Statistics in Biopharmaceutical Research

ISSN: (Print) 1946-6315 (Online) Journal homepage: https://www.tandfonline.com/loi/usbr20

The Role of *p*-Values in Judging the Strength of Evidence and Realistic Replication Expectations

Eric W. Gibson

To cite this article: Eric W. Gibson (2020): The Role of *p*-Values in Judging the Strength of Evidence and Realistic Replication Expectations, Statistics in Biopharmaceutical Research, DOI: 10.1080/19466315.2020.1724560

To link to this article: https://doi.org/10.1080/19466315.2020.1724560

Published online: 23 Jun 2020.

STATISTICS IN BIOPHARMACEUTICAL RESEARCH
2020, VOL. 00, NO. 0, 1–13
https://doi.org/10.1080/19466315.2020.1724560

Taylor & Francis
Taylor & Francis Group

The Role of *p*-Values in Judging the Strength of Evidence and Realistic Replication Expectations

Eric W. Gibson

Clinical Development and Analytics, Novartis Pharmaceuticals, East Hanover, NJ

ABSTRACT

p-Values are viewed by many as the root cause of the so-called replication crisis, which is characterized by the prevalence of positive scientific findings that are contradicted in subsequent studies. The spectrum of proposed solutions includes redefining statistical significance, abandoning the concept of statistical significance, or eliminating the use of *p*-values altogether. The unintended consequence of these proposals has been confusion within the scientific community, especially in the absence of consensus or clear alternatives. The goal of this article is to reframe the perceived replication crisis. I argue that this crisis is to a large extent the result of excessive optimism based on unknowingly (and sometimes knowingly) overstated evidence. As a remedy, I suggest a four-part guide to navigating statistical inference with *p*-values that is accessible for scientists. Examples taken from pharmaceutical drug development for heart failure illustrate key concepts.

ARTICLE HISTORY
Received November 2019
Accepted January 2020

KEYWORDS
False discovery rate; Multiplicity; Selective inference; Significance test

1. Introduction

I was recently called upon to explain why a pivotal clinical trial had failed, after an earlier trial with the same assessment in a similar design was "statistically significant" according to $p < 0.05$. The findings of the earlier trial had been promptly published in a top-tier medical journal and funding was secured for a second confirmatory trial. The research team was surprised to observe $p = 0.386$ for the primary assessment in the subsequent trial. Understanding why the second trial failed required a discussion of how *p*-values measure evidence, what constitutes strong evidence, and how evidence can be overstated to create an unrealistic expectation of future replication. Unfortunately, excessive optimism for the replication of positive initial findings has led many to believe that science is suffering from a replication crisis.

To address the growing perception of a replication crisis, the American Statistical Association (ASA) recently published 43 articles on the misuses of *p*-values in statistical inference. The adjoining editorial (Wasserstein, Schirm, and Lazar 2019) recommended that scientists stop using the term "statistically significant" entirely, but noted the articles "do not sing as one" and reflect "deep dissonance." The current ASA president (Kafadar 2019) cautioned that the editorial and special issue may have had the unintended consequence of creating confusion among nonstatisticians, even leading some to "abandon statistical methods altogether." In contrast to the ASA, the *New England Journal of Medicine* (Harrington et al. 2019) recommended, "despite the difficulties they pose, *p*-values continue to have an important role in medical research, and we do not believe that *p*-values and significance tests should be eliminated altogether." An editorial in the journal *Clinical Trials* (Cook et al. 2019) also cautioned that "there is still a place for significance testing in clinical trials."

The objective of this article is to offer a guide for the role of *p*-values in judging whether the strength of evidence reflected in a set of data is persuasive. I begin by reframing the replication crisis as the consequence of excessive optimism, and follow with a brief summary of *p*-values versus fixed significance levels and the relationship with false discoveries. After a short note on interpreting *p*-values, I provide four recommendations for judging the strength of evidence. Although I focus on *p*-values, it is important to note that other statistical tools such as confidence intervals and Bayes factors are also prone to misuse. Two case studies from heart failure drug development help illustrate these concepts.

1.1. Replication Crisis or Excessive Optimism?

Replication is the deliberate repetition of an initial experiment to confirm its findings, and is a cornerstone of the scientific method. Ioannidis (2005a) suggested there is a crisis in scientific replication based on a statistical model for false positive findings, concluding that "most published research findings are false." Goodman and Greenland (2007) noted in their correspondence that Ioannidas' model "dramatically diminishes a study's evidential impact," driving the circular conclusion that most findings are false. However, the model was not the reason this article would be downloaded over 2.9 million times and give rise to the popular perception of a crisis in replication. The article connected to the emotional disappointment experienced by

CONTACT Eric W. Gibson eric.gibson@novartis.com Clinical Development and Analytics, Novartis Pharmaceuticals, One Health Plaza, East Hanover, NJ 07936.

so many scientists when a promising initial finding disappears in the next study.

The popular perception of a replication crisis has caused many to focus on p-values as the main problem. Siegfried (2010) suggested that contradicted scientific findings are more prevalent because scientists depend on p-values to declare their results significant. The journal of *Basic and Applied Social Psychology* (BASP) banned the use of p-values and null hypothesis significance testing (Trafimow and Marks 2015). The ASA issued a statement on best practices for p-values (Wasserstein and Lazar 2016). This was followed by a series of proposals to either redefine statistical significance (Benjamin et al. 2018), remove statistical significance (Amrhein and Greenland 2018; Wasserstein, Schirm, and Lazar 2019), or require researchers to justify their level of significance (Lakens et al. 2018).

A deeper examination of replication failures, such as highly cited positive clinical trials that were contradicted or attenuated in follow-up trials (Ioannidis 2005b), reveals four drivers of the excessive optimism that feed the perception of a crisis. Investigating a multiplicity of research questions on the same data causes selective inference and inflates the evidence. Publishing only successful studies, in particular those with very small sample size, exaggerates the effect size. Reporting p-values without the effect size clouds the actual strength of evidence. Ignoring the distinction between exploratory and confirmatory research also inflates the evidence. Bans on p-values and significance tests have no effect on the economic and career-related incentives to overstate evidence from scientific investigations (Fricker et al. 2019). As Amrhein, Trafimow, and Greenland (2019) aptly stated, "there is no replication crisis if we don't expect replication."

1.2. p-Values Versus Fixed Significance Levels

To understand the controversy around null hypothesis significance testing it is helpful to review its origins (Lehmann 1993; Hubbard and Bayarri 2003; Christensen 2005). The p-value is usually credited to Pearson (1900), but it was Fisher (1925) who developed significance testing as a procedure to validate a hypothesis using proof by contradiction. A null hypothesis is simply a "straw man," such as the hypothesis of no effect, and there is no reference to any alternative hypothesis. An observed effect that is large enough to be highly improbable when there is actually no effect would suggest that the null hypothesis is not valid. The p-value is the probability of seeing an effect as large as or larger than the observed effect assuming that the null hypothesis is true. A small p-value indicates that either we have observed something highly unusual or that the null hypothesis is not true.

Fisher regarded p-values as a measure of evidence against the null hypothesis, the smaller the p-value, the greater the evidence. He wanted scientists to make their own determination as to how small the p-value must be to establish sufficient evidence to disprove the null hypothesis, a threshold referred to as the level of significance (denoted by α). Fisher (1926) suggested the level of 5% (one in twenty), or more rigorous thresholds such as 1% (one in a hundred), depending on the research objectives. Fisher (1956) also advised that "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects (null) hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas." Reaching a threshold does not confer any scientific importance, it is merely a gatekeeper that invites a deeper examination of all the data in the context of the experiment. When the observed data are insufficient to refute the null hypothesis, this is not taken as proof that the null is true.

Consider a sample of patients from a target population that are randomly assigned to receive active treatment or placebo in a 1:1 ratio, resulting in n patients per group and a continuous outcome with common known variance σ^2. The null hypothesis is expressed as no difference in the mean response for the two groups, H_0: $\mu_1 - \mu_2 = 0$. The observed means $\bar{x}_1$ and $\bar{x}_2$ are estimates of the unknown population means μ_1 and μ_2 for the two groups, respectively. The sampling distribution of the observed difference $\delta_{\text{obs}} = \bar{x}_1 - \bar{x}_2$ when the null is true is a normal distribution with mean zero and variance $2\sigma^2/n$, and determines what is unusual when there is no effect. A two-sided significance test is illustrated in Figure 1, which shows the sampling distribution under the null hypothesis for $n = 40$ and $\sigma^2 = 2$. The two-sided p-values corresponding to observed differences of $\delta_{\text{obs}} = 0.6$ and $\delta_{\text{obs}} = 0.8$, are $p = 0.0578$ and $p = 0.0114$, respectively. For example, differences as large as $|\delta_{\text{obs}}| \geq 0.8$ have only a 1.14% probability (1 in 88) and are unexpected if there is no effect, and thus, suggest evidence against H$_0$.

