P-Values on Trial: Selective Reporting of (Best Practice Guides Against) Selective Reporting

Deborah Mayo¹

¹Department of Philosophy, College of Liberal Arts and Human Sciences, Virginia Tech, Blacksburg, Virginia, United States of America

Published on: Jan 31, 2020

DOI: <https://doi.org/10.1162/99608f92.e2473f6a>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

In an attempt to stem the practice of reporting impressive-looking findings based on data dredging and multiple testing, the American Statistical Association's (ASA) 2016 guide to interpreting p values (Wasserstein & Lazar) warns that engaging in such practices “renders the reported p -values essentially uninterpretable” (pp. 131-132). Yet some argue that the ASA statement actually frees researchers from culpability for failing to report or adjust for data dredging and multiple testing. We illustrate the puzzle by means of a case appealed to the Supreme Court of the United States: that of Scott Harkonen. In 2009, Harkonen was found guilty of issuing a misleading press report on results of a drug advanced by the company of which he was CEO. Downplaying the high p value on the primary endpoint (and 10 secondary points), he reported statistically significant drug benefits had been shown, without mentioning this referred only to a subgroup he identified from ransacking the unblinded data. Nevertheless, Harkonen and his defenders argued that “the conclusions from the ASA Principles are the opposite of the government's” conclusion that his construal of the data was misleading (*Harkonen v. United States*, 2018, p. 16). On the face of it, his defenders are selectively reporting on the ASA guide, leaving out its objections to data dredging. However, the ASA guide also points to alternative accounts to which some researchers turn to avoid problems of data dredging and multiple testing. Since some of these accounts give a green light to Harkonen's construal, a case might be made that the guide, inadvertently or not, frees him from culpability.

Keywords: statistical significance, p values, data dredging, multiple testing, ASA guide to p values, selective reporting

1. Introduction

The biggest source of handwringing about statistical inference boils down to the fact it has become very easy to infer claims that have not been subjected to stringent tests. Sifting through reams of data makes it easy to find impressive-looking associations, even if they are spurious. Concern with spurious findings is considered sufficiently serious to have motivated the American Statistical Association (ASA) to issue a guide to stem misinterpretations of p values (Wasserstein & Lazar, 2016; hereafter, ASA guide). Principle 4 of the ASA guide asserts that:

Proper inference requires full reporting and transparency. P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain p -values (typically those passing a significance threshold) renders the reported p -values essentially uninterpretable. (pp. 131–132)

An intriguing example is offered by a legal case that was back in the news in 2018, having made it to the U.S. Supreme Court (*Harkonen v. United States*, 2018). In 2009, Scott Harkonen (CEO of drug company

InterMune) was found guilty of wire fraud for issuing a misleading press report on Phase III results of a drug Actimmune in 2002, successfully pumping up its sales. While Actimmune had already been approved for two rare diseases, it was hoped that the FDA would approve it for a far more prevalent, yet fatal, lung disease (whose treatment would cost patients \$50,000 a year). Confronted with a disappointing lack of statistical significance ($p = .52$)¹ on the primary endpoint—that the drug improves lung function as reflected by progression free survival—and on any of ten prespecified endpoints, Harkonen engaged in postdata dredging on the unblinded data until he unearthed a non-prespecified subgroup with a *nominally* statistically significant survival benefit. The day after the Food and Drug Administration (FDA) informed him it would not approve the use of the drug on the basis of his post hoc finding, Harkonen issued a press release to doctors and shareholders optimistically reporting Actimmune’s statistically significant survival benefits in the subgroup he identified from ransacking the unblinded data.

What makes the case intriguing is not its offering yet another case of *p*-hacking, nor that it has found its way more than once to the Supreme Court. Rather, it is because in 2018, Harkonen and his defenders argued that the ASA guide provides “compelling new evidence that the scientific theory upon which petitioner’s conviction was based [that of statistical significance testing] is demonstrably false” (Goodman, 2018, p. 3). His appeal alleges that “the conclusions from the ASA Principles are the opposite of the government’s” charge that his construal of the data was misleading (*Harkonen v. United States*, 2018, p. 16).

Are his defenders merely selectively reporting on the ASA guide, making no mention of Principle 4, with its loud objections to the behavior Harkonen displayed? It is hard to see how one can hold Principle 4 while averring the guide’s principles run counter to the government’s charges against Harkonen. However, if we view the ASA guide in the context of today’s disputes about statistical evidence, things may look topsy turvy. None of the attempts to overturn his conviction succeeded (his sentence had been to a period of house arrest and a fine), but his defenders are given a leg to stand on—wobbly as it is. While the ASA guide does not show that the theory of statistical significance testing ‘is demonstrably false,’ it might be seen to communicate a message that is in tension with itself on one of the most important issues of statistical inference.

Before beginning, some caveats are in order. The legal case was not about which statistical tools to use, but merely whether Harkonen, in his role as CEO, was guilty of intentionally issuing a misleading report to shareholders and doctors. However, clearly, there could be no hint of wrongdoing if it were acceptable to treat post hoc subgroups the same as prespecified endpoints. In order to focus solely on that issue, I put to one side the question whether his press report rises to the level of wire fraud. Lawyer Nathan Schachtman argues that “the judgment in *United States v. Harkonen* is at odds with the latitude afforded companies in securities fraud cases” even where multiple testing occurs (Schachtman, 2020, p. 48). Not only are the intricacies of legal precedent outside my expertise, the arguments in his defense, at least the ones of interest here, regard only the data interpretation. Moreover, our concern is strictly with whether the ASA guide provides grounds to exonerate Harkonen-like interpretations of data.

