The Bernoulli Fallacy

When a small p-value is read as proof, and why the advice keeps changing

Murray Cantor PhD

Fault 4 of six · healthy-skepticism.com · September 2026

The fault

The purpose of medical research is to show that a claim is valid given the data. A trial proposes that a drug prevents heart attacks. The data should tell us how likely that is. That is not what the papers report. They report something called statistical significance, a small p-value, and the reader takes it as proof the claim is true. It is not, and taking it as proof is the fault. Significance answers a different question, and the difference has led to a crisis in medicine: a large share of published findings do not hold up when someone tries to repeat them.

The symptom is familiar. Why does the published medical advice keep changing? This paper explains what significance actually measures, why it is so often mistaken for proof, and why a healthy skeptic should be wary of any claim that rests on it.

What significance measures

Fault 2 introduced the notation P(X | Y), the probability of X given Y. A researcher wants P(hypothesis | data): the probability the claim is true, given the trial results. The p-value runs the other way. It is P(data | hypothesis is false): the probability of getting these results if the claim were false. A small p says it is unlikely we would get this data if the hypothesis were false. When p is below 0.05 the result is called significant and gets published.

These are the two directions of one conditional probability. They are different numbers. A small enough p does make the hypothesis hard to dismiss, but small enough is not 0.05. Physicists will not claim a discovery until p is below one in 3.5 million, five sigma. Medicine accepts one in twenty, and p below 0.01 is not much better. At that bar the claim of significance is simply wrong. It says nothing about whether the hypothesis is true. To get from P(data | hypothesis is false) to P(hypothesis | data) you need a second number, how plausible the hypothesis was before the trial, and the p-value does not contain it. Reading it as if it did is fault 2, probability reversal, committed by the profession that should know best.

The crisis it produced

When landmark cancer findings were rerun, 6 of 53 held up. In psychology, 35 of 97 findings stayed significant on replication. This is the irreproducibility crisis, and its arithmetic is simple. Honest testing at the 0.05 bar produces a false finding one time in twenty on its own, with no misconduct at all. What is observed is far worse than one in twenty, and the excess is the evidence of a second mechanism: searching subsets of the data until one clears the bar by luck, a practice with the honest name of data torturing and the polite one of p-hacking.

Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature 2012;483:531–533. Open Science Collaboration. Estimating the reproducibility of psychological science. Science 2015;349:aac4716. Clayton A. Bernoulli's Fallacy. Columbia University Press, 2021.

The example: baby aspirin

In 1988 the Physicians' Health Study reported that aspirin cut first heart attacks by 44%, with p below 0.00001. It was stopped early for benefit, and decades of daily-aspirin advice followed. In 2018 the retests arrived. ARRIVE reported p = 0.60. ASPREE found no net benefit and more major bleeding. In 2022 the U.S. Preventive Services Task Force recommended against starting aspirin after age 60.

Trials stopped early tend to stop on a lucky high. On the retests, p never crossed the line again. The advice changed because the finding would not repeat and the harm outweighed the benefit. This concerns primary prevention; for people who have already had a heart attack or stroke the evidence is a different body of work and has not reversed. Do not stop a prescribed medication on the strength of a paper.

Steering Committee of the Physicians' Health Study. N Engl J Med 1989;321:129–135. Gaziano JM et al. (ARRIVE). Lancet 2018;392:1036–1046. McNeil JJ et al. (ASPREE). N Engl J Med 2018;379:1509–1518. USPSTF. JAMA 2022;327:1577–1584.

Where the bar sits, and where it came from

Particle physics demands five sigma, a p-value below 0.000000287, one chance in 3.5 million; the Higgs boson waited for that bar. Medicine and the social sciences accept one chance in twenty. The 0.05 threshold was never derived; Fisher offered it as a convenience in 1925, and it hardened into a publication rule. Medicine cannot simply adopt five sigma, since patients are scarce and trials are costly. But note what the one in twenty is: a rate per test of nothing. The share of published findings that are false is a different quantity, and getting to it is a reversal question, how many of the tested ideas were true to begin with. The prevalence, yet again.

ATLAS and CMS required five sigma for the 2012 Higgs discovery. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2:e124. Fisher RA. Statistical Methods for Research Workers, 1925.

Why it happens

Researchers must publish, and journals publish what clears the bar. A career is made of published findings, a finding is defined by a p-value, and a p-value can be cleared by luck, by trying enough analyses, or by stopping at the right moment. The honest case for the method is that sometimes there is no prior to be had, in quality control or in particle physics. The less flattering reason, which Clayton develops, is that a more rigorous standard would disrupt the academic publication world. Either way the incentive runs toward the bar and not toward the truth, and the patient meets the consequence years later as a reversal of advice.

What to ask

- Have these findings been replicated? A single result clears the 0.05 bar by luck one time in twenty.
- How many analyses were run, and how many were reported?
- Was the trial stopped early, and on what?
- How plausible was the idea before the study? A surprising finding with a small p is still a surprising finding.