ADVERTISEMENT

How Shoddy Data Becomes Sensational Research

Academics are addicted to p-hacking, data torturing, and other statistical sins.

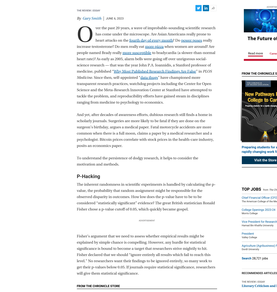

ILLUSTRATION BY JOE MCKENDRY FOR THE CHRONICLE

By Gary Smith | JUNE 8, 2022

For the past 20 years, a team of independent-minded scientists have been tracking cancer deaths in the state of Georgia. Are Atlanta Hawks fans more prone to heart attacks on the Sunday day-after-a-game-loss? Do opposite sexes really increase testosterone? Do men really eat more when women are around? Are people natural freely more susceptible to heartbreaks in above-mean normal hours-of-sleep? In early 2019, Atlanta folk were going all over outrageous social science research — that was the year John P.A. Ioannidis, a Stanford professor of medicine, published "Retracting a University's Reputation" in Chronicle of Higher Education. Since then, other epidemiologists have been championed as the transparent research pioneers, watchdog projects including the Center for Open Science and the Meta Research Innovation Center at Stanford have attempted to tackle the problems, and reproducibility efforts have gained steam in disciplines ranging from medicine to psychology to economics.

And yet, after decades of awareness efforts, dubious research still finds a home in scholarly journals. Surgeries are more likely to be fatal if they are done in the morning rather than the afternoon, and "power poses" can increase testosterone, common sense flows to a full moron, claims a paper by a cardiac researcher and psychologist. Bitcoin prices correlate with stock prices in the health care industry, proclaim an economics paper.

To understand the persistence of dodgy research, it helps to consider the motivation and methods.

P-Hacking

The inherent randomness in scientific experiments is handled by calculating the p-value, the probability that random assignment might be responsible for the observed results. If that probability is less than 5 percent, the result is the considered "statistically significant" evidence? The great British statistician Ronald Fisher chose a p-value cutoff of 0.05, which quickly became gospel.

ADVERTISEMENT

Fisher's argument that we need to assess whether empirical results might be explained by simple chance is compelling. However, any hurdle for statistical significance is bound to become a target that researchers strive mightily to hit. Fisher declared that we should "ignore results of experiments which fail to reach this level." No researchers want their findings to be ignored entirely, so many work to push their p-values below 0.05. If p-values require statistical significance, researchers will give them statistical significance.

FROM THE CHRONICLE STORE

ESSAY
Restructuring a University
How one institution is shaping stable

Visit the Store

With terabytes of data and lightning-fast computers, it is too easy to calculate first, think later. This is a flaw, not a feature.

The result is p-hacking — trying different combinations of variables, looking at subsets of the data, discarding contradictory data, and generally doing whatever it takes until something with a low p-value is found and then pretending that this is what you were looking for in the first place. As Donald Coxon, an economics Nobel laureate, cynically observed, "If you torture data long enough, they will confess."

Consider a 2016-2017 article lyricized up by diverse health-medical claiming that surgeries are more likely to be fatal if they are done on the surgeon's birthday. It is a truly damning indictment of true, that patients are dying because surgeons are distracted by birthday plans and good wishes from colleagues. The conclusion is plausible, but it is provocative and media-friendly — something that is often true of p-hacked studies.

The result is p-hacking — trying different combinations of variables, looking at subsets of the data, discarding contradictory data, and generally doing whatever it takes until something with a low p-value is found and then pretending that this is what you were looking for in the first place. As Donald Coxon, an economics Nobel laureate, cynically observed, "If you torture data long enough, they will confess."

Consider a 2016-2017 article lyricized up by diverse health-medical claiming that surgeries are more likely to be fatal if they are done on the surgeon's birthday. It is a truly damning indictment of true, that patients are dying because surgeons are distracted by birthday plans and good wishes from colleagues. The conclusion is plausible, but it is provocative and media-friendly — something that is often true of p-hacked studies.