I will begin by describing the case in relation to the ASA guide. I then make the case that Harkonen's defenders mislead by omission of the relevant principle in the guide. I will then reopen my case by revealing statements in the guide that have thus far been omitted from my own analysis. Whether they exonerate Harkonen's defenders is for you, the jury, to decide.

2. The Harkonen Trial and His Interpretation of Data

The study in the Harkonen trial is an example of a randomized double-blind study: 330 patients with idiopathic pulmonary fibrosis (IPF) are randomly assigned to receive Actimmune or placebo. They were mainly looking for improved lung function, or at least a delay in the progression of scar tissue build-up that is caused by IPF. This was the “primary endpoint” of the trial. They observe differences in the proportions in the two groups, treated and control, who enjoy “progression-free survival.” The problem of course is that there is ordinary variability in disease progression; it might be that, just by chance, more patients who would have shown improvement even without the drug happen to be assigned to the treated group. The random assignment allows determining the probability that an even larger difference would be observed even if it is correct to hold H_0 : the two groups, treated and control, do not differ with respect to lung function.² This is the p value. If even larger differences than observed occur fairly frequently even under H_0 (the p value is not small), there is little evidence of incompatibility with H_0 .

The first principle in the ASA guide states³:

P-values can indicate how incompatible the data are with a specified statistical model... The smaller the p -value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p -value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions. (p. 131)

The difference between sample data $\mathbf{X}$ and H_0 is measured by a test statistic $d(\mathbf{X})$, its observed value being $d(\mathbf{x}_0)$. The p value is $\Pr(d(\mathbf{X}) \geq d(\mathbf{x}_0); H_0)$. If this is low, then there's a high probability that a smaller difference would have occurred, were H_0 in fact adequate: $\Pr(d(\mathbf{X}) < d(\mathbf{x}_0); H_0) = 1 - p$ is high. So low p values indicate at least *some* incompatibility or discrepancy with ascribing the observed difference to chance variability as in H_0 .

Requiring a small p value before inferring an indication of a genuine discrepancy from H_0 controls the probability of a false positive or a *Type I error*: erroneously finding evidence against H_0 :

[The p value] is the probability that we would mistakenly declare there to be evidence against H_0 , were we to regard the data under analysis as just decisive against H_0 . (Cox & Hinkley, 1974, p. 66)

The p value on the primary endpoint in the Actimmune trial was .52, "entirely consistent with [Actimmune having] no effect" on progression-free survival (*Harkonen v. United States*, 2013[B], p. 5). Chance variability alone would fairly often produce an even larger difference, even if Actimmune had no effect. In the same brief, InterMune's chief biostatistician testified, "it was immediately apparent that the study had missed its primary endpoint as well as all ten of the secondary endpoints" (p. 4), listed in order of clinical relevance (the seventh being survival), along with eight "exploratory" endpoints.

Even if the p value had been small, the ASA guide correctly warns (Principle 5), it does not, by itself, tell us the effect size or discrepancy from H_0 that is indicated. We should supplement the p value with the p value distribution under various discrepancies from the hypothesized mean μ_0 , say in a test of $H_0: \mu = \mu_0$. If you very probably would have observed a more impressive (smaller) p value than you did, were μ as great as μ_1 (where $\mu_1 = \mu_0 + \gamma$), then the data indicate that $\mu < \mu_1$.⁴ However, this depends on a valid p value. The ASA guide warns that selectively reporting a small p value computed on a post-data subgroup is illicit (Principle 4). It would be to report what is sometimes called a *nominal* p value, in contrast to an *actual* one.

According to testimony, Harkonen set out to "cut that data and slice it until [he] got the kind of results [he was] looking for" (*Harkonen v. United States*, 2013[A], p. 4a). Aware that he was violating normal procedure, he blocked his own scientists from reviewing the report before issuing it. He instructed those performing the data dredging not to tell the FDA about the full extent of his post hoc analysis because he "didn't want to make it look like we were doing repeated analyses looking for a better result" (*Harkonen v. United States*, 2013[A], p. 4a). When he found what he was looking for—a nominally significant improvement in a group he labeled "mild to moderate IPF"—he issued a press release for doctors and investors touting "a significant survival benefit in patients with mild to moderate disease randomly assigned to Actimmune versus control treatment ($p = 0.004$)" (*Harkonen v. United States* 2013[A], p. 84a). His press release misled "readers into believing that the reported result was a statistically significant conclusion of a double-blind trial, rather than a post hoc analysis generated only after intentional repackaging of the data" (*Harkonen v. United States*, 2013[B], p. 19). An impressive bottomline for the company was all but assured, thanks to the "positive" trials he declared.⁵ Now this was merely a press release, not a paper published in a scholarly journal. The FDA admitted it was making an example of this case. It was deliberately showing it was prepared to hold individual researchers accountable for using misleading statistical practices for monetary gain, not just companies (it also fined InterMune).

It is important not to run together the different null hypotheses, each with their corresponding alternatives. Focus on the two main ones: (1) *the primary null*, H_0 : no improved lung function among experimental patients with IPF, and (2) *the postdata (PD) null* constructed from data dredging, H_0^{PD} : no survival benefit among IPF patients in the "mild to moderate category." Finding "incompatibility" with these null hypotheses indicates, respectively, H_1 : some improved lung function among the treated IPF patients, and H_1^{PD} : some survival

benefit in the PD subgroup of patients. It was Harkonen’s inference about the latter that was deemed blameworthy (he did not allege the former).