Neyman and Pearson (1928a, 1928b) proposed a decision framework for choosing between two competing hypotheses, the null hypothesis (H_0) and alternative hypothesis (H_1). A scientist can commit two errors in this framework, namely the Type I error of rejecting the null hypothesis of no effect when it is true (a false positive), or the Type II error of failing to reject the null hypothesis when it is false (a false negative). The significance level α in this framework is the probability of a Type I error, and β is the probability of a Type II error. Scientists fix

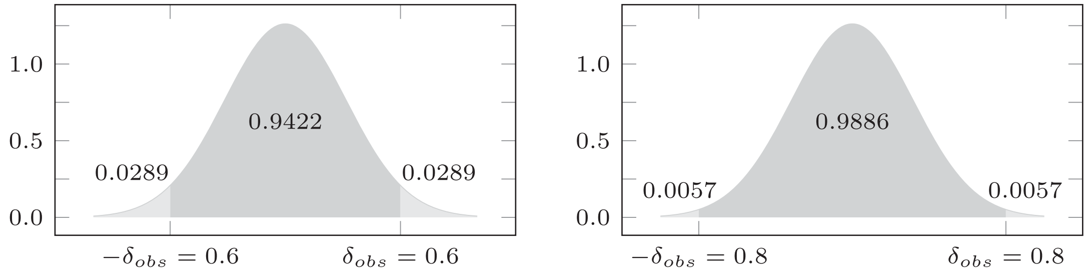

Figure 1. Significance tests for observed effects $\delta_{\text{obs}} = 0.6$ and 0.8, based on $n = 40$ and $\sigma^2 = 2$, result in $p = 0.0578$ and 0.0114, respectively.

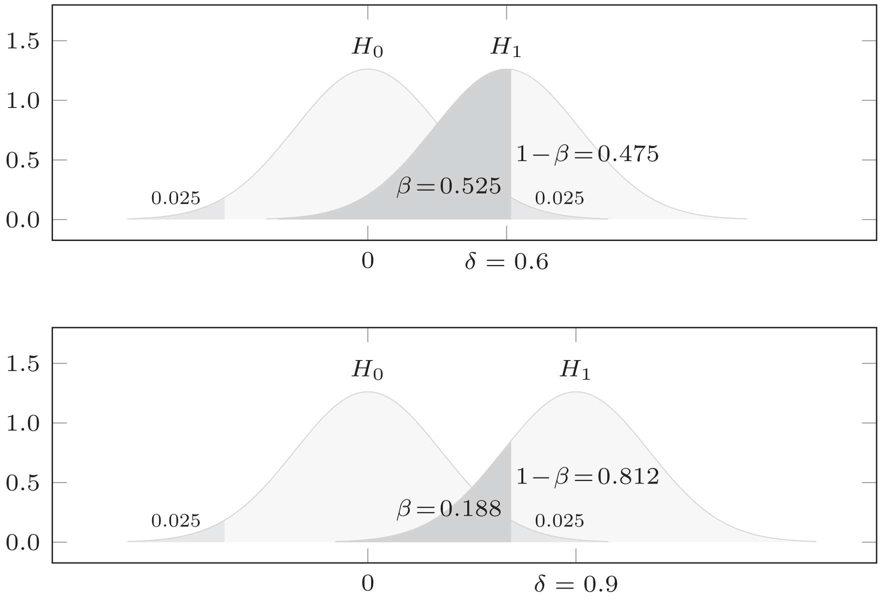

Figure 2. Hypothesis tests based on $n = 40$ per group, assuming $\sigma^2 = 2$ and $\alpha = 0.05$ (2-sided), with 47.5% and 81.2% power to detect minimum differences of $\delta = 0.6$, and 0.9, respectively. H_0 is rejected in favor of H_1 when observed differences fall in the upper or lower rejection region, each with area 0.025.

α in advance and plan experiments with adequate power $1 - \beta$ to correctly reject H_0 for an effect size of interest. The rejection region is defined by $(\alpha, 1 - \beta)$ and determines which observed effects are large enough to reject H_0. The previous example is illustrated in Figure 2 as a decision between H_0: $\mu_1 - \mu_2 = 0$ versus H_1: $\mu_1 - \mu_2 \neq 0$. A sample size of $n = 40$ per group, assuming $\sigma^2 = 2$ and $\alpha = 0.05$, provides 47.5% and 81.2% power to detect minimum effect sizes of $\delta = 0.6$ and 0.9, respectively.

The ubiquitous bright line of $p < 0.05$ originated with the Neyman–Pearson concept of prespecifying a fixed level of significance, an arbitrary choice to control risk of false positives. The relative attitudes toward false positives versus false negatives depend on the consequences of wrong decisions and where one is in the research continuum. The goal of exploratory research is to generate questions, whereas the goal of confirmatory research is to definitively answer a question (De Groot 2014). The impact of too many false positives in the exploratory phase is to waste resources on chasing too many false discoveries, while too many false negatives results in passing over meaningful discoveries too quickly.

1.3. The 5% Level and False Discoveries

The 5% level for "statistical significance" emerged in the 19th century before Pearson or Fisher (Stigler 2008), when economist Francis Edgeworth used values such as 1.5%, 3.25%, and 7% "as a criterion for how firm evidence should be before considering a matter seriously." Although Fisher advised against using the same fixed level of significance in every circumstance, the tables of quantiles in his famous 1925 book offered a limited number of choices due to the absence of a computer to calculate them and the constraints of space on the printed page. The 5% level was a simple way to convey evidence against the null as an unusual finding expected only 1 in 20 times when there is no effect, or alternatively as an observed effect that is approximately two standard deviations from the null effect. Scientists were dependent on printed tables before the era of personal computers, and so 5% became the traditional criterion for firm evidence.

The common use of the 5% significance level led to the practice of publishing only those findings meeting the $p < 0.05$ criterion, resulting in a binary division of all research findings as either positive (significant) or negative (nonsignificant). The drive in academia to "publish or perish" moved the focus away from a continuous measure of evidence to a binary measure of academic success. While the probability of a Type I error is the risk of a false positive finding for an individual study, the risk of publishing false research findings from thousands of studies depends on the prevalence of true effects as well as the power and fixed significance level for each study.

Suppose we investigate 1000 experimental medicines in independent clinical trials each designed with 80 power to detect some specified effect versus placebo. Assume the prevalence of real effects is 10 (with effect size being exactly the same as the one used in the respective power calculation), and a threshold of $p < 0.05$ is used to screen trials for publication, as illustrated in Figure 3. The false discovery proportion (FDP) is the proportion of false positive research findings among all positive research findings (Staquet, Rozencweig, and Von Hoff 1979; Simon 1982; Oakes 1986; Benjamini and Hochberg 1995; Sterne and Smith 2001). Here, the FDP is $\frac{0.05*900}{0.05*900+0.8*100} = 45/(45+80) = 36\%$. The effects of low power, multiplicity, and selective inference can drive the FDP considerably higher. A more rigorous threshold (e.g., $p < 0.001$) might be a tempting solution to reduce the FDP; however, this requires a major increase in sample size to maintain the same level of power.

False discoveries are an issue in drug development (FDA 2017), which consists of laboratory discovery, preclinical safety testing, first in human trials of safety and tolerability (Phase 1), dose ranging trials (Phase 2), and confirmatory testing of safety and efficacy (Phase 3). DiMasi et al. (2010) surveyed 1738 development programs to estimate the probability of success

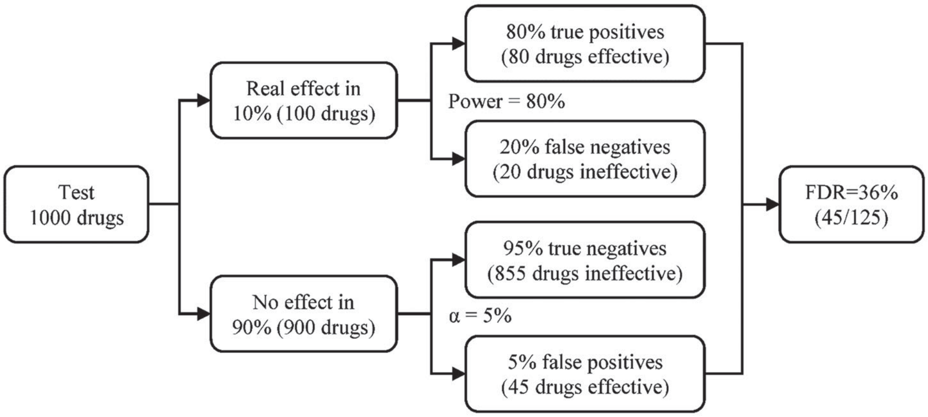

Figure 3. False discovery rate using $p < 0.05$ as the screening threshold to test 1000 experimental drugs, assuming a real effect is prevalent in only 10%, based on clinical trials designed with 80% power.

(POS) for transitions from Phase 1 to Phase 2, Phase 2 to Phase 3, Phase 3 to submission, and submission to approval. They reported the following estimated transition probabilities: $\mathrm{POS}_{1,2} = 71\%$, $\mathrm{POS}_{2,3} = 45\%$, $\mathrm{POS}_{3,S} = 64\%$, $\mathrm{POS}_{S,A} = 93\%$, and thus, $\mathrm{POS}_{1,A} = 19\%$ for the entire program. Hay et al. (2014) and Wong, Siah, and Lo (2019) report similar transition probabilities. Benjamini and Hechtlinger (2014) suggested $1 - \mathrm{POS}_{3,S} = 36\%$ as a plausible estimate of the confirmatory phase FDR, since 36% of the Phase 3 studies failed to confirm the successful results of Phase 2. Conversely, 64% estimates the replication rate for Phase 2 findings.