It is difficult to prove p-hacking, but one sign is when the research involves many selection choices, what Andrew Gelman, professor of statistics and political science at Columbia University, has likened to a "garden of forking paths." The Berkeley study included thousands patients who underwent one of 17 common types of surgery between 2011 and 2014. Four cardiovascular surgeries and the 13 most common cancer surgeries, nonelective surgeries in the Medicare database. The use of 2011-14 data is a proper, published in 2020 is perplexing. The choice of 17 surgeries is baffling. P-hacking would explain all of this.

The authors justified their surgery selections by referencing several studies that had used Medicare data to investigate the relationship between surgery and mortality and other variables. One of those four cited papers considered 14 cardiovascular or cancer operations but reported results for only four cardiovascular procedures and four cancer resections; two papers examined four cardiovascular and four cancer operations and the fourth paper considered four cardiovascular operations and 16 most common noncardiovascular surgeries in the Medicare population.

The four cardiovascular procedures considered in the birthday paper are identical or nearly identical to those reported in the four cited papers. However, the inclusion of 13 other procedures is suspicious. Why didn't they use a more natural number, like 10, or perhaps 16, so that the total would be 20? Did 13 procedures give the lowest p-value? It is also striking that none of the four referenced studies excluded patients with cancer, but the birthday study did. The authors' explanation was that they wanted "to be consistent" with prior conference studies — and of the cancer, after being prospective mortality.

Even with all these possible p-hacks, the reported p-value is 0.01, only marginally under Fisher's 5 percent rule. One sign of widespread p-hacking by researchers is the suspicious clustering of reported p-values slightly below 0.05. A 0.01p-value does not necessarily mean that there was p-hacking — but when there are many looking-pathes and p-value bulls are chosen, a marginal p-value is not compelling evidence.

Brian Wansink retired from his position as a professor of marketing at Cornell University and director of the university's Food and Brand Lab after a number of problems were discovered with his studies, including outrageous p-hacking. One mindless eater was so proud to a co-author lamenting that a p-value was 0.06: "If you can get the data, and it needs some tweaking, it would be good to get that one value below 0.05."

HARKing

In Gelman's garden of forking paths strategy, p-hacking occurs when a researcher seeks empirical support for a theory by trying several paths and reporting the path with the lowest p-value. Other times, a researcher begins with a theory and looks for patterns and make up a theory after crunching a dataset with a low p-value. This is hypothesizing after the results are known — HARKing.

A good example is a 2018 National Bureau of Economic Research study of bitcoin prices. Bitcoin is particularly interesting because there is theoretical reason why bitcoin prices should be related to anything other than investor expectations about future prices, or perhaps random speculation. Unlike bonds that pay interest and stocks that pay dividends, bitcoin doesn't yield any income at all, so there is no logical way to value bitcoin the way investors value stocks and bonds.

Nonetheless, the NBER sleeping paper reported hundreds of estimated statistical relationships between bitcoin prices and various variables, including such seemingly random items as the Canadian dollar-U.S. dollar exchange rate, the price of crude oil, and stock returns in the automobile, book, and beer industries. I am not making this up.

Of the 810 statistical relations they report, 64 percent are significant at the 5 percent level — which is somewhat fewer than the 81 statistically significant relationships that would be expected if they had just correlated bitcoin prices with random numbers.

ADVERTISEMENT

The occasional justification the authors offer are seldom persuasive. For example, they acknowledge that, unlike stocks, bitcoins don't generate income or pay dividends, so they "proxy" this value using the number of bitcoin wallet users:

> Obviously, there is no direct measure of dividend for the cryptocurrencies. However, in its essence, the price is dividend ratio is a measure of the gap between the market value and the fundamental value of an asset. The market value of cryptocurrency is just the observed price. We proxy the fundamental value by using the number of Bitcoin wallet users.

The number of bitcoin wallet users is not analogous to the income earned on stocks or the dividends paid to stockholders and is not a valid proxy for fundamental value of bitcoin — which is a big hot area.

