outcome. In the cases examined in this Article, the losing parties overestimated the p-values that favored their cases, ignoring other relevant data as well as factors like sample size, multiple

---

252. *Id.* In this context, commentators have highlighted the potential for misuse of visual evidence, and the need for rigorous analysis by lawyers. *See* Dan M. Kahan et al., *Whose Eyes Are You Going to Believe?* Scott v. Harris *and the Perils of Cognitive Illiberalism*, 122 HARV. L. REV. 837, 839–40 (2009) (discussing Scott v. Harris, 550 U.S. 372 (2007), which involved a finding of whether police force was excessive in the context of a car chase). While the Eleventh Circuit had found that reasonable minds could differ on whether lethal force was needed by watching a video of the car chase taken from a state patrol car, Harris v. Coweta Cnty., 433 F.3d 807, 810 (11th Cir. 2005), the Supreme Court disagreed. Watching the same state patrol car video that was considered in the lower court, the Supreme Court concluded that no reasonable juror could have found that lethal force was not required under these circumstances, stating that the defendant's "version of events is so utterly discredited by the record that no reasonable jury could have believed him. The Court of Appeals should not have relied on such visible fiction; it should have viewed the facts in the light depicted by the videotape." *Scott*, 550 U.S. at 380–81. In so doing, the Supreme Court arguably usurped the power of the jury. As Richard Sherwin has explained:

> The Supreme Court majority was taken in by a mental phenomenon that social scientists call "naïve realism": the natural tendency to spot other people's biases, but not your own. No one's perceptions are perfectly neutral. They are at least to some extent shaped by beliefs and expectations, which in turn are influenced by such factors as education, social and cultural background, and ideology—all of which help us to construct the reality we see. By ruling that no reasonable juror could see anything other than what the majority of the Supreme Court saw when they watched the chase video, the justices consigned the Eleventh Circuit majority and Justice Stevens [in the minority], together with all citizens who happen to share a similar set of perceptual and cognitive characteristics, to a special kind of cognitive purgatory. They cordoned off a no man's zone for people whose perceptions don't count.

Richard K. Sherwin, *Visual Literacy for the Legal Profession*, 68 J. LEGAL EDUC. 55, 60 (2018).

comparison adjustments, primary outcome measures, and FDA determinations. While consideration of these issues may not have won those cases, acknowledgement of them would have strengthened the judicial presentations. When the math matters, lawyers must embrace the complexity presented by p-values and conquer the communication processes that are necessary to best serve their clients.