3. ASA Guide in Defense of Harkonen: FDA Ignores Context, Is Too Rigid

3.1. Importance of Context

Defenders of Harkonen note the ASA guide’s insistence that p values be interpreted in context. “[Rather than seeking bright-line rules,] [r]esearchers should bring many contextual factors into play to derive scientific inferences” (Goodman, 2018, p. 7). True enough, but the case *against* Harkonen turned precisely on context. To properly interpret a p value, the prosecutors stated, “it is necessary to know the context in which that p -value was generated” (*Harkonen v. United States*, 2013[A], p. 19a). In particular, they emphasized, one needs to know whether the primary endpoint was met (i.e., statistically significant), the number of secondary endpoints, and if hypotheses were constructed postblinding. Far from being an unthinking exercise, the choice of primary endpoints is intended to reflect the background understanding of the drug’s mechanism of action. Failure of the primary endpoint, the company’s own biostatistician explained, casts doubt on the causal theory upon which the trial was built, that the drug improved lung function. He saw “no apparent way from the data that the drug would be working” to increase survival, “given the absence of any indication that it actually improved lung function” (*Harkonen v. United States*, 2013[B], p. 7). He went further: not only was the “mild to moderate” classification invented for the purpose, it “defined the subgroups in such a way that patients with severe IPF had a *higher* deathrate if they had used Actimmune than if they had used the placebo—not what one would expect to see if Actimmune truly helped IPF patients live longer” (p. 8). It is precisely because the p value is a context-dependent measure that Harkonen’s misdescription of it was flagged.

Nor is it correct to charge the FDA is denying we can learn from factors discovered postdata. In fact, the FDA immediately took the survival benefit in the post hoc group as a worthy exploratory factor, and helped InterMune to subject the claim to a legitimate test. The new trial enrolled only patients satisfying the category Harkonen identified (“mild to moderate” lung damage). Over twice as large as the earlier study, 826 patients from 81 hospitals were randomly assigned to Actimmune or placebo, with survival now the primary endpoint. It had to be stopped early, after about a year, because the treated group had a higher death rate than the placebo group. Our focus is just on the 2002 data, not the new trial.

The government’s position against Harkonen, notice, merely denies that H_1^{PD} was well-tested by the same data that was used to find the subgroup. Denying the claim that H_1^{PD} was well-tested (by the particular trial and data) is importantly different from denying the claim H_1^{PD} altogether. The latter would be to assert H_1^{PD} is false and H_0^{PD} true. Inferring from a nonstatistically significant difference (which his $p = .5$ clearly was) to claiming there’s zero association is a well-known fallacy. The government does not commit this fallacy.

3.2. Overly Rigid?

According to Harkonen's 2013 defenders, in an amicus brief, the government's position is based upon "an extreme, rigid view within statistics" wherein "multiple post-hoc testing renders subgroup findings 'meaningless'" (Rothman, Lash, & Schachtman, 2013, p. 20). If so, they would have to denounce the guide as equally rigid. The ASA guide's Principle 4 states:

Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p -values computed. Valid scientific conclusions based on p -values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p -values) were selected for reporting. (Wasserstein & Lazar, 2016, p. 132)

Remember that our issue is whether *the ASA guide offers grounds* to exonerate Harkonen's interpretation of data (not whether such grounds might be found elsewhere). It cannot do so insofar as Principle 4 expressly condemns treating post hoc hypotheses in the same way as if they were predesignated. (If there is post hoc dredging, then the study is no longer double-blind.) This principle is well-integrated into best practice manuals.

Registering clinical trials and pre-specifying their outcomes are now mandated by legislation in numerous territories, including the US, with strong support from global organisations, ... and an extensive range of professional bodies, [including] the Consolidated Standards of Reporting Trials (CONSORT) guidelines, on best practice in trial reporting, which are endorsed by 585 academic journals. (Goldacre et al., 2019, p. 2)

In the *Reference Manual on Scientific Evidence* for judges, David Kaye and David Freedman emphasize the importance of asking:

How many tests have been performed? Repeated testing complicates the 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. (Kaye & Freedman, 2011, p. 127)

Social scientists increasingly look to best practices in medicine to put their own houses in order. The new move in psychology to preregister sampling plans, unsurprisingly, has led to many more negative results and improved replication by independent groups.

Nevertheless, rules can be changed if their rationale is overturned. It would be ironic if a guide, intended to block flexible practices that are known to lead to erroneous conclusions and irreproducible results, could be construed as grounds to be more permissible.

3.3. Defending Harkonen Before and After the 2016 ASA Guide

Interestingly, prior to the 2016 ASA guide, Harkonen's defenders, in an amicus brief, point to the fact that: "There is no consensus whether, when, or how to adjust p -values or Type I error rates for multiple testing" (Rothman et al., 2013, p. 21). Surely, one cannot be blamed for interpreting data in a way that is still open to debate—or so his defenders argued in appeals prior to 2016. After the guide, in the 2018 appeal, no mention is made of lack of consensus on multiple testing, nor even that Harkonen's culpability centered on his data dredging. Moreover, this new appeal goes further than the others⁶ by declaring that the ASA guide shows his "actual innocence" (*Harkonen v. United States*, 2018, p. 2). Perhaps it is supposed the guide settles the issue—in favor of those who deny the multiple testing adjustment. How can this be? The guide, with its Principle 4, clearly comes down on the side of those who think it matters. The ASA guide itself maintains: "Nothing in the ASA statement is new". It is merely "a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the p -value"—not changing or adjudicating any controversial principles (Wasserstein and Lazar, 2016, pp. 130-131).