Among the 64 statistical relationships that were significant at the 10 percent level, the researchers reported finding that bitcoin returns were positively correlated with stock returns in the consumer goods and health care industries, and negatively correlated with stock returns in the fabricated products and metal mining industries. These correlations don't make any sense, and the authors did not try to explain them. "We don't give explanations, we just document this behavior." Academics rarely have better things to do than document coincidental correlations.

Dry Labbing

Some are tempted by an even easier strategy — simply make up whatever data are needed to support the desired conclusion. When Diederik Stapel, a prominent social psychologist, was exposed in 2011 for having made up data, it led to his firing and the eventual retraction of 58 papers. His explanation: "I was not able to withstand the pressure to write papers, to publish, to always have to be better." He confessed: "I wanted too much, too fast."

It is not a short hop, skip, and jump from making up data to making up entire papers. In 2005, three MIT graduate students created a prank program they called SCIgen that used randomly selected words to generate bogus computer science papers. Their goal was to "maximize amusement, rather than coherence" and, also, to demonstrate that some academic conferences will accept almost anything.

They submitted a SCIgen paper with this gibberish abstract to the World Multiconference on Systemics, Cybernetics and Informatics:

> Many physicists would agree that, had it not been for congestion control, the evaluation of web browsers might never have occurred. In fact, few hackers worldwide would disagree with the essential unification of voice-over IP and public-private key pair. In order to solve this riddle, we confirm that SMPs can be made smoothed, cachable, and interposable.

The conference organizers accepted the prank paper and then withdrew their acceptance when the graduate students revealed their hoax. The conferences have gone on to bigger and better things, but SCIgen lives on. Believe it or don't, but some researchers have used SCIgen to bolster their CVs.

ADVERTISEMENT

Cyril Labbé, a computer scientist at Grenoble Alpe University, wrote a program to detect hoax papers published in real journals. Working with Guillaume Cabanac, a computer scientist at the University of Toulouse, they found 243 hoaxes published in mainstream journals. Some of the hoaxes were obvious nonsense, and people could tell from the gibberish that the paper was a prank or perhaps a hoax. One of the embarrassed publishers, Springer, subsequently expressed that it was working with Labbé to develop a tool that would identify computer papers. The obvious question is why such a tool is needed. In the peer review system so bodies that someone cannot recognize nonsense when they read it?

P-hacking and HARKing were less of a problem when it was not practical to assume zillions of models. Now, computers can do a second-what it takes to run thousands of models, and the temptation to do so is irresistible. It is unsurprising, since odd-ear that the half-life of data-science-based fact is calculations by hand were actually a blessing. The calculations were so hard that people thought hard before calculating. Today, with terabytes of data and lightning-fast computers, it is too easy to calculate first, think later. This is a flaw, not a feature.

P-hacking, HARKing, and dry labbing inevitably lead to the publication of fragile studies that do not hold up when tested with fresh data, which has caused far-reaching problems in recent years. In a non-ironical that 28% of the 1,637 randomized clinical trials published in three prestigious medical journals were medical reversals, that conclusion that previously recommended medical treatments were worthless, or worse.

In 2015, Brian Nosek's Reproducibility Project: retracted the results of attempts to replicate 100 studies that had been published in what are arguably the top three psychology journals. Only 36 continued to have p-values below 0.05 and to have effects in the same direction as the original studies.

In December 2021, the Center for Open Science (co-founded by Nosek, a psychology professor at the University of Virginia) and Science Exchange retracted the results of an eight-year project attempting to replicate 23 highly cited in vitro or animal-based preclinical cancer biology studies. The 23 papers involved 158 estimated effects. Only 46 percent replicated, and the median effect size was 85 percent smaller than originally estimated.

ADVERTISEMENT

In 2019, a team led by Colin Camerer, a behavioral economist at Caltech, attempted to replicate 18 experimental economics studies published in two prestigious journals. Only 11 were successfully replicated. In 2018 another Camerer team reported that it had attempted to replicate 21 experimental social science studies published in Science and Nature and found only 13 continued to be statistically significant and in the same direction with fresh data.

The skepticism that psychology researchers have for work in their field is sobering — and justified.