1.4. The Bayes Factor Alternative

The Bayesian alternative to the p-value as a measure of evidence is the Bayes factor (Jeffreys 1935, 1961; Kass and Raftery 1995; Goodman 1999; Held and Ott 2018). The Bayes factor (BF) measures the evidence provided by the observed data $\boldsymbol{D}$ against the null hypothesis H_0 and in favor of the alternative H_1, and is derived as a ratio of posterior and prior odds. The data $\boldsymbol{D}$ are assumed to have originated under one of the two competing hypotheses according to a probability density $P(\boldsymbol{D}|H_0)$ or $P(\boldsymbol{D}|H_1)$. Each hypothesis has a prior probability $P(H_0)$ and $P(H_1) = 1 - P(H_0)$, and the posterior probabilities after observing the data are $P(H_0|\boldsymbol{D})$ and $P(H_1|\boldsymbol{D}) = 1 - P(H_0|\boldsymbol{D})$. Using Bayes' theorem, the posterior probabilities can be expressed as

$$P(H_i|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|H_i)\,P(H_i)}{P(\boldsymbol{D}|H_0)\,P(H_0) + P(\boldsymbol{D}|H_1)\,P(H_1)}\ (i = 0, 1)\,.$$

From the ratio of posterior probabilities, the posterior odds are obtained as follows

$$\frac{P(H_1|\boldsymbol{D})}{P(H_0|\boldsymbol{D})} = \frac{P(\boldsymbol{D}|H_1)}{P(\boldsymbol{D}|H_0)}\frac{P(H_1)}{P(H_0)}.$$

The Bayes factor is given by $B_{10} = P(\boldsymbol{D}|H_1)/P(\boldsymbol{D}|H_0)$. Since the posterior odds = Bayes factor × prior odds, the Bayes factor is the ratio of the posterior odds of H_1 to its prior odds. Computing the Bayes factor depends on the prior distribution under the alternative hypothesis. If the competing hypotheses are equally probable prior to observing the data, the Bayes factor is simply equal to the posterior odds against H_0 and in favor of H_1 (Kass and Raftery 1995). A Bayes factor of $B_{10} = 100$, for example, suggests that the odds based on the observed data are 100 to 1 against the null hypothesis.

Benjamin and Berger (2019) recommend reporting p-values together with the corresponding Bayes factor (upper) bound (BFB), which is independent of any prior and easy to calculate. They recommend the upper bound $\mathrm{BF} \le \mathrm{BFB} \equiv 1/\left(-e\,p\ln p\right)$ for $p < 1/e$, and 1 otherwise, which is valid under general conditions across a large class of reasonable alternatives. The BFB is the largest odds against the null hypothesis that is consistent with the data. Benjamin and Berger (2019) suggest "converting a p-value into interpretable odds." For example, the Bayes factor bound for a p-value of 0.005 corresponds to odds of at most 13.9 to 1 against the null hypothesis, whereas a p-value of 0.05 corresponds to odds of at most 2.45 to 1 against the null hypothesis.

Gelman and Carlin (2017) cautioned that simply replacing p-values with Bayes factors is not a solution because "the use of Bayes factors for hypothesis testing is also subject to many of the problems of p-values when used for the same purpose." For example, Mandel and Rinott (2009) illustrated the challenges in adjusting for selection bias for both frequentist and Bayesian methods. Goodman (2019) suggested that "the Bayes factor alternative is attractive but may be the bitcoin equivalent; people are not sure what it means, have little clue where it will be accepted, and it has variations in value." Their use, or the use of a Bayes factor bound, to supplement p-values for better interpretation of the evidence is up to the researcher.

2. Using *p*-Values to Judge the Strength of Evidence

This section begins with a note on interpreting p-values as a continuous measure of evidence. This is followed by a four-part guide for using p-values to judge the strength of evidence without distorting the expectations for replication. Of course, judging the evidence requires looking beyond p-values to examine the totality of all the data in the context of the design and the quality and consistency of a study.

2.1. p-Values Measure Evidence on a Log Scale

Evidence is information indicating the degree to which a proposition is valid. p-Values are random variables that measure

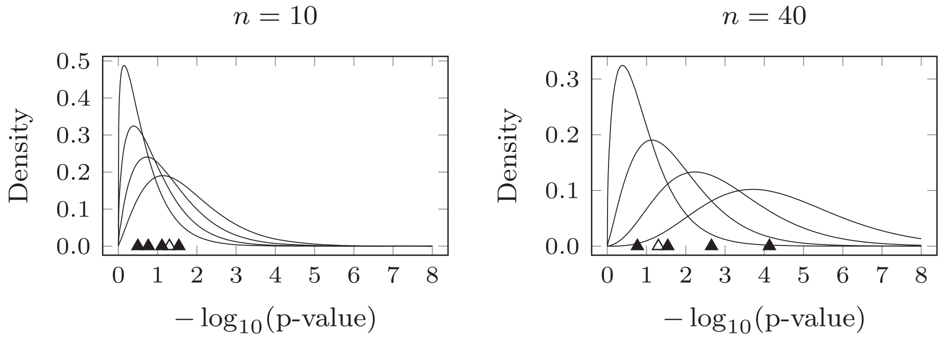

Figure 4. The probability densities for $-\log_{10}(p\text{-value})$ corresponding to a two-sample test for four different alternative effect sizes (0.3, 0.6, 0.9, 1.2), based on $n = 10$ and 40, and $\sigma^2 = 2$. The open triangle represents the threshold of $p < 0.05$, and closed triangles represent the median p-value for each effect size.

evidence against the null hypothesis. The sampling distribution of the p-value under the null hypothesis is well known to be uniform over the unit interval [0,1] for a single point null hypothesis and continuous test statistics. Under the alternative hypothesis, the sampling distribution of p-values is highly skewed and depends on both the sample size and effect size. Lambert and Hall (1982) showed for large samples that the sampling distribution of the p-value under the alternative hypothesis is approximately lognormal, or equivalently that $-\log(p)$ is asymptotically normal. The behavior of p-values when the null hypothesis is false is essential to understanding how to judge the evidence they convey.

Hung et al. (1997) derived the probability density of the p-value under the alternative for the simple case of a large-sample test of two means for given values of δ, σ, and n as $g_p(p) = \phi(Z_p - \sqrt{n/2}\delta/\sigma)/\phi(Z_p)$ for $0 < p < 1$, where ϕ is the standard normal density, Φ is the cumulative normal distribution, and $Z_p = \Phi^{-1}(1-p)$. The density of $y = -\log_{10}(p)$ is thus

$$f_\delta(y) = \left(10^{-y}\right)\phi(Z_{10^{-y}} - \sqrt{n/2}\delta/\sigma)/\phi(Z_{10^{-y}}) \text{ for } y > 0,$$

and is displayed in Figure 4 for given values of δ, σ, and n. The distribution of $-\log_{10}(p)$ is increasingly bell-shaped for larger effect sizes corresponding to more extreme departures from the null, and for larger sample sizes. The concentration of extremely small p-values increases as the effect size increases, and for well-powered studies, it is relatively easy to observe p-values well below 0.05 when the alternative is true. In fact, the median p-value for a study planned with 90% power against any alternative is 0.001, and thus, $p = 0.05$ might even be considered in this case negative evidence of a true difference since it would be relatively unusual to observe a $p \geq 0.05$ (Hung et al. 1997).

p-Values are naturally interpreted on a log scale since $-\log_{10}(p)$ is approximately asymptotically normal. The p-value can be expressed as $p = \text{c} \times 10^{-k}$ so that $-\log_{10}(p) = -\log_{10}(c) + k$, where c is a constant and k is an integer, which implies that only the magnitude k measures the actual strength of evidence (Boos and Stefanski 2011). For example, a scientist might assume the evidence associated with $p = 0.02$ is twice as strong relative to $p = 0.04$, since it is half the size. However, this is misleading because they both have the same magnitude on the log scale ($k = 2$). This would suggest that $p = 0.01(k = 2)$ could be interpreted as twice the evidence of $p = 0.10(k = 1)$. Working with raw p-values or log-transformed p-values is up to the researcher, what is important is understanding their behavior under the alternative. It is worth noting the Bayes factor (upper) bound, $\text{BFB} \equiv 1/\left(-e\, p \ln p\right)$, can also be expressed as a function of k as $\text{BFB} = 10^k/[ec(k - \log_{10}(c))\log_e(10)]$. This is perhaps why Jeffreys (1961) suggested Bayes factors could be interpreted on the $\log_{10}$ scale.

Reproducibility probabilities are another way of calibrating the strength of evidence measured by p-values. The reproducibility probability (RP) is the probability of replicating the statistically significant results of an initial study in a subsequent identical study conducted under the same conditions (Goodman 1992; Senn 2002). It is denoted by $\text{RP} = P(p_{\text{new}} < 0.05)$, where p_{new} represents the p-value from an independent replication of the original study. An estimate of RP is computed by estimating the power of the subsequent trial conditioned on the observed data from the initial trial (Goodman 1992; Shao and Chow 2002; Boos and Stefanski 2011). The estimated reproducibility probability for the simple case of a two-sample test of means with known common variance σ^2 is given by

$$\begin{aligned}\widehat{\text{RP}} &= \text{power}(z_{\text{obs}}) = P\left(\text{reject } H_0 \,|z_{\text{obs}}\right)\\ &= 1 - P\left(Z < z_{\alpha/2} - z_{\text{obs}}\right) + P\left(Z < -z_{\alpha/2} - z_{\text{obs}}\right),\end{aligned}$$

where $z_{\text{obs}} = \delta_{\text{obs}}/\sqrt{2\sigma^2/n}$ based on the observed data from the first study.

Since $|z_{\text{obs}}| = -\Phi^{-1}(p_{\text{obs}}/2)$, the estimated reproducibility probability is simply a monotone function of the observed p-value. However, reproducibility probabilities do provide another way to calibrate p-values and are often much lower than scientists would expect. For example, the estimated reproducibility probability given an initial finding of $p_{\text{obs}} = 0.05$ indicates that there is only a 50% probability that $p_{\text{new}} < 0.05$, suggesting that $p_{\text{obs}} = 0.05$ is relatively weak evidence. In contrast, the estimated reproducibility probability given an initial finding of $p_{\text{obs}} = 0.001$ indicates that there is a 90% probability that $p_{\text{new}} < 0.05$, suggesting that $p = 0.001$ is relatively strong evidence. Estimated reproducibility probabilities for a range of p-values are given in Table 1. Reproducibility probabilities do not calibrate evidence below 0.001 as well as $-\log_{10}(p\text{-value})$ and the Bayes factor bound.