It appears that those taking the guide as exonerating Harkonen's interpretation mislead by omission of the principle that Harkonen is flouting. The alleged basis for the conviction is shifted. Harkonen's conviction, it is claimed, is "premised on the fundamentally flawed view that a non-significant p -value, by itself, falsifies a claim that a relationship exists" (Goodman, 2018, p. 10). The only relationship the press report claimed to have evidence for is H_1^{PD} : some survival benefit in the PD subgroup of patients. But, as noted, the prosecution does not rest on claiming H_1^{PD} is false, but rather very poorly tested by dint of Harkonen's data dredging. Even calling it "a non-significant P -value" hides the real problem: In the language of Principle 4: a valid p value "cannot be drawn." On the face of it, the 2018 appeal declares his 'actual innocence' only by changing what Harkonen was found guilty of.

Nor will it suffice for the defense to declare that "no rational jury could convict Dr. Harkonen in light of the ASA's rejection of the government's false statistical theory" (*Harkonen v. United States*, 2018, p. 28). Nowhere does the ASA Guide reject statistical significance testing.

4. Other Approaches and Alternative Measures of Evidence

But wait. My report thus far is itself guilty of selectively reporting the ASA guide, and failing to take account of the general context in which it was issued. I have omitted a relevant portion of the guide, notably the final section, immediately after the principles:

Other Approaches. In view of the prevalent misuses of and misconceptions concerning p -values, some statisticians prefer to supplement or even replace p -values with other approaches. These include methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian

methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors. (Wasserstein & Lazar, 2016, p. 132)

Tests are usually supplemented with confidence intervals (CIs), as they should be. However, there is a duality between CIs and tests. The parameter values within the 95% CI are those that would not be rejected at the 0.05 level (with the usual two-sided test). Thus, the same selective reporting will result in CIs failing to cover the true value with the claimed coverage probability.⁷ So Harkonen's postdata inference still would not be countenanced if tests were replaced by CIs.

4.1. Alternative Measures of Evidence

Can any of the alternative measures of evidence serve the significance test function while freeing Harkonen's interpretation from the admonition of Principle 4? Here is where different philosophies behind the alternative measures of evidence enter. Guide readers aware of long-standing battles among Bayesians, likelihoodists, and frequentists know that a key issue is how to treat multiple-testing and data-dependent selection. Are the alternative measures of evidence free from those concerns? The ASA guide does not directly tell us. If the guide is assenting to these other methods (for the task of distinguishing signal from noise), then it is going significantly *beyond* rehearsing agreed-upon principles for avoiding misinterpreting p values. It may be averred that the guide is merely reporting that some statisticians 'prefer to supplement or even replace' tests with alternative methods—not endorsing these methods. Given that this is a best practice guide, the reader may suppose the alternatives are regarded as at least acceptable.

Consider likelihood ratios or Bayes factors. The same data-dredged alternative that caused all the brouhaha can find its way into a likelihood ratio or Bayes factor. Suppose Harkonen opted to form a likelihood ratio using the post hoc hypothesis, H_1^{PD} : Actimmune increases survival in IPF patients in the postdata subgroup:

$$Pr(x|H_1^{PD})/Pr(x|H_0^{PD}).$$

The alternative H_1^{PD} would be comparatively better supported (for the likelihoodist) or made more probable (for the Bayes factor theorist). We know this because H_0^{PD} has been deliberately selected as a subset in which those with Actimmune do better than the placebo group. The significance tester, by contrast, must go beyond the ratio of likelihoods to the overall method that found evidence for H_1^{PD} in the particular case. As Pearson and Neyman (1930, p. 106) put it: "In order to fix a limit between 'small' and 'large' values of [the likelihood ratio] we must know how often such values appear when we deal with a true hypothesis". Significance testers ask: 'What is the probability Harkonen's method would identify some such subgroup *or other* that gives a high likelihood ratio, even if spurious?'—an error probability. Discovering this probability is high, they deny the high likelihood ratio counts as evidence for H_1^{PD} .

4.2. Error Probabilities and Error Statistical Accounts

Accounts that employ error probabilities to control and assess the capability of a method to avoid error, I call *error statistical*. This is more apt and inclusive than *frequentist*. The basis for Principle 4 is that unless you know the number of endpoints, any data-dependent subgroups—and more generally, what is called the *sampling plan*—you cannot correctly assess a method's error probabilities. The Type I error, erroneously finding evidence against a null hypothesis, is of main interest in this case. Surprisingly, according to some alternative measures of evidence, such considerations violate a key principle of evidence. Subjective Bayesian Dennis Lindley observes: "Sampling distributions, significance levels, power, all depend on...something that is irrelevant in Bayesian inference—namely the sample space" (Lindley, 1971, p. 436). In other words, once the data are in hand, outcomes other than the one observed are irrelevant. The evidence from the data is contained in the likelihood ratio, which *conditions* on the actual outcome. The formal principle stating this is the *Likelihood Principle*.

Either the other approaches require the same treatment of multiple testing and postdata subgroups as statistical significance tests or they do not. If they do, then the basis for criticizing Harkonen remains. If they do not, then the message from the ASA guide may well be seen to absolve Harkonen of blame for his interpretation—remembering that we are only considering this issue. Principle 4 of the ASA guide (p. 131) is said to apply to "*p*-values and related analyses". This suggests that the threat to validity applies only to error statistical accounts. The message would appear to be that although *p* values and significance levels require adjusting or at least reporting multiple testing and postdata dredging, these other approaches do not.