An interesting side study was done while Nosek's Reproducibility Project was underway. Psychologists were contacted before 64 of the replication studies were scheduled to be completed, auction markets were set up for researchers in the field of psychology to bet on whether each replication would be successful. People doing the studies were not allowed to participate. The final market prices indicated that researchers believed that these papers had, on average, a 55 percent chance of being successfully replicated, and the results showed that dismal expectations turned out to be overly optimistic: Only 33 of the 64 studies that were completed on time replicated. The skepticism that psychology researchers have for work in their field is sobering — and justified.

There are several ways to alleviate the replication crisis and restore the faith of science. Here are four of the more promising directions.

1. The first step for slowing the p-hacking/HARKing epidemic is for researchers to recognize the seriousness of the problem. In 2017, Joseph Simmons, Leif Nelson, and Uri Simonsohn wrote:

> We know many researchers — including ourselves — who readily admitted to dropping dependent variables, conditions, or participants so as to achieve significance. Everyone knew it was wrong, that they thought it was wrong, that it was wrong to jaywalk. ... Simulations revealed it was wrong the way it's wrong to rob a bank.

ADVERTISEMENT

Michael Inzlicht, a professor of psychology at the University of Toronto, spoke for many but not all others, he wrote that:

> I want social psychology to change. But, the only way we can really change is if we reckon with our past coming clean that we erred, and erred badly. ... Our problems are not small and they will not be resolved by small fixes. Our problems are systemic and they are at the core of how we conduct our science.

Statistics courses in all disciplines should include substantial discussion of p-hacking and HARKing.

2. It doesn't way to fight p-hacking and HARKing is to eliminate the incentive by removing statistical significance as a hurdle for publication. P-values can help us assess the extent to which chance might explain empirical results, but they should not be the primary measure of a model's success. Artificial thresholds like p = 0.05 encourage unwanted practices.

3. Peer review is often cursory. Compensating reviewers for thoroughly review might help screen out flawed research.

ADVERTISEMENT

4. Replication tests need replicators, and would be replicators need incentives. Highly skilled researchers are generally rewarded for their own work and have little reason to spend their time trying to replicate other peoples' research. One alternative is to make a replication study of an important paper a prerequisite for a Ph.D. or other degree in an empirical field. Such a requirement would allow students to see first-hand how research is done and would also generate thousands of replication tests.

None of these steps are easy, but they are all worth trying.

We welcome your thoughts and questions about this article. Please email the editors or submit a letter for publication.

FIELD GUIDES & RESOURCES | OPINION | DATA

Gary Smith

Gary Smith is a professor of economics at Pomona College. He is the author of *Distrust: Big Data, Data Torturing, and the Assault on Science* (Oxford University Press).

ADVERTISEMENT

The Future of Science

Read more stories | Carnegie Mellon campus

Visit the Store

FROM THE CHRONICLE STORE

Preparing students for a rapidly changing world. Series

Visit the Store

TOP JOBS See The Records

Chief Financial Officer (CFO)
The Ground College of the Mediterranean

College Openings 2023-24
Morris College

Vice President for Research
Hampton Roads University

President
Hampton

Open Rank Tenure/Tenured Faculty
Saint University

Search 20,721 jobs

RECOMMENDED ARTICLES

THE WEEKLY, EDITOR'S PICK
Literary Criticism and its Caricatures

THE WEEKLY, EDITOR'S PICK
The Mirror of Phong

THE WEEKLY, EDITOR'S PICK
John Guillory's Nonalignment Pact

ADVERTISEMENT

NEWSLETTER
Race in Admissions
Stay on top of the ever-changing criticisms administrators with Race in Admissions, a newsletter from Inside Higher Ed. Learn more. Free weekly.

Sign Up

IN THE CHRONICLE STORE

Restructuring a University

The Accessible Campus

Advancing Women's Leadership in Higher Education

Trouble at the Top

1255 23rd Street, N.W. Washington, D.C. 20037
© 2022 The Chronicle of Higher Education

Twitter | Instagram | Youtube | Facebook | LinkedIn