2.2. *The Force of a* p*-Value Depends on Effects of Interest*

Kempthorne (1976) explained that "the force of an observed p-value depends on the distribution of the p-value under

Table 1. *p*-Values, Bayes factor bounds, and estimated reproducibility probabilities.

p-Value	0.10	0.05	0.01	0.001	0.0001	0.00001
BFB	1.6	2.5	8	53	399	3195
$\widehat{RP}$	0.38	0.50	0.73	0.91	0.97	0.99
$-\log_{10}$(*p*-value)	1	1.3	2	3	4	5

alternative hypotheses which are worth entertaining." This principle is illustrated in Figure 5, which shows the distribution of $-\log_{10}(p)$ under two different effect sizes for $n = 40$ and $\sigma^2 = 2$. This design provides 88% power to detect effects as small as $\delta = 1$, and when the true effect is $\delta = 1$ the distribution of p-values is characterized by strong evidence against the null. For example, 80% of the p-values are less than $0.01(k > 2)$, 53% of the p-values are less than $0.001(k > 3)$, and 29% of the p-values are less than $0.0001(k > 4)$. The same design provides only 47% power to detect effects as small as $\delta = 0.6$, and this effect size results in a distribution of p-values that represents relatively weak evidence against the null. In this case, only 33% of the p-values are less than 0.01 and only 11% are less than 0.001.

Betensky (2019) proposed judging the strength of a p-value by the scope of meaningful effects that it supports for a given study design. Instead of using $p < 0.05$ to rule out no effect, she derives a p-value threshold for concluding a meaningful effect. Solve $p = 2[1 - \Phi\left(Z \le \sqrt{n/2}\delta_{\text{obs}}/\sigma\right)]$ for the observed effect corresponding to the observed p-value $\delta_{\text{obs}} = Z_{p/2}\sigma\sqrt{2/n}$. Substitute into the lower $(1-\alpha)$ 100% confidence limit $\delta^* = \delta_{\text{obs}} - Z_{\alpha/2}\sigma\sqrt{2/n}$ to obtain $\delta^* = (Z_{p/2} - Z_{\alpha/2})\sigma\sqrt{2/n}$. Since smaller p-values result in larger values of δ^*, define p^* as the value of p for which δ^* represents a meaningful effect size. Reject the null hypothesis of no effect in favor of a meaningful effect if and only if $p < p^*$, or equivalently, when the lower limit of a 95% confidence interval exceeds the meaningful effect.

For example, the design of Figure 5 is based on $n = 40$ and $\sigma^2 = 2$, with a minimum meaningful effect $\delta = 1$. As one would expect, excluding a zero effect $\delta^* = 0$ with 95% confidence requires a p-value below $p^* = 0.05$. Ruling out small nonnull effects requires stronger evidence. To conclude a nonnull effect size at least as large as $\delta^* = 0.42$ with 95% confidence would require a p-value below $p^* = 0.001(k > 3)$, observed 53% of the time with this design. Evidence of an effect at least as large as $\delta^* = 0.61$ would require a p-value below $p^* = 0.0001(k > 4)$, observed 29% of the time. Evidence of an effect at least as large as $\delta^* = 0.78$ would require a p-value below $p^* = 0.00001(k > 5)$, observed 13% of the time. Finally, evidence of an effect as large as $\delta^* = 1$ would require a p-value below $p^* = 0.0000001(k > 7)$, observed just 2% of the time.

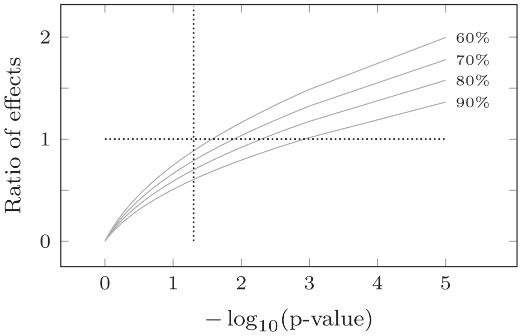

Figure 6. Ratio of the observed to expected effect size ($\delta_{\text{obs}}/\delta_e$), for observed p-values corresponding to designs based on power of 60–90% with $\alpha = 0.05$. The horizontal dashed line represents a ratio of 1, and the vertical dashed line represents the threshold of $p < 0.05$.

Hung and O'Neill (2003) expressed the relationship between the observed p-value and the ratio of the observed effect size δ_{obs} to the expected (hypothesized) effect size δ_e assumed during the design stage as $\delta_{\text{obs}}/\delta_e = z_p/(z_\alpha + z_\beta)$, where z_p is the standardized difference associated with the observed p-value. Figure 6 shows the observed effect is smaller than the expected effect when the observed p-value is not significant (i.e., $p > \alpha$) or is near the threshold (i.e., $p \cong \alpha$). For a study planned with 90% power to detect an expected effect size δ_e for $\alpha = 0.05$ (2-sided), the observed effect will be 60% of the expected effect for $p = 0.05$, 50% of the expected effect for $p = 0.10$, and 40% of the expected effect for $p = 0.20$. Conversely, the observed effect associated with $p = 0.001$ will be 100% of the expected effect. Marginal evidence in a well-powered study indicates an effect size smaller than anticipated.

While researchers almost universally adopt the hypothesis of no effect as their null hypothesis, it is important to realize that quite often, as is the case in drug development, we know there is some effect and the magnitude of the effect is the key question. By the start of Phase 2, we usually have good reason to believe the drug has some effect. However, if the effect is smaller than we think, this changes the distribution of p-values. When results are not confirmed in larger Phase 3 trials, it is often because the effect is smaller than was predicted, and not because it was zero.

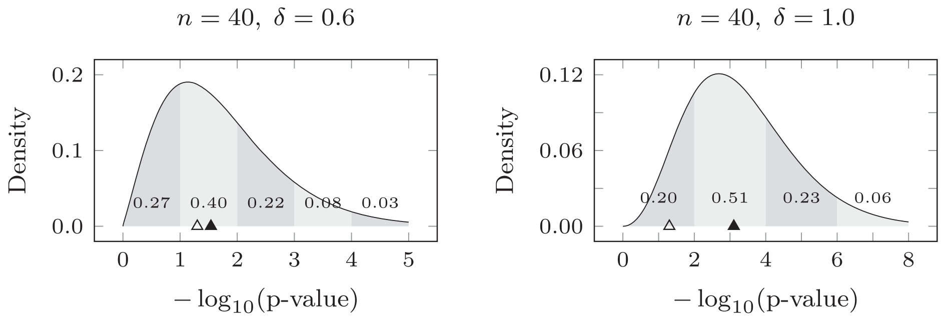

Figure 5. The distribution of $-\log_{10}$(*p*-value) based on $n = 40$ and $\sigma^2 = 2$ under two alternative effect sizes (0.6, 1.0). The open triangle represents the threshold of $p < 0.05$, and closed triangle represents the median *p*-value.

2.3. *Multiplicity and Selective Inference Inflate the Evidence*

Tukey (1977) cautioned against multiplicity and selective inference. He described this as “asking multiple questions and concentrating on the most favorable answers.” Clinical trials, for example, often evaluate multiple endpoints between multiple treatment groups among multiple subgroups of patients, resulting in a potentially large number of significance tests. Multiplicity refers to the risk of at least one false positive when testing multiple research questions. Selective inference is the biased practice of choosing the primary research question after the experiment is completed from the most promising finding among the data. While multiplicity inflates the Type I error rate, the bias of selective inference affects p-values, point estimates, confidence intervals, as well as Bayesian statistics. Both phenomena exaggerate the evidence and increase the risk of initial research findings that are contradicted in future research.

Multiplicity inflates the probability of making at least one Type I error when performing a series of m independent hypothesis tests, each at the significance level α. The family-wise error rate (FWER) is the probability of at least one Type I error in a single study with a family of m hypothesis tests, and is equal to $1-(1-\alpha)^m$ in the case of independent tests. For example, a clinical study designed to test 10 hypotheses arising from multiple treatment comparisons, endpoints, or populations, would result in a FWER as large as 40%, which is concerning given that most studies evaluate much larger numbers of hypotheses. Whatever the FWER, if $p < 0.05$ is used to screen for real effects when the actual rate of false positives is much higher than $\alpha = 0.05$, then the risk of false discoveries rises dramatically, and is exacerbated further for underpowered studies, as shown in Table 2.

Bretz and Westfall (2014) illustrated the impact of selective inference when selecting the subgroup with the best effect observed from among four nonoverlapping subgroups in an initial study, and conducting a subsequent study focused on the chosen subgroup as the new study population. This scenario is often encountered in clinical drug development, where the initial study is an exploratory study at the end of Phase 2, and a promising subgroup finding is the basis for a subsequent confirmatory Phase 3 study. Their simulations showed that the effect sizes observed in the second study were on average much smaller than the effect sizes observed in the initial study. This shows the impact of selective inference not only on replicability but also on estimating expected effects for a new study based on effect sizes observed in an initial study that are exaggerated.

Benjamini (2019) warned that “selective inference is the silent killer of replicability,” because it inflates the evidence, making it appear stronger than it actually is unless the nature of the selection is reported and adjusted for in the analysis. It is important to realize that multiplicity and selective inference can impact any statistical method or tool, and not just p-values. For example, Benjamini (2019) notes that “adjusting for selection in estimation and confidence intervals is rarely practiced, even when done for testing.” Greenland (2019) notes that “curtailing selection biases will still require additional drastic measures rather than just a change in inferential method.” Bauer (2017) also cautions that “multiplicity seems to remain a serious challenge for any type of statistical inference.”