Yoav Benjamini (2016), in his commentary on the ASA guide, objects to the *p* value being singled out as high risk in this manner. Although the *p* value can be distorted by selection effects, Benjamini argues, this is a problem affecting all of today's industrialized uses of statistics. Yet it is routinely described as an unfortunate affliction of *p* values. Regina Nuzzo (2018), an influential statistics communicator at ASA, describes as "one little-known requirement" for valid use of *p* values: "is that the data analysis be planned before looking at the data" and "all analyses and results be presented." In her view, "these seem like strange, nit-picky rules, but they're part of the deal when using *p*-values." Users of alternate measures of evidence are apparently free of such rules. For example, we hear that "Bayes factors can be used in the complete absence of a sampling plan, or in situations where the analyst does not know the sampling plan that was used" (Bayarri, Benjamin, Berger, & Sellke, 2016, p. 100). Had Harkonen given a likelihood ratio or Bayes factor analysis, he would apparently be free of the main charge against him.⁸

The fact that a formal measure of evidence—say, a likelihood ratio—may be unchanged by data dredging or other selection effects, does not make it immune to selection effects. The probability of that measure resulting in erroneously declaring evidence in favor of a genuine effect is still altered, perhaps radically. An account that does not assess error probabilities simply will not register this. To the error statistician, acceptable measures of evidence must have antennae that directly pick up on gambits that alter a method's ability to control

erroneously taking spurious effects as real. So a deeper question is whether an account ought to register shifts to error probabilities. This issue should be at the forefront of today's debates about competing reforms.

Admittedly, there are a number of different ways to adjust for multiplicity: it is a topic of ongoing research. One, the Bonferroni adjustment, multiplies the attained p value with the number of predesignated endpoints. So, for example, if the significance level of .05 is used and 10 endpoints are tested, a p value of .005 (as opposed to .05) would be required to claim statistical significance. (I take that's why Harkonen dredged for a nominal p value of .004.) A less conservative approach developed by Benjamini and Hochberg (1995) is the false-discovery rate (FDR): the expected proportion of the N hypotheses tested that are falsely rejected.

Even with disagreement on how best to deal with multiplicity, many would agree that "failure to acknowledge the multiple testing is poor statistical practice" (Schachtman, 2014). However, if alternative notions of evidence, legitimate for the ASA, free us from having to take into account multiple testing, then perhaps there is no call to do so in reporting p values either. Steven Goodman—who also wrote a 'friend of the court' (amicus) brief in support of Harkonen—maintains:

Two problems that plague frequentist inference: multiple comparisons and multiple looks, or, as they are more commonly called, *data dredging* and peeking at the data. The frequentist solution to both problems involves adjusting the P value...But adjusting the measure of evidence because of considerations that have nothing to do with the data defies scientific sense. (Goodman, 1999, p. 1010)

This is in tension with Principle 4 of the ASA guide. It would not be correct to claim p values cannot be interpreted without knowing of the data dredging, if their influence on error probabilities did not affect the import of the data.

4.3. Evidence, Action, and Belief

Those who reject a concern with error probabilities generally give the reason that they only matter to a concern with the long-run performance of a method. However, the real rationale for Principle 4 is not a concern for long-run error control. Although Harkonen's trying and trying again (until he found a way to make Actimmune look good) would result in high error rates in the long run of uses, that is not the reason we would question the scientific standing of a resulting inference. It is rather that the claim that Actimmune improves IPF in the postdata subgroup (H_1^{PD}) has not been well-tested: it has not been *severely* probed (see Mayo 2018).

Some will say, agreeing with the likelihoodist Richard Royall (1997, 2004), that if you reject the inference to H_1^{PD} , despite its high comparative likelihood, then you are going beyond evidence to either action—which requires considering error probabilities—or belief—which requires considering prior probabilities in the hypotheses (in a Bayesian computation). Suppose for the sake of argument we go along with these distinctions. Consider the category of action. Since the FDA is concerned with action (to approve a drug), Harkonen is once again not spared from culpability for misreporting the test's error probabilities. What about invoking prior

beliefs in the truth of hypotheses? An inference to the postdata hypothesis H_1^{PD} may be blocked by giving the null hypothesis (of no benefit) a suitably high prior probability. A common assignment is .5. There is much disagreement as what these prior probabilities should mean and how to arrive at them. Many claim a point null of no effect is always false, and thus would oppose assigning it the high prior needed to block the inference. Some try to keep to quasi-frequentist priors. Here they try to estimate the frequency of true experimental effects among those thought to be similar to the one in hand. The trouble is that the prevalence of purportedly "true" hypotheses varies greatly depending on what reference class the hypothesis in question is placed in (e.g., all drug hypotheses, results of RCTs, hypotheses put forward by Harkonen).⁹ Others appeal to a prior degree of belief in the claim's truth, or a researcher's willingness to bet on it. However, data-dredged hypotheses are often extremely convincing—which is what makes them so seductive—and 'prior' probabilities may themselves be data-dependent. In our view, looking to priors to block the inference to a data-dredged hypothesis misidentifies what the problem really is. The influence of the biased selection is not on the plausibility of the claim, but on the capability of the test to have unearthed errors.

