When Are Results Too Good to Be True?

Gary A. Churchill
The Jackson Laboratory, Bar Harbor, Maine 04609

SCIENCE is in crisis. Many published studies appear to be irreproducible (Prinz *et al.* 2011; Begley and Ellis 2012). What can be done? How concerned should we be?

Is the reproducibility crisis something new? Suppose we randomly sampled studies from the past 100 years of scientific literature and attempted to replicate their findings. I expect that many of these studies would fail to replicate. We must be careful not to conflate "irreproducible" with "false." The experiments may require specific conditions that are difficult to reproduce; the original studies may have been underpowered; or they may have addressed a hypothesis that turned out to be false. "False" studies are part and parcel of the scientific method in which falsifiable hypotheses are repeatedly put to the test. Many potentially transformative studies have been published and later discredited (*e.g.*, Fleischmann and Pons 1989); others have stood the test of time and have become integrated into the fabric of knowledge (*e.g.*, Luria and Delbruck 1943).

Something has changed, however: the industry of science has grown exponentially, at a rate of ~4% per year as measured in the number of articles published (Larsen and Von Ins 2010). At this rate of growth, the doubling time is ~15 years, which means that very soon more scientific publications will have appeared in the 21st century than in all of prior history. This growth has resulted in fierce competition for coveted spots in high-profile journals. Being good, or even very good, is no longer good enough. One also needs to be lucky.

Nature magazine receives ~10,000 submissions annually and (necessarily) rejects >90% of them. Suppose that all of the 10,000 studies submitted to *Nature* in a year had evaluated false scientific hypotheses. The *p*-values reported would be uniformly distributed between zero and one[1]. Some of these would be significant by chance. If the *Nature* editors evaluated these studies based solely on the reported *p*-values, the probability that at least one of these lucky papers would report "$p < 0.0001$" is 67%. Of course, it takes more than an impressive *p*-value to be published in *Nature*. A truly outstanding article will have something novel, even surprising, to say.

To understand the impact of the novelty criterion, consider the following three experiments (see Greenhouse 2012). In the first experiment, a music expert claims that she can distinguish a score written by Mozart from one written by Haydn. Presented with 10 scores in a double-blinded and randomized order, she identifies each one correctly. In the second experiment, a tea-drinking lady claims that she can tell if the milk or the tea was poured first into the cup. Given 10 carefully prepared cups of tea in randomized order, she correctly identifies each one. Finally, an inebriated customer at a bar claims he can predict the outcome of a coin toss. He proceeds to toss the coin and calls heads or tails correctly 10 times in a row.

The *p*-value in each case is 0.001 ($1/2^{10}$). But what conclusions are we to draw? Here we cannot avoid our subjective opinions about the prior plausibility of each claim. In my opinion, the music experiment was not really necessary because the claim is believable *a priori*. In the case of the tea-drinking lady, while I may have initially doubted this unusual talent, the evidence is convincing and I would consider the matter settled. As for the drunken coin tosser, after carefully examining the coin, I would ask him to do it again.[2] This claim is just too hard to believe; a higher standard of evidence seems justified.

Which of these studies, if properly vetted by review, would make an exciting paper? My vote is with "Alcohol induces clairvoyance." Experiments based on hypotheses that are *a priori* implausible can be potentially groundbreaking, but are also the most likely to be false (Ioannidis 2005). Can we determine which are which?

The shortcomings of *p*-values as a measure of evidence are well documented (Berger and Sellke 1987) and continue to inspire debate (Nuzzo 2014). Imagine, then, a statistic

[1]By definition, under the null hypothesis a *p*-value is randomly and uniformly distributed between zero and one. But for this hypothetical exercise, you must suspend your disbelief about the audacity of the scientists who submitted their nonsignificant findings for publication.

doi: 10.1534/genetics.114.169912
Address for correspondence: The Churchill Group, The Jackson Laboratory, 600 Main Street Bar Harbor, Maine 04609. E-mail: garyc@jax.org

[2]Bear in mind that every day one-in-a-million events will happen to ~7000 people (Littlewood 1986).

that provides an optimal and objective measure of evidence based only on the experimental data at hand without recourse to prior beliefs and opinions. Let us call it the "o-value." Using the o-value as our criterion for publication, we would still publish false conclusions. Otherwise, we would have to be so stringent in our evaluation that an excessive number of true findings would be rejected or our requirements for evidence would be so prohibitive that progress would grind to a halt.

A paradoxical consequence of having access to an optimal measure of evidence is the impossibility of distinguishing a true claim from a false one. If that were possible, it would follow that there is additional information in the data, which contradicts our stipulation that the o-value is optimal. We could use that extra information to create an improved o-value, but we would end up in the same conundrum. Even with an ideal measure of evidence, it is impossible to objectively establish truth or falsehood of a claim based solely on the available data. This is a troubling reality.