Table 2. FDR by power and significance levels inflated by multiple tests.

Power	1 test FWER = 0.05	2 tests FWER = 0.10	4 tests FWER = 0.185	10 tests FWER = 0.40
30%	60%	75%	85%	92%
50%	47%	64%	77%	88%
80%	36%	53%	68%	82%
90%	33%	50%	65%	80%

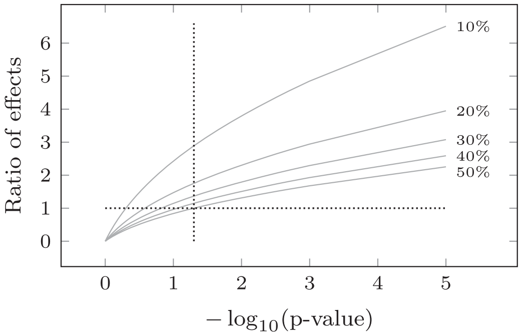

Figure 7. Ratio of the observed to expected effect size ($\delta_{\text{obs}}/\delta_e$), for observed p-values corresponding to designs based on power of 10–50% with $\alpha = 0.05$. The horizontal dashed line represents a ratio of 1, and the vertical dashed line represents the threshold of $p < 0.05$.

2.4. *Publication Bias and Low Power Exaggerate the Effect Size*

Publication bias (Sterling 1959; Rosenthal 1979) is a type of selection bias in which only those studies that reach the conventional threshold for statistical significance of $p < 0.05$ are published. This results in a higher rate of false discoveries in the literature. Lane and Dunlap (1978) showed that estimating the effect size using only published studies can considerably overestimate the true effect size, and lead to an under-powered follow-up study. Advances in clinical trial registration, for example, are reducing the selective publication of research outcomes (Miller et al. 2017). The International Committee of Medical Journal Editors (ICMJE) requires the registration of a clinical trial in a public registry at or before the time of first patient enrollment for the results of the completed trial to be considered for publication. The sharing of protocols, statistical analysis plans, and patient data will have an even bigger impact on the transparency of research findings (Rockhold et al. 2019). These principles of open science should be expanded to other scientific disciplines.

Small studies with low power occur in many areas of science and are prone to exaggerated observed effect sizes. Pereira, Horwitz, and Ioannidis (2012) reviewed large treatment effects reported in 85,002 comparative analyses of medical interventions in the Cochrane Database, and found that most large effects are observed in small studies and are usually attenuated in follow-up trials. The Open Science Collaboration (2015) conducted a large project to replicate findings in psychological science, and reported that “replication effects were half the magnitude of original effects.” The relationship between the observed p-value and the ratio of the observed effect relative to the expected effect $\delta_{\text{obs}}/\delta_e = z_p/(z_\alpha + z_\beta)$ is illustrated for small and noisy studies with low power in Figure 7. Significant

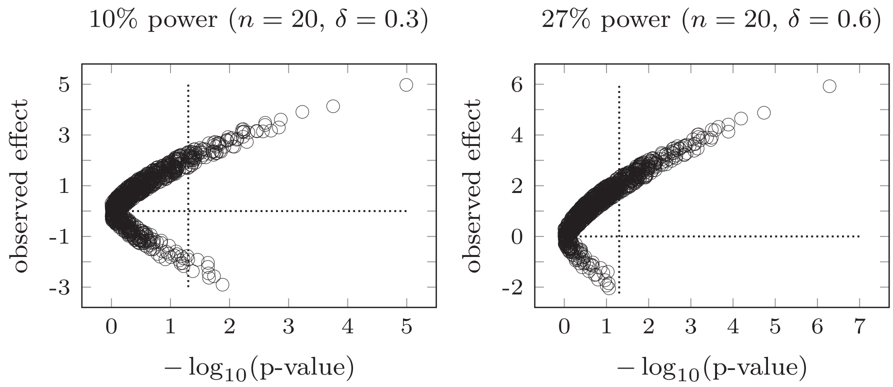

Figure 8. The standardized observed effect size $\sqrt{n/2}\delta/\sigma$ (*z*-score) versus the corresponding observed *p*-value for 1000 repeated trials with small samples ($n = 20$) and low power to detect small effects ($\delta = 0.3,\ 0.6$), with $\alpha = 0.05$ and $\sigma^2 = 2$. The horizontal dashed line represents zero effect, and the vertical dashed line represents the threshold of $p < 0.05$.

($p < \alpha$) effects observed in a study with low power, for example, below 50%, are increasingly exaggerated relative to what was expected as the power decreases.

The behavior of observed effects in studies with extremely low power is also illustrated in Figure 8, which presents 1000 simulated two-sample tests for small effects ($\delta = 0.3$ and $\delta = 0.6$) based on small samples of $n = 20$. Statistically significant effects ($p < 0.05$) are to the right of the vertical line, and may be as much as two to six times larger than expected depending on the magnitude of the *p*-value. Significant effects may even occur in the wrong direction. Gelman and Carlin (2014) described these types of misleading results as Type M (magnitude) errors and Type S (sign) errors.

The observed effect size in a given study may be prone to exaggeration due to low power or publication bias. Without adjustment, the postulated effect size for a follow-up study may result in an underpowered design, resulting in a diminished or vanishing effect. For publication bias, Hedges (1984) proposed a shrinkage method for correcting the estimated effect size. For studies with low power, discounting is a common adjustment (Chuang-Stein and Kirby 2014). Kirby et al. (2012) recommended at least a 10% discount to the Phase 2 estimate of treatment effect when planning a Phase 3 study. Wang, Hung, and O'Neill (2006) recommended using the lower endpoint of the observed 95% confidence interval from Phase 2 when planning Phase 3.

2.5. Exploratory Versus Confirmatory Evidence

Tukey (1977, 1980) described exploratory research as a flexible attitude that transforms an abstract idea into a well-formulated question that is ready to be confirmed in a more rigorous study. Exploratory research is focused on inquiry, data exploration, hypothesis generation, modeling, and estimation. Hypothesis tests in this setting are often data-driven and not predefined. Drug development, for example, rarely begins with a focused confirmatory question. Instead, the confirmatory question is shaped by a series of exploratory studies. Exploratory objectives include topics such as establishing whether the new drug has any measurable effect, identifying biomarkers that predict response, estimating a plausible range of anticipated effects, and characterizing the shape of the dose–response curve. The hypotheses and analysis methods for confirmatory testing are specified in advance, based on the learnings of the exploratory phase (ICH 1998).

The regulated field of medical research operates under a well-developed body of methods that maintains a distinction between exploratory and confirmatory research (Wellek 2017). Many other disciplines are not bound by such a framework. Wagenmakers et al. (2012), for example, characterized psychological research as a discipline that routinely extracts confirmatory conclusions from exploratory findings. They note that "almost without exception, psychologists do not commit themselves to a method of data analysis before they see the actual data." This is confounded by condensing the research continuum into a small single study. This has some similarity to the early years of clinical research in which clinical trials were routinely conducted without a protocol or an analysis plan (Temple 2005). Although psychological research is changing, the "one-and-done tradition of theory confirmation by a single small randomized trial, while weakening, is still dominant" (Goodman 2019). Bishop (2019) also describes progress but notes the need to bring publication bias, low power, multiplicity, and selective inference under control through improved rigor and open science.

The appropriate standard of evidence for exploratory research depends on the volume of questions, the prevalence of interesting discoveries, and the tolerance for false discoveries. Traditional methods for addressing multiplicity provide strong control of the familywise error rate at the expense of losing power and missing real differences. Benjamin et al. (2018) suggested controlling the FDR in exploratory research by requiring a higher level of evidence, namely $\alpha = 0.005$. If π represents the proportion of real effects (false null hypotheses), then for a large number of questions $FDR \approx \alpha(1 - \pi)/[\alpha(1-\pi) + (1-\beta)(\pi)]$. Testing at $\alpha = 0.005$ with 80% power when the prevalence of real effects is $\pi = 10\%$ would limit the FDR to 5%, but at the expense of a 70% increase in sample size and fewer studies within existing budgets. Alternatives include softer thresholds, or reduced power, or both.

Exploratory research with big data is often characterized by the selection of a subset of variables or features, from hundreds of thousands or millions of candidates, that are associated with an outcome or trait of interest. Candes et al. (2018) described

Table 3. Lack of replication in all-cause mortality results between PRAISE-I and PRAISE-II.

Amlodipine studies	Placebo	10 mg	HR (95% CI)	p-Value
PRAISE-1 overall ($N = 1153$)	38.3% (223/582)	33.3% (190/571)	0.84 (0.69, 1.02)	0.07
PRAISE-1 ischemic ($N = 732$)	40.3% (149/370)	40.1% (145/362)	1.02 (0.81, 1.29)	0.87
PRAISE-1 nonischemic ($N = 421$)	34.9% (74/212)	21.5% (45/209)	0.54 (0.37, 0.79)	<0.001
PRAISE-2 ($N = 1654$)	31.7% (262/827)	33.6% (278/827)	1.09 (0.92, 1.29)	0.33

this as "panning for gold." Berger (2012) and Benjamin et al. (2018) noted that early genome-wide association studies "almost universally failed to replicate (estimates of the replication rate are as low as 1%) because they were doing extreme multiple testing at nonextreme p-values." This led researchers to adopt $p \leq 5 \times 10^{-8}$ as a genome-wide significance threshold (Jannot, Ehret, and Perneger 2015). The combination of average power and strong evidence of association improved the replication of identified associations in subsequent studies. Methods for controlling the false discovery rate are also useful when screening numerous endpoints, as they offer a compromise by reducing the risk of too many false discoveries without overlooking too many real differences (Benjamini et al. 2001).