4.4. The 2019 Special Issue of *The American Statistician* (TAS) on Statistical Inference

Guest editors of the March 2019 supplemental issue "A World Beyond $p < 0.05$ "—R. Wasserstein, A. Schirm, and N. Lazar—recommend going much further than the 2016 ASA guide.¹⁰ They suggest that "declarations of 'statistical significance' be abandoned" (p. 2). They do not ban p values, but oppose using prespecified p value thresholds as the basis for interpreting results. They recognize this is at odds with the FDA's long-established drug review procedures. They think that by removing the p value thresholds, researchers will no longer be incentivized to data dredge, and otherwise exploit researcher flexibility. In fact, Harkonen still could not take the large (non-significant) p value to indicate a genuine effect. It would be tantamount to saying: Even though larger differences would frequently be expected by chance variability alone, the data provide evidence they are not due to chance variability. In short, he would still need to report a reasonably small p value. But this is to use a threshold. Say he ransacks the data just as he did (although he might not dredge quite as far). In a world without thresholds, it would be hard to hold him accountable for reporting a nominally small p value. According to the Wasserstein et al. 2019 editorial, "whether a p -value passes any arbitrary threshold should not be considered at all" in interpreting data (p. 2) (see Mayo, 2019 a, b). Thresholds via confidence levels and Bayes factors are also blocked. The problem is, if you cannot say about any results, ahead of time, that they will not be allowed to count in favor of a claim—if you reject all thresholds—then you do not have a test of it.¹¹ That is why John Ioannidis (2019) declares that "retiring statistical significance would give bias a free pass".

Even without the word "significance," and without p value thresholds, the numbers Harkonen reported as p values are not valid p values. So it is still possible, though more difficult, to find him culpable. If, however, the p value is regarded merely as a descriptive or logical measure between data and hypotheses, with no error probability assessment, one may concur with the authors of the 2013 amicus brief in support of Harkonen that

“Multiple Testing Does Not Undermine the Meaning of P-Values” (Rothman et al., p. 19). Yet this is at odds with Principle 4. A more careful analysis of the consequences of the guide’s recommendations is needed if it is not to boomerang. Else, there is a risk of relapsing to the problems the FDA (2017) has struggled to combat:

In the past, it was not uncommon, after the study was unblinded and analyzed, to see a variety of post hoc adjustments of design features (e.g., endpoints, analyses), usually plausible on their face, to attempt to elicit a positive study result from a failed study — a practice sometimes referred to as data-dredging. ... The results of such analyses can be biased .. The results also represent a multiplicity problem because there is no way to know how many different analyses were performed and there is no credible way to correct for the multiplicity of statistical analyses and control the Type I error rate. (pp. 7–8)

5. Brief Summation

The ASA guide cannot be taken to absolve a researcher from failing to mention, if not also to adjust for, data dredging, multiple endpoints, outcome switching—without selectively reporting on its principles—entirely omitting the principle around which the case against Harkonen turns: Principle 4. Pointing out that some accounts of evidence are unaltered by multiple testing and subgroup selection is a fair legal strategy. It might convince some jury members that he should never have been held to that standard, and in law it is appropriate to see winning as the primary goal. This would still not render Harkonen’s report a non-misleading construal of the p values attained for his inference. (At the actual trial, Harkonen brought no expert statistical witnesses to testify to the non-misleadingness of his reports.)

However, the fact that the ASA guide points to alternative accounts to which researchers may turn to avoid central problems of p values—the most notable being threats from data dredging and multiple testing—a case can be made that the guide, inadvertently or not, frees the practitioner from culpability in a case such as Harkonen’s. This explains why the authors of Wasserstein et al. (2019) could see the 2016 ASA guide as having “stopped just short” of their recommendation to abandon p value thresholds. We have shown this to be in tension with the guide’s Principle 4—at least if selection effects are thought to call for a p value adjustment. If the 2016 ASA guide opens the door to giving data dredging a free pass, the recommendation in Wasserstein et al. (2019) swings the door wide open. We can expect further cases to be brought before courts with the ASA guide in hand. The onus is on the ASA to clarify its standpoint on this important issue of statistical evidence.

Acknowledgments

I thank N. Schachtman, A. Spanos, P. Stark, and anonymous reviewers for helpful comments, corrections, and guidance on the statistical analysis. I thank Jean Miller for significant copyediting.

Disclosure Statement

Deborah G. Mayo has no financial or non-financial disclosures to share for this article.

Appendix

Appendix A: The ASA 2016 Guide's Six Principles (Wasserstein & Lazar, 2016)

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.

Appendix B: P values on Trial in the Case of Matrixx and Its Cold Remedy, Zicam

Harkonen charged that the government's ruling in his case was at odds with its ruling in the case of a drug company Matrixx (*Matrixx Initiatives, Inc. v. Siracusano* 2011), just before his sentencing in 2011. Matrixx was found guilty of securities fraud for failing to reveal information they had on adverse effects (loss of sense of smell) related to its cold remedy Zicam. Even though they were battling lawsuits by Zicam users, and had received FDA letters to stop issuing clean bills of health on Zicam, Matrixx assured investors there were no such regulatory threats. Since the evidence was anecdotal, and no statistical study has been done, the data were, strictly speaking, 'non-statistically significant.' Although it would have been proper for Matrixx to report these strictly non-statistically significant effects, it does not follow, despite what Harkonen contends, that it was proper for him to report evidence of benefit in his post hoc subgroup (where there was also no valid assessment of statistical significance). This was part of several of Harkonen appeals (see *Harkonen v. United States* 2013[A],[B] and 2018). Besides, Matrixx's violation of disclosure rules did not require showing anything about a causal connection between Zicam and loss of smell; it sufficed that Matrixx hid information of relevance to investors' risk-benefit analysis. Anyone who trades in biotech stocks knows that the slightest piece of news, rumors of anecdotal side effects, and much else, can radically alter a stock price in the space of a few hours. No formal p value is needed. But that does not sanction illicit uses of p values.