The Case of Dias and Ressler

A recent publication in *Nature Neuroscience* (Dias and Ressler 2013) put forward a provocative hypothesis that epigenetic inheritance can be modified by olfaction. The paper understandably drew a great deal of attention. In this issue of *GENETICS*, we publish a critique by G. Francis who argues that the evidence presented is "too good to be true." Francis claims that the study could not plausibly have been as successful as reported and that some of the experiments reported should have failed to reject the null hypothesis. These concerns are reminiscent of Fisher's claim that Mendel cooked his data (Hartl and Fairbanks 2007). While Mendel's data may remain the subject of debate, Mendel's laws—appropriately modified to conform to subsequent observations—have stood the test of time.

The *post hoc* power analysis provided by Francis is enlightening, but the evidence required to evaluate the study is entirely contained in the reported *p*-values (Hoenig and Heisy 2001). If the same analysis were applied to a randomly selected study, it might cast doubt on the integrity of the study. But the Dias and Ressler study was selected from among thousands of studies competing for limited publication space by an evaluation process that selects papers with inflated *p*-values. "Extraordinary claims require extraordinary evidence."[3] Skeptical reviewers are going to balk at any sign of weakness in a controversial manuscript. And yet, extraordinary evidence can occur by chance. This suggests that improbable *p*-values are to be expected in controversial, high-profile papers.

Opinions play an important role in deciding what gets published. As illustrated in our hypothetical example of three experiments that produce identical statistical evidence, subjectivity plays a crucial role in our interpretation of that evidence. The proposal that epigenetic inheritance can be modified by olfaction stretches the imagination, but it is not outside the realm of possibility. Expert scientists vetted the experimental procedures, and we must assume that the data reported are accurate and complete. But is the claim true? Without further study, it is a matter of opinion.

Progress in science requires an influx of new ideas balanced by skepticism that compels us to re-examine the evidence. When we seek more evidence to corroborate or refute hypotheses, some will prove to be wrong. We should not subvert this process by reaching for an unattainable ideal of perfectly reproducible studies. Ironically, statistical evidence presented in the original study may be of little help in determining which hypotheses will hold up. Dias and Ressler have proposed an intriguing hypothesis, and they have reported their evidence to support it. Of course, the study should be repeated—perhaps by using different approaches and methods that address the same hypothesis from different angles. The findings of Dias and Ressler warrant further study, and the scientific method compels us to try to topple this hypothesis. Is it true or too good to be true? Only time—and further investigation—will tell.

Note added in proof: See Dias and Ressler 2014 (pp. 453) and Francis 2014 (pp. 449–451) in this issue for a related work.

Literature Cited

Begley, C. G., and L. M. Ellis, 2012 Drug development: raise standards for preclinical cancer research. Nature 483(7391): 531–533.

Berger, J. O., and T. Sellke, 1987 "Testing a point null hypothesis: the irreconcilability of *p* values and evidence" (with discussion). J. Am. Stat. Assoc. 82: 112–139.

Dias, B. G., and K. J. Ressler, 2014 Reply to Gregory Francis. Genetics 198: 453.

Dias, B. G., and K. J. Ressler, 2014 Parental olfactory experience influences behavior and neural structure in subsequent generations. Nat. Neurosci. 17: 89–96.

Fleischmann, M., and Pons, S. (1989). Electrochemically induced nuclear fusion of deuterium. J. Electroanal. Chem. 261(2, Part 1): 301–308.

Francis, G., 2014 Too much success for recent groundbreaking epigenetic experiments. Genetics 198: 449–451.

Greenhouse, J. B., 2012 On becoming a Bayesian: early correspondences between J. Cornfield and L. J. Savage. Stat. Med. 31: 2782–2790.

Hartl, D. L., and D. J. Fairbanks, 2007 Mud sticks: on the alleged falsification of Mendel's data. Genetics 175(3): 975–979.

Hoenig, J. M., and D. M. Heisy, 2001 The abuse of power: the pervasive fallacy of power calculations for data analysis. Am. Stat. 55(1): 19–24.

Ioannidis, J. P. A., 2005 Why most published research findings are false. PLoS Med. 2(8): e124.

Larsen, P. O., and M. von Ins, 2010 The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84: 575–603.

Littlewood, J. E., 1986 *Littlewood's Miscellany.* Cambridge University Press, Cambridge, UK.

Luria, S. E., and M. Delbrück, 1943 Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28(6): 491–511.

Nuzzo, R., 2014 Scientific method: statistical errors. *Nature* 506: 150–152.

Prinz, F., T. Schlange and K. Asadullah, 2011 Believe it or not: How much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. 10:712.

[3]Carl Sagan, from the TV series *Cosmos*.

Communicating editor: M. Johnston