A well-known standard for confirmatory evidence is the requirement of two successful (two-sided $p < 0.05$, with both estimates in the favorable direction) adequate and well-controlled studies for new drug approval (FDA 1998). Adopting 0.05 was not a specific regulatory decision, it was simply the "conventional" statistical practice followed by scientists at the time (Kennedy-Shaffer 2017). Requiring two replicate studies below $p < 0.05$ limits the false approval rate to no more than $2 \times (1/40) \times (1/40) = 0.00125$. This standard of evidence is an order of magnitude higher than a decision based on one study with $p < 0.05$ and illustrates the value of independent replication. The replicate study does not have to be identical and may employ a different design or population. Regulatory approval is also possible based on a single study with the same level of evidence of as the two-study paradigm, namely $p < 0.00125$ (Fisher 1999; CPMP 2001). This is not an absolute requirement, and a single study "in the neighborhood of $p = 0.001$" might open the door for discussion (Temple 2005).

Pocock, McMurray, and Collier (2015) offered guidelines for "using p-values wisely to assess the strength of evidence" in the context of randomized clinical trials. They recommended using the p-value together with an estimate of the treatment effect and its 95% confidence interval to assess the magnitude of the effect, the degree of uncertainty, and the strength of evidence that the effect is genuine. Pocock and Stone (2016a) recommended $p < 0.001$ when proof beyond reasonable doubt is required. They also offer guidelines for interpreting evidence when the primary outcome of a trial fails to achieve statistical significance, noting that it is "hard to think of an example in which an apparent benefit in a subgroup in a trial with a negative outcome has led to a confirmation in a subsequent trial (Pocock and Stone 2016b)."

3. Case Study: Heart Failure

Disappointing results in Phase 3 that fail to confirm the promising results previously observed in Phase 2 are common in the challenging field of heart failure drug development, where there continues to be a high unmet need for effective therapies. Vaduganathan, Butler, and Gheorghiade (2016) noted that publication bias limits the availability of data from failed programs in heart failure drug development. The following section presents two case studies of how the perceived strength of evidence plays a critical role in making decisions based on promising initial results.

3.1. The Praise Trials

The PRAISE-1 trial randomized 1153 patients to receive amlodipine 10 mg or placebo, stratified by ischemic or nonischemic heart failure (Packer et al. 1996). The prospectively defined primary outcome was the combination of major morbidity or mortality from any cause. The trial was designed with 90% power to detect a 25% reduction in the primary endpoint. The observed reduction in morbidity/mortality was only 9% ($p = 0.31$), with a 16% ($p = 0.07$) reduction in mortality. The nonischemic stratum was more compelling, with a 31% reduction in morbidity/mortality ($p = 0.04$) and a 46% reduction in mortality ($p < 0.001$). To confirm the mortality finding in this stratum, 1654 patients with nonischemic heart failure were enrolled in PRAISE-2 (Packer et al. 2013). The primary outcome was mortality from any cause, and the trial was designed with 90% power to detect a 25% reduction in mortality. The results failed to replicate the mortality finding, and the two studies are summarized in Table 3.

The failure of the PRAISE-2 trial illustrates the impact of power, multiplicity, and selective inference on the perceived level of evidence. PRAISE-1 suggested a reduction in mortality in the nonischemic stratum, which was a finding in one of eight different subgroups examined within a negative trial. PRAISE-2 was designed with 90% power to confirm the reduction in mortality, whereas the first trial was not powered to examine mortality within a subgroup of only 36.5% of the patients. The initial finding of a 46% reduction in mortality in the nonischemic subgroup was exciting in the context of high unmet medical need, but was likely an exaggerated effect subject to selection bias. In planning the second trial, the investigators adjusted for the inflated effect by powering the second trial to detect a smaller reduction. Nevertheless, the small p-value in PRAISE-1 overstated the true strength of evidence and lacked adjustment for the number of subgroups explored within a trial that did not meet its primary endpoint.

3.2. The Elite Trials

The ELITE-1 trial randomized 722 elderly patients with chronic symptomatic heart failure, untreated with ACE inhibitors, to receive either the ARB losartan 50 mg once daily or the ACE

Table 4. Lack of replication in all-cause mortality results between ELITE-I and ELITE-II.

Studies	Captopril	Losartan	Estimate (95% CI)	*p*-Value
ELITE-1 (*N* = 722)	8.6% (32/370)	4.8% (17/352)	0.46* (0.05, 0.69)	0.035
ELITE-2 (*N* = 3152)	15.9% (250/1574)	17.7% (280/1578)	1.13** (0.95, 1.35)	0.16

*Risk reduction based upon Mantel–Haenszel adjusted (for age category) relative risk estimate.
**Hazard ratio based upon Cox regression. Final analysis adjusted, after several interim analyses, to 0.043 (two-sided) and 95.7% CI.

inhibitor captopril 50 mg three times daily for 48 weeks (Pitt et al. 1997). The prospectively defined primary assessment was a measure of renal dysfunction to evaluate tolerability, the secondary assessment was the composite of death and/or hospitalized heart failure, as well as five other assessments, including mortality from any cause. The study was designed for an analysis of renal dysfunction, and was not powered for an analysis of mortality or the composite of death and/or hospitalized heart failure. However, the observed risk reduction in death and/or hospitalized heart failure was 32% $\left(p = 0.075\right)$ and was driven by an unexpected 46% ($p = 0.035$) reduction in mortality from all causes.

The ELITE-2 trial was designed to confirm the finding of a reduced risk of mortality from any cause with losartan compared to captopril (Pitt et al. 2000). A total of 3152 elderly patients with symptomatic heart failure were randomized to either losartan 50 mg once daily or captopril 50 mg three times daily, and followed for a median time of 1.5 years. The primary and secondary assessments were prospectively defined as mortality from any cause and the composite of sudden death or resuscitated cardiac arrest. The study was designed with 90% power to detect a relative 25% difference between treatments. The results did not confirm the mortality benefit of losartan observed in ELITE-1, and actually indicated a trend that suggested the possibility of harm. There was also no significant difference in composite of sudden death or resuscitated cardiac arrest. The results of these two studies are summarized in Table 4, and note that risk reductions were reported for ELITE-1, whereas hazard ratios were reported for ELITE-2.

The failure of the ELITE-2 trial also illustrates the impact of power, multiplicity, and selective inference on the perceived level of evidence. While the initial finding of a 46% risk reduction in mortality ($p = 0.035$) seemed exciting, it is important to view the evidence in a wider context. The investigators focused selectively on one of multiple secondary endpoints that the study was not designed to evaluate. The power of an unplanned analysis of mortality in ELITE-1, based on only one tenth of the events observed in the ELITE-2, would be extremely low and prone to an exaggerated effect. The evidence is less compelling when viewed in the context of multiplicity, selective inference, and the limited length of follow-up.

4. Conclusion

It is not realistic to expect every interesting experimental finding will be replicated, but it is natural for every scientist to hope for it. When the apparent evidence is inflated by factors such as multiplicity, selective inference, bias, and low power, the hope for replication can develop into an unrealistic expectation. This exacerbates the disappointment when promising initial findings are not confirmed in a follow-up study. A view of the data in the transparent light of factors that may distort the level of evidence is the best remedy for the excessive optimism that can lead to heightened disappointment. I do not believe there is a replication crisis, but there are misunderstandings, and there is always an opportunity to strive for the best scientific practice in judging the strength of the data. As Bauer (2017) asked rhetorically in his commentary on the *p*-value controversy, “is there any statistical concept without the potential for being misused?”

A proposal for the best practices for using *p*-values to judge the persuasiveness of the evidence includes the following steps. Illustrate the force of the *p*-value by the scope of nonnull effects that it excludes for a given study design, and the range of meaningful benefits that it supports. Apply the appropriate adjustments for the effects of multiplicity and selective inference. Discount the postulated effect size when planning a follow-up study to account for the effects of publication and selection bias and low power. Interpret the *p*-value as a continuous measure of evidence on a log scale as $-\log_{10}(p)$, where only the magnitude of k is reliably determined as the measure of evidence (Boos and Stefanski 2011). Finally, differentiate between exploratory and confirmatory research and apply the appropriate standards of evidence. Proof beyond a reasonable doubt is not required in every setting, but it is critical to be transparent about the true strength of evidence, and to avoid overstating the evidence through omission of key details.

These key principles are not part of the curriculum taught to students and scientists. Maurer et al. (2019) proposed a framework for updating the introductory statistics curriculum to include greater emphasis on the best practices for the use and interpretation of *p*-values for statistical inference. This has the potential to impact, for example, nearly one million students in the United States who enroll each year in introductory and upper level undergraduate statistics courses (Blair, Kirkman, and Maxwell 2018). Some of these students will go on to become scientists and policy makers. Statisticians likewise have an opportunity to step forward and assert their leadership as they collaborate with multidisciplinary research teams to develop the best designs and make decisions based on evidence (Gibson 2019). As we collaborate, so must we continue to teach and educate our partners in science.

Making decisions invariably leads to a judgment of whether the evidence is persuasive enough to support an affirmative decision. Scientific conclusions and business or policy decisions should not be based simply on whether a *p*-value passes a specific threshold. However, a *p*-value can serve as “a first line of defense against being fooled by randomness,” like a gatekeeper that opens the door to a deeper examination of the data in terms of the benefits and risks (Benjamini 2016). The strength of this “first line of defense” depends on whether the research is exploratory or confirmatory and the consequences associated with making the wrong decision. Perhaps this is why Fisher (1956) cautioned against using a fixed level of significance "in all circumstances."