References

Bayarri, M., Benjamin, D., Berger, J., & Sellke, T. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72, 90–103.
<https://doi.org/10.1016/j.jmp.2015.12.007>

Benjamini, Y. (2016). It's not the p -values' fault. Comment on Wasserstein and Lazar (2016), "The ASA's statement on p -values: Context, process, and purpose" (the fourth entry in supplemental materials). *The American Statistician*, 70(2), 129–133. <https://doi.org/10.1080/00031305.2016.1154108>

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300. Retrieved from: www.jstor.org/stable/2346101

Cox, D. & Hinkley, D. (1974). *Theoretical statistics*. London, UK: Chapman and Hall. <https://doi.org/10.1201/b14832>

FDA (U. S. Food and Drug Administration) (2017). Multiple endpoints in clinical trials: Guidance for industry (*DRAFT GUIDANCE*). Retrieved from <https://www.fda.gov/media/102657/download>

Gelman, A., & Shalizi, C. (2013). "Philosophy and the practice of Bayesian statistics" and "rejoinder." *British Journal of Mathematical and Statistical Psychology*, 66(1), 8–38; 76–80. <https://doi.org/10.1111/j.2044-8317.2011.02037.x>; <https://doi.org/10.1111/j.2044-8317.2012.02066.x>

Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., ... Mahtani, K. R. (2019). COMPare: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, 20(1), 118. <https://doi.org/10.1186/s13063-019-3173-2>

Goodman, S. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, 130, 1005–1013. <https://doi.org/10.7326/0003-4819-130-12-199906150-00019>

Goodman, S. (2018). Brief to the United States Supreme Court of *Amici Curiae* in support of petitioner. W. Scott Harkonen v. United States of America, No. 18-417 (filed October 31, 2018). Retrieved from <https://www.law360.com/articles/1098870/attachments/3>

Harkonen v. United States, No. 13– (Supreme Court of the United States, filed August 5, 2013). Petition for a Writ of Certiorari (Haddad & Philips). (2013[A]). Retrieved from <http://sblog.s3.amazonaws.com/wp-content/uploads/2013/11/HarkonenPetitionAppendix-2.pdf>

Harkonen v. United States, No. 13–180 (Supreme Court of the United States, filed November, 2013). Brief to United States Supreme Court for the United States in opposition to petitioner. (Verrilli, Raman, Rao) (2013[B]). Retrieved from <https://www.justice.gov/sites/default/files/osg/briefs/2013/01/01/2013-0180.resp.pdf>

Harkonen v. United States, No. 18– (Supreme Court of the United States, filed October 1, 2018). Petition for a Writ of Certiorari. Retrieved from <https://errorstatistics.files.wordpress.com/2019/06/harkonen-v-us-scotus-2018-petn-cert.pdf>

- Ioannidis, J. (2019). Correspondence: Retiring statistical significance would give bias a free pass. *Nature*, 567(7749), 461. <https://doi.org/10.1038/d41586-019-00969-2>
- Kafadar, K. (2019, December 1). The year in review...and more to come. President's Corner, *AMSTATNEWS*. Retrieved from https://magazine.amstat.org/blog/2019/12/01/kk_dec2019/
- Kaye, D., & Freedman, D. (2011). Reference guide on statistics. In *Reference manual on scientific evidence* (3rd ed., pp. 83–178). <https://doi.org/10.17226/13163>
- Lindley, D. (1971). The estimation of many parameters. In V. Godambe & D. Sprott (Eds.) *Foundations of statistical inference*. (pp. 435–447). Toronto: Holt, Rinehart and Winston of Canada.
- Matrixx Initiatives, Inc. v. Siracusano*, 563 U.S. 27, 131 S.Ct. 1309 (2011). Retrieved from <https://www.courtlistener.com/opinion/212969/matrixx-initiatives-inc-v-siracusano/>
- Mayo, D. (1996). *Error and the growth of experimental knowledge*. Chicago, IL: University of Chicago Press.
- Mayo, D. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge, UK: Cambridge University Press. Retrieved from <https://doi.org/10.1017/9781107286184>
- Mayo, D. (2019a). P-value thresholds: Forfeit at your peril. *European Journal of Clinical Investigation*, 49(10), Article e13170. <https://doi.org/10.1111/eci.13170>
- Mayo, D. (2019b, June 17) “The 2019 ASA Guide to P-values and Statistical Significance: Don’t Say What You Don’t Mean” (Some Recommendations) (ii) [Blog post]. *Error Statistics Philosophy*. Retrieved from <https://errorstatistics.com/2019/06/17/the-2019-asa-guide-to-p-values-and-statistical-significance-dont-say-what-you-dont-mean-some-recommendations/>
- Mayo, D., & Cox, D. (2006). Frequentist statistics as a theory of inductive inference. *Lecture Notes-Monograph Series, Institute of Mathematical Statistics*, 49, 77–97. <https://doi.org/10.1214/074921706000000400>
- Mayo, D., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *British Journal for the Philosophy of Science*, 57(2), 323–357. <https://doi.org/10.1093/bjps/axl003>
- New England Journal of Medicine* (2019). New Manuscripts. Retrieved from <https://www.nejm.org/author-center/new-manuscripts>
- Nuzzo, R. (2018, October 15). *Tips for communicating statistical significance*. National Institutes of Health. <https://www.nih.gov/about-nih/what-we-do/science-health-public-trust/perspectives/science-health-public-trust/tips-communicating-statistical-significance>

Pearson, E. & Neyman, J. (1930). On the problem of two samples. *Bulletin of the Academy of Polish Sciences*, 73–96. Reprinted 1967 in J. Neyman & E. Pearson (Eds.) *Joint Statistical Papers of J. Neyman and E. S. Pearson* (pp. 99–115). London, UK: Cambridge University Press.

