MedicalXpress

Misleading p-values showing up more often in biomedical journal articles, study finds

March 15 2016

A review of p-values in the biomedical literature from 1990 to 2015 shows that these widely misunderstood statistics are being used increasingly, instead of better metrics of effect size or uncertainty.

A study of millions of journal articles shows that their authors are increasingly reporting p-values but are often doing so in a misleading way, according to a study by researchers at the Stanford University School of Medicine. P-values are a measure of statistical significance intended to inform scientific conclusions.

Because p-values are so often misapplied, their increased use probably doesn't indicate an improvement in the way biomedical research is conducted or the way data are analyzed, the researchers found.

"It's usually a suboptimal technique, and then it's used in a biased way, so it can become very misleading," said John Ioannidis, MD, DSc, professor of disease prevention and of health research and policy and co-director of the Meta-Research Innovation Center at Stanford.

The study will be published March 15 in *JAMA*. Ioannidis is the senior author. The lead author is David Chavalarias, PhD, director of the Complex Systems Institute in France.

When p-values = embarrassment

The Ioannidis team used automated text mining to search the biomedical databases MEDLINE and PubMed Central for the appearance of p-values in millions of abstracts, and also manually reviewed 1000 abstracts and 100 full papers. All the papers were published between 1990 and 2015.

The widespread misuse of p-values—often creating the illusion of credible research—has become an embarrassment to several academic fields, including psychology and biomedicine, especially since Ioannidis began publishing critiques of the way modern research is conducted.

Reports in Nature, STAT and FiveThirtyEight, for example, have covered the weaknesses of p-values. On March 7, the American Statistical Association issued a statement warning against their misuse. In one of a series of essays accompanying the statement, Boston University epidemiologist Kenneth Rothman, DMD, DrPH, wrote, "These are pernicious problems. ... It is a safe bet that people have suffered or died because scientists (and editors, regulators, journalists and others) have used significance tests to interpret results, and have consequently failed to identify the most beneficial courses of action."

At Stanford, Ioannidis' team found that among all the millions of biomedical abstracts in the databases, the reporting of p-values more than doubled from 7.3 percent in 1990 to 15.6 percent in 2014. In abstracts from core medical journals, 33 percent reported p-values, and in the subset of randomized controlled clinical trials, nearly 55 percent reported p-values.

The meaning of p-values

P-values are designed to illuminate a fundamental statistical conundrum. Suppose a clinical trial compares two drug treatments, and drug A appears to be 10 percent more effective than drug B. That could be

because drug A is truly 10 percent more effective. Or it could be that chance just happened to make drug A appear more effective in that trial. In short, drug A could have just gotten lucky. How do you know?

A p-value estimates how likely it is that data could come out the way they did if a "null hypothesis" were true—in this case, that there is no difference between the effects of drugs A and B. So, for example, if drugs A and B are equally effective and you run a study comparing them, a p-value of 0.05 means that drug A will appear to be at least 10 percent more effective than drug B about 5 percent of the time.

In other words, assuming the drugs have the same effect, the p-value estimates how likely it is to get a result suggesting A is at least 10 percent better.

"The exact definition of p-value," said Ioannidis, "is that if the null hypothesis is correct, the p-value is the chance of observing the research result or some more extreme result." Unfortunately, many researchers mistakenly think that a p-value is an estimate of how likely it is that the null hypothesis is not correct or that the result is true.

P-values

"The p-value does not tell you whether something is true. If you get a p-value of 0.01, it doesn't mean you have a 1 percent chance of something not being true," Ioannidis added. "A p-value of 0.01 could mean the result is 20 percent likely to be true, 80 percent likely to be true or 0.1 percent likely to be true—all with the same p-value. The p-value alone doesn't tell you how true your result is."

For an actual estimate of how likely a result is to be true or false, said Ioannidis, researchers should instead use false-discovery rates or Bayes factor calculations.

Despite the serious limitations of p-values, they have become a symbol of good experimental design in the current era. But unfortunately, they are little more than a symbol. Ioannidis and his team found that practically the only p-values reported in abstracts were those defined somewhat arbitrarily as "statistically significant"—a number typically set at less than 0.05. The team found that 96 percent of abstracts with p-values had at least one such "statistically significant" p-value.

"That suggests there's selective pressure favoring more extreme results. The fact that you have so many significant results is completely unrealistic. It's impossible that 96 percent of all the hypotheses being tested would be significant," said Ioannidis.

But how big was the effect?

Despite increasing numbers of papers reporting that results were statistically significant, few papers reported how much of an effect a treatment had compared to controls or placebos. For example, suppose 10,000 patients showed an average improvement in symptoms that was statistically significant compared with another 10,000 who didn't get the drug. But if patients on the drug were only 1 percent better, the statistical significance derived from the p-value would likely have no practical value.

Of the 796 papers manually reviewed by the Ioannidis team that contained empirical data, only 111 reported effect sizes and only 18 reported confidence intervals (a measure of the uncertainty about the magnitude of the effect). Finally, none reported Bayes factors or false-discovery rates, which Ioannidis said are better-suited to telling us if what is observed is true. Fewer than 2 percent of abstracts the team reviewed reported both an effect size and a confidence interval.

In a manual review of 99 randomly selected full-text articles with data, 55 reported at least one p-value, but only four reported confidence intervals for all effect sizes, none used Bayesian methods and only one used false-discovery rates.

Ioannidis advocates more stringent approaches to analyzing data. "The way to move forward," he said, "is that p-values need to be used more selectively. When used, they need to be complemented by effect sizes and uncertainty [confidence intervals]. And it would often be a good idea to use a Bayesian approach or a false-discovery rate to answer the question, 'How likely is this result to be true?'"

Suboptimal technique

P-values are a suboptimal technique, and they often are used in a biased, misleading way, he said. "Across the entire literature, the statistical approaches used are often suboptimal. P-values are potentially very misleading, and they are selectively reported in favor of more significant results, especially in the abstracts. And authors underuse metrics that would be more meaningful and more useful to have—effect sizes, confidence intervals and other metrics that can add value in understanding what the results mean."

Provided by Stanford University Medical Center