Acknowledgments

The author thanks the editor, associate editor, and both reviewers for their careful review and helpful suggestions which significantly improved this article. The author is also grateful to his colleagues Paul Gallo and Ekkehard Glimm for their many insights and suggestions.

References

Amrhein, V., and Greenland, S. (2018), "Remove, Rather Than Redefine, Statistical Significance," *Nature Human Behaviour*, 2, 4–4. [2]

Amrhein, V., Trafimow, D., and Greenland, S. (2019), "Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis If We Don't Expect Replication," *The American Statistician*, 73, 262–270. [2]

Bauer, P. (2017), "Comment on 'A Critical Evaluation of the Current *p*-Value Controversy," *Biometrical Journal*, 59, 873–874. [7,10]

Benjamin, D. J., and Berger, J. O. (2019), "Three Recommendations for Improving the Use of *p*-Values," *The American Statistician*, 73, 186–191. [4]

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2018), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6–10. [2,8,9]

Benjamini, Y. (2016), "It's Not the *p*-Values' Fault," *The American Statistician*, available at *http://amstat.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108?scroll=top*. [10]

——— (2019), "Selective Inference: The Silent Killer of Replicability," in *The Henry L. Rietz Lecture, ASA Joint Statistical Meetings*, Denver, Colorado. [7]

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001), "Controlling the False Discovery Rate in Behavior Genetics Research," *Behavioral Brain Research*, 125, 279–284. [9]

Benjamini, Y., and Hechtlinger, Y. (2014), "Discussion: An Estimate of the Science-wise False Discovery Rate and Applications to Top Medical Journals by Jager and Leek," *Biostatistics*, 15, 13–16. [4]

Benjamini, Y., and 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, 289–300. [3]

Berger, J. O. (2012), "Reproducibility of Science: *p*-Values and Multiplicity," in *Eighth International Purdue Symposium on Statistics*, available at *http://www.stat.purdue.edu/symp2012/docs/Purdue_Symposium_2012_Jim_Berger_Slides.pdf*. [9]

Betensky, R. A. (2019), "The *p*-Value Requires Context, Not a Threshold," *The American Statistician*, 73, 115–117. [6]

Bishop, D. (2019), "Rein in the Four Horsemen of Irreproducibility," *Nature*, 568, 435–435. [8]

Blair, R., Kirkman, E. E., and Maxwell, J. W. (2018), *Statistical Abstract of Undergraduate Programs in the Mathematical Sciences in the United States: Fall 2015 CBMS Survey*, Washington, DC: Mathematical Association of America. [10]

Boos, D. B., and Stefanski, L. A. (2011), "*p*-Value Precision and Reproducibility," *The American Statistician*, 65, 213–221. [5,10]

Bretz, F., and Westfall, P. H. (2014), "Multiplicity and Replicability: Two Sides of the Same Coin," *Pharmaceutical Statistics*, 13, 343–344. [7]

Candes, E. J., Fan, Y., Janson, L., and Lv, J. (2018), "Panning for Gold: Model-X Knockoffs for High-Dimensional Controlled Variable Selection," *Journal of the Royal Statistical Society*, Series B, 80, 551–577. [8]

Chuang-Stein, C., and Kirby, S. (2014), "The Shrinking or Disappearing Observed Treatment Effect," *Pharmaceutical Statistics*, 13, 277–280. [8]

Christensen, R. (2005), "Testing Fisher, Neyman, Pearson, and Bayes," *The American Statistician*, 59, 121–126. [2]

Committee for Proprietary Medicinal Products (CPMP) (2001), *Points to Consider on Application With 1. Meta-Analyses; 2. One Pivotal Study*. [9]

Cook, J. A., Fergusson, D. A., Ford, I., Gonen, M., Kimmelman, J., Korn, E. L., and Begg, C. B. (2019), "There Is Still a Place for Significance Testing in Clinical Trials," *Clinical Trials*, 16, 223–224. [1]

De Groot, A. D. (1956/2014), "The Meaning of Significance for Different Types of Research," Translated and Annotated by Eric–Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas," *Acta Psychologica*, 148, 188–194. [3]

DiMasi, J. A., Feldman, L., Seckler, A., and Wilson, A. (2010), "Trends in Risks Associated With New Drug Development: Success Rates for Investigational Drugs," *Clinical Pharmacology & Therapeutics*, 87, 272–277. [3]

FDA (1998), "The Quantity of Evidence to Support Effectiveness, Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drugs and Biological Products," available at *https://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm072008.pdf*. [9]

——— (2017), "22 Case Studies Where Phase 2 and Phase 3 Trials Had Divergent Results," available at *https://www.fda.gov/downloads/AboutFDA/ReportsManualsForms/Reports/UCM535780.pdf*. [3]

Fisher, L. D. (1999), "One Large, Well-Designed, Multicenter Study as an Alternative to the Usual FDA Paradigm," *Drug Information Journal*, 33, 265–271. [9]

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, London: Oliver and Boyd. [2]

——— (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture for Great Britain*, 33, 503–513. [2]

——— (1956), *Statistical Methods and Scientific Inference*, Edinburgh: Oliver & Boyd. [2,11]

Fricker, R. D., Jr., Burke, K., Han, X., and Woodall, W. H. (2019), "Assessing the Statistical Analyses Used in Basic and Applied Social Psychology After Their *p*-Value Ban," *The American Statistician*, 73, 374–384. [2]

Gelman, A., and Carlin, J. B. (2014), "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors," *Perspectives on Psychological Science*, 9, 641–651. [8]

——— (2017), "Some Natural Solutions to the *p*-Value Communication Problem—And Why They Won't Work," *Journal of the American Statistical Association*, 112, 899–901. [4]

Gibson, E. W. (2019), "Leadership in Statistics: Increasing Our Value and Visibility," *The American Statistician*, 73, 109–116. [10]

Goodman, S. N. (1992), "A Comment on Replication, *p*-Values, and Evidence," *Statistics in Medicine*, 11, 875–879. [5]

——— (1999), "Towards Evidence-Based Medical Statistics, II: The Bayes Factor," *Annals of Internal Medicine*, 130, 1005–1013. [4]

——— (2019), "Why Is Getting Rid of *p*-Values So Hard? Musings on Science and Statistics," *The American Statistician*, 73, 26–30. [4,8]

Goodman, S. N., and Greenland, S. (2007), "Why Most Published Research Findings Are False: Problems in the Analysis," *PLoS Medicine*, 4, e168. [1]

Greenland, S. (2019), "Valid *p*-Values Behave Exactly as They Should: Some Misleading Criticisms of *p*-Values and Their Resolution with S-Values," *The American Statistician*, 73, 106–114. [7]

Harrington, D., D'Agostino, R. B., Sr., Gatsonis, C., Hogan, J. W., Hunter, D. J., Normand, S. T., Drazen, J. M., and Hamel, M. B. (2019), "New Guidelines for Statistical Reporting in the Journal," *New England Journal of Medicine*, 381, 285–286. [1]

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J. (2014), "Clinical Development Success Rates for Investigational Drugs," *Nature Biotechnology*, 32, 40–51. [4]

Hedges, L. V. (1984), "Estimation of Effect Size Under Nonrandom Sampling: The Effects of Censoring Studies Yielding Statistically Insignificant Mean Differences," *Journal of Educational Statistics*, 9, 61–85. [8]

Held, L., and Ott, M. (2018), "On p-Values and Bayes Factors," *Annual Review of Statistics and Its Application*, 5, 393–419. [4]

Hubbard, R., and Bayarri, M. J. (2003), "Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing," *The American Statistician*, 57, 171–178. [2]

Hung, H. M. J., and O'Neill, R. T. (2003), "Utilities of the p-Value Distribution Associated With Effect Size in Clinical Trials," *Biometrical Journal*, 45, 659–669. [6]

Hung, H. M. J., O'Neill, R. T., Bauer, P., and Kohne, K. (1997), "The Behavior of the p-Value When the Alternative Hypothesis Is True," *Biometrics*, 53, 11–22. [5]

ICH (1998), "E9 Topic: Statistical Principles for Clinical Trials," available at *http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf*. [8]

Ioannidis, J. P. A. (2005a), "Why Most Published Research Findings Are False," *PLoS Medicine*, 2, e124. [1]

——— (2005b), "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research," *JAMA*, 294, 218–228. [2]

Jannot, A. S., Ehret, G., and Perneger, T. (2015), "$p < 5 \times 10^{-8}$ Has Emerged as a Standard of Statistical Significance for Genome-Wide Association Studies," *Journal of Clinical Epidemiology*, 68, 460–465. [9]

Jeffreys, H. (1935), "Some Tests of Significance, Treated by the Theory of Probability," *Mathematical Proceedings of the Cambridge Philosophical Society*, 31, 203–222. [4]

——— (1961), *Theory of Probability* (3rd ed.), Oxford, UK: Oxford University Press. [4,5]

Kafadar, K. (2019), "Statistics and Unintended Consequences," *Amstat News*, 504, 3–4. [1]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 430, 773–795. [4]

Kempthorne, O. (1976), "Of What Use Are Tests of Significance and Tests of Hypothesis," *Communication in Statistics—Theory and Methods*, 5, 763–777. [5]

Kennedy-Shaffer, L. (2017), "When the Alpha Is the Omega: p-Values, 'Substantial Evidence,' and the 0.05 Standard at FDA," *Food and Drug Law Journal*, 72, 595–635. [9]