Raghu, G., Brown, K., Bradford, W., Starko, K., Noble, P. W., Schwartz, D. A., & King, T. E., Jr. (2004). A placebo-controlled trial of Interferon Gamma-1b in patients with idiopathic pulmonary fibrosis. *New England Journal of Medicine*, 350(2), 125–133. <https://doi.org/10.1056/NEJMoa030511>

Rothman, K., Lash, T., & Schachtman, N. (2013). Brief to United States Supreme Court in *Harkonen v. United States*, 2013 WL 5915131, 2013 WL 6174902 (Supreme Court Sept. 9, 2013) of *Amici Curiae* in Support of Petitioner. W. Scott Harkonen v. United States, No. 13–180, (9th Cir., filed September 4, 2013). Retrieved from <http://schachtmanlaw.com/wp-content/uploads/2010/03/KJR-TLL-NAS-Amicus-Brief-in-US-v-Harkonen-090413A.pdf>

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. Boca Raton, FL: Chapman and Hall, CRC Press. <https://doi.org/10.1201/9780203738665>

Royall, R. (2004). “The likelihood paradigm for statistical evidence” and “Rejoinder.” In M. Taper & S. Lele (Eds.), *The nature of scientific evidence* (pp. 119–138; 145–151). Chicago, IL: University of Chicago Press. Retrieved from <https://doi.org/10.7208/chicago/9780226789583.003.0005>

Schachtman, N. (2014, October 14). Courts can and must acknowledge multiple comparisons in statistical analyses [Blog post]. *Tortini Blog*. Retrieved from <http://schachtmanlaw.com/courts-can-and-must-acknowledge-multiple-comparisons-in-statistical-analyses/>

Schachtman, N. (forthcoming 2020) Chapter 29 Statistical evidence in products liability litigation. In S. Scharf, S. Marmor, & G. Sax (Eds.), *Product liability litigation: Current law, strategies and best practices*. NYC, NY: Practising Law Institute.

Wasserstein, R. & Lazar, N. (2016). The ASA’s statement on P-values: Context, process and purpose. *The American Statistician*, 70(2), 129–133. [ASA Guide] <https://doi.org/10.1080/00031305.2016.1154108>

Wasserstein, R., Schirm, A. & Lazar, N. (2019). Moving to a world beyond “ $p < 0.05$ ” [Editorial]. *The American Statistician*, 73(S1), 1–19. <https://doi.org/10.1080/00031305.2019.1583913>

©2020 Deborah G. Mayo. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.

Footnotes

1. Comparison of the study groups and estimation of the hazard ratio, with 95% confidence intervals (CI), were performed with the likelihood score test from the Cox proportional-hazards model, stratified according to smoking status. The 95% CI for the hazard rate was (0.6–1.2). Raghu et al. (2004). I thank an HDSR reviewer for this analysis. [↵](#)
2. While the test is specified as a two-sided test, looking for improvement or worsening, the FDA makes it clear they are interested in controlling the error of erroneously inferring a drug benefit. Following Cox and Hinkley, the analysis might be viewed as conducting two tests, one in each direction, and adjusting the p value for selection (Cox & Hinkley, 1974, p. 79). [↵](#)
3. For a list of the ASA guide's principles, see Appendix A. [↵](#)
4. This is part of what I call a *severity* assessment. While akin to confidence limits, we do not limit the analysis to a single confidence level. Moreover, the evidential warrant associated with different points in any interval are distinguished (Mayo, 1996, 2018; Mayo & Cox, 2006; Mayo & Spanos, 2006). [↵](#)
5. Harkonen sent a fax blast of the press release to more than 2,000 pulmonologists. Actimmune sales went from 11 million dollars in 2000 to 141 million in 2003 primarily due to prescriptions for IPF patients. [↵](#)
6. One of the earlier appeals points to another case where p values were “on trial.” See Appendix B. [↵](#)
7. The *New England Journal of Medicine* (NEJM, 2019) recently updated its guidelines for authors, taking the ASA guide into account. It stipulates that when tests are adjusted, so must CIs. However, “[w]hen no method to adjust for multiplicity of inferences or controlling false discovery rate was specified in the protocol ... the report of all secondary and exploratory endpoints should be limited to point estimates of treatment effects with 95% confidence intervals. In such cases, the Methods section should note that the widths of the intervals have not been adjusted for multiplicity and that the inferences drawn may not be reproducible. No P values should be reported for these analyses.” For these cases, in effect, they allow confidence levels to be used descriptively. [↵](#)
8. I thank Philip Stark for this point. [↵](#)
9. Typically, the comparison is between a point null and a ‘meaningful’ discrepancy for the alternative. Since these are two simple or point hypotheses, the Bayes factor is the same as the likelihood ratio. However, all other alternatives are effectively given 0 prior. [↵](#)
10. Karan Kafadar (2019), President of the ASA, emphasizes that although the March 2019 editorial was co-authored by an official of the ASA, it does not represent ASA policy. Thanks to her efforts, a new ASA working group plans to take up issues of statistical significance tests and reproducibility in 2020. [↵](#)

11. It is not clear what would happen to the use of statistical significance tests in testing assumptions of statistical models—found in both Bayesian and error statistical accounts (see Gelman and Shalizi 2013). [↗](#)