Kirby, S., Burke, J., Chuang-Stein, C., and Sin, C. (2012), "Discounting Phase 2 Results When Planning Phase 3 Clinical Trials," *Pharmaceutical Statistics*, 11, 373–385. [8]

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., Cross, E. S., Daniels, S., Danielsson, H., DeBruine, L., Dunleavy, D. J., Earp, B. D., Feist, M. I., Ferrell, J. D., Field, J. G., Fox, N. W., Friesen, A., Gomes, C., Gonzalez-Marquez, M., Grange, J. A., Grieve, A. P., Guggenberger, R., Grist, J., van Harmelen, A.-L., Hasselman, F., Hochard, K. D., Hoffarth, M. R., Holmes, N. P., Ingre, M., Isager, P. M., Isotalus, H. K., Johansson, C., Juszczyk, K., Kenny, D. A., Khalil, A. A., Konat, B., Lao, J., Larsen, E. G., Lodder, G. M. A., Lukavský, J., Madan, C. R., Manheim, D., Martin, S. R., Martin, A. E., Mayo, D. G., McCarthy, R. J., McConway, K., McFarland, C., Nio, A. Q. X., Nilsonne, G., de Oliveira, C. L., de Xivry, J.-J. O., Parsons, S., Pfuhl, G., Quinn, K. A., Sakon, J. J., Saribay, S. A., Schneider, I. K., Selvaraju, M., Sjoerds, Z., Smith, S. G., Smits, T., Spies, J. R., Sreekumar, V., Steltenpohl, C. N., Stenhouse, N., Świątkowski, W., Vadillo, M. A., Van Assen, M. A. L. M., Williams, M. N., Williams, S. E., Williams, D. R., Yarkoni, T., Ziano, I., and Zwaan, R. A. (2018), "Justify Your Alpha," *Nature Human Behaviour*, 2, 168–171. [2]

Lambert, D., and Hall, W. J. (1982), "Asymptotic Lognormality of p-Values," *The Annals of Statistics*, 10, 44–64. Corrected Vol. 11, p. 348. [5]

Lane, D. M., and Dunlap, W. P. (1978), "Estimating Effect Size: Bias Resulting From the Significance Criterion in Editorial Decisions," *British Journal of Mathematical and Statistical Psychology*, 31, 107–112. [7]

Lehmann, E. L. (1993), "The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?," *Journal of the American Statistical Association*, 88, 1242–1249. [2]

Mandel, M., and Rinott, Y. (2009), "A Selection Bias Conflict and Frequentist Versus Bayesian Viewpoints," *The American Statistician*, 63, 211–217. [4]

Maurer, K., Hudiburgh, L., Werwinski, L., and Bailer, J. (2019), "Content Audit for p-Value Principles in Introductory Statistics," *The American Statistician*, 73, 385–391. [10]

Miller, J. E., Wilenzick, M., Ritcey, N., Ross, J. S., and Mello, M. M. (2017), "Measuring Clinical Trial Transparency: An Empirical Analysis of Newly Approved Drugs and Large Pharmaceutical Companies," *BMJ Open*, 7, e017917, DOI: 10.1136/bmjopen-2017-017917. [7]

Neyman, J., and Pearson, E. S. (1928a), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part I," *Biometrika*, 20A, 175–240. [2]

——— (1928b), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part II," *Biometrika*, 20A, 263–294. [2]

Oakes, M. (1986), *Statistical Inference: A Commentary for the Social and Behavioral Sciences*, Chichester: Wiley. [3]

Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. [7]

Packer, M., Carson, P., Elkayam, U., Konstam, M. A., Moe, G., O'Connor, C., Rouleau, J. L., Schocken, D., Anderson, S. A., DeMets, D. L., for the PRAISE-2 Study Group (2013), "Effect of Amlodipine on the Survival of Patients with Severe Chronic Heart Failure Due to a Nonischemic Cardiomyopathy," *JACC: Heart Failure*, 1, 308–314. [9]

Packer, M., O'Connor, C. M., Ghali, J. K., Pressler, M. L., Carson, P. E., Belkin, R. N., Miller, A. B., Neuberg, G. W., Frid, D., Wertheimer, J. H., Cropp, A. B., DeMets, D. L., for the Prospective Randomized Amlodipine Survival Evaluation Study Group (1996), "Effect of Amlodipine on Morbidity and Mortality in Severe Chronic Heart Failure," *New England Journal of Medicine*, 355, 1107–1114. [9]

Pearson, K. (1900), "On the Criterion That a Given System of Deviations From the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen From Random Sampling," *Philosophical Magazine Series*, 5, 157–175. [2]

Pereira, T. V., Horwitz, R. I., and Ioannidis, J. P. A. (2012), "Empirical Evaluation of Very Large Treatment Effects of Medical Intervention," *JAMA*, 308, 1676–1684. [7]

Pitt, B., Poole-Wilson, P. A., Segal, R., Martinez, F. A., Dickstein, K., Camm, A. J., Konstam, M. A., Riegger, G., Klinger, G. H., Neaton, J., Sharma, D., Thiyagarajan, B., on behalf of the ELITE II Investigators (2000), "Effects of Losartan Compared With Captopril on Mortality in Patients With Symptomatic Heart Failure: Randomized Trial—The Losartan Heart Failure Survival Trial Study ELITE II," *The Lancet*, 355, 1582–1587. [10]

Pitt, B., Segal, R., Martinez, F. A., Meurers, G., Cowley, A. J., Thomas, I., Deedwania, P. C., Ney, D. E., Snavely, D. B., Chang, P. I., on behalf of ELITE Study Investigators (1997), "Randomized Trial of Losartan Versus Captopril in Patients Over 65 With Heart Failure (Evaluation of Losartan in the Elderly Study, ELITE)," *The Lancet*, 349, 747–752. [10]

Pocock, S. J., McMurray, J. J. V., and Collier, T. J. (2015), "Making Sense of Statistics in Clinical Trial Reports," *Journal of the American College of Cardiology*, 66, 2536–2549. [9]

Pocock, S. J., and Stone, G. W. (2016a), "The Primary Outcome Fails—What Next?," *New England Journal of Medicine*, 375, 861–870. [9]

——— (2016b), "The Primary Outcome Is Positive—Is That Good Enough," *New England Journal of Medicine*, 375, 971–979. [9]

Rockhold, F., Bromley, C., Wagner, E. K., and Buyse, M. (2019), "Open Science: The Open Clinical Trials Data Journey," *Clinical Trials*, 16, 539–546. [7]

Rosenthal, R. (1979), "The File Drawer Problem and Tolerance for Null Results," *Psychological Bulletin*, 86, 638–641. [7]

Senn, S. (2002), "Letter to the Editor: 'A Comment on Replication, p-Values and Evidence by S.N. Goodman,'" *Statistics in Medicine*, 21, 2437–2444. [5]

Shao, J., and Chow, S. C. (2002), "Reproducibility Probability in Clinical Trials," *Statistics in Medicine*, 21, 1727–1742. [5]

Siegfried, T. (2010), "Odds Are, Its Wrong: Science Fails to Face the Shortcomings of Statistics," *Science News*, 177, 26–29, available at *https://www.sciencenews.org/article/odds-are-its-wrong*. [2]

Simon, R. (1982), "Randomized Clinical Trials and Research Strategy," *Cancer Treatment Reports*, 66, 1083–1087. [3]

Staquet, M. J., Rozencweig, M., and Von Hoff, D. D. (1979), "The Delta and Epsilon Errors in the Assessment of Cancer Clinical Trials," *Cancer Treatment Reports*, 63, 1917–1921. [3]

Sterling, T. D. (1959), "Publication Decisions and their Possible Effects on Inferences Drawn From Tests of Significance—Or Vice Versa," *Journal of the American Statistical Association*, 285, 30–34. [7]

Sterne, J. A. C., and Smith, D. G. (2001), "Sifting the Evidence—What's Wrong With Significance Tests?," *BMJ*, 322, 226–231. [3]

Stigler, S. (2008), "Fisher and the 5% Level," *CHANCE*, 21, 12–12. [3]

Temple, R. (2005), "How FDA Currently Makes Decisions on Clinical Studies," *Clinical Trials: Journal of the Society for Clinical Trials*, 2, 276–281. [8,9]

Trafimow, D., and Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology*, 37, 1–2. [2]

Tukey, J. W. (1977), "Some Thoughts on Clinical Studies, Especially Problems of Multiplicity," *Science*, 198, 679–684. [7,8]

——— (1980), "We Need Both Exploratory and Confirmatory," *The American Statistician*, 34, 23–25. [8]

Vaduganathan, M., Butler, J., and Gheorghiade, M. (2016), "Transforming Drug Development in Heart Failure," *Circulation: Heart Failure*, 9, e003192. [9]

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., and Kievit, R. A. (2012), "An Agenda for Purely Confirmatory Research," *Perspectives on Psychological Science*, 7, 632–638. [8]

Wang, S.-J., Hung, H. M. J., and O'Neill, R. T. (2006), "Adapting the Sample Size Planning of a Phase III Trial Based on Phase II Data," *Pharmaceutical Statistics*, 5, 85–97. [8]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on *p*-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [2]

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019), "Moving to a World Beyond $p < 0.05$," *The American Statistician*, 73, 1–19. [1,2]

Wellek, S. (2017), "A Critical Evaluation of the Current *p*-Value Controversy" (with discussion), *Biometrical Journal*, 59, 854–872. [8]

Wong, C. H., Siah, K. W., and Lo, A. W. (2019), "Estimation of Clinical Trial Success Rates and Related Parameters," *Biostatistics*, 20, 273–286. [4]