NEWS

Statistical errors may taint as many as half of mouse studies

BY **BAHAR GHOLIPOUR**

15 MARCH 2018

Seven years ago, **Peter Kind**, a neuroscientist at the University of Edinburgh in Scotland, found himself in an uncomfortable situation. He was reading a study about **fragile X syndrome**, a developmental condition characterized by severe intellectual disability and, often, autism. The paper had appeared in a high-profile journal, and the lead scientist was a reputable researcher — and a friend. So Kind was surprised when he noticed a potentially serious statistical flaw.

The research team had looked at 10 neurons from each of the 16 mice in the experiment, a practice that in itself was unproblematic. But in the statistical analysis, the researchers had analyzed each neuron as if it were an independent sample. That gave them 160 data points to work with, 10 times the number of mice in the experiment.

“The question is, are two neurons in the brain of the same animal truly independent data points? The answer is no,” Kind says. “The problem is that you are increasing your chance of getting a false positive.”

The more times an experiment is replicated, the more likely it is that an observed effect is not just a lucky roll of the dice. That’s why more animals (or people) means more reliable results. But in the fragile X study, the scientists had artificially inflated the number of replications — a practice known as ‘pseudoreplication.’

This practice makes it easier to reach the sweet spot of statistical significance, especially in studies involving small numbers of animals. But treating measurements taken from a single mouse as independent samples goes against a fundamental principle of statistics and can lead scientists to find effects that don’t actually exist.

Persistent problem:

If a high-quality study such as this one included such a fundamental error, Kind wondered, how many others did? He gathered his students, and together they pored over hundreds of mouse studies published between 2001 and 2011 in six top neuroscience journals.

The problem, they found, was “shockingly” widespread, Kind recalls. By conservative measures, more than half the papers they looked at seemed to have committed pseudoreplication. Even more worrisome, 82 percent of the studies had not provided enough information about how the data had been gathered and analyzed for Kind and his students to know for sure.

Their survey suggests that pseudoreplication generates substantial false results and muddies the scientific record in many areas of neuroscience, including autism research, Kind says. “There are a lot of papers in the literature that say that they got a finding but they actually probably don’t.”

Kind is far from the only scientist who’s concerned. Since an ecologist first pointed out the problem in 1984, several studies have found pseudoreplication to be common in many areas of research[1,2].

In a study submitted for publication, researchers at Iowa State University examined 200 mouse studies published between 2000 and 2016 and found indications of pseudoreplication in about half of them. In study accepted for publication, **Stanley Lazic**, a statistician at the U.K.-based pharmaceutical company AstraZeneca, **surveyed 200 studies** published between 2011 and 2016.

Lazic and his team focused on animal studies with a specific design that comes with clear guidelines about how to treat the data. And yet they found that only 22 percent of the studies had followed the guidelines correctly, half had pseudoreplication and one-third didn’t provide enough information to know for sure.

“It’s a persistent problem that has been around for ages, and it doesn’t seem to be getting any better,” Lazic says. “The potential impact is pretty concerning.”

Radical stance:

In the majority of cases, the statistical faux pas is unintentional: Some researchers are simply unaware of statistical best practices, and others may have trouble because of the complexity of their studies. “They just don’t realize they are breaking statistical rules,” Kind says.

There are several solutions for handling data points that are not completely independent. In some cases, the measurements need to be averaged as one data point for each animal. When researchers don’t want to lose data in an average, they need to use a statistical model that adjusts for a group of measurements coming from a single animal.

Much of autism research is done with rodents, in studies that typically employ fewer than 10 animals, often just 3 or 4. Scientists tend to use a small number of mice for both ethical and

financial reasons — specially raised and often genetically modified, lab rodents are expensive.

But using too few mice can make it impossible to pick up an effect — even when one exists. “If you are doing any statistical analysis, such as p-values, you probably need a sample size of larger than 30,” says **David Vaux**, professor of cell biology at the Walter and Eliza Hall Institute of Medical Research in Melbourne, Australia.

Some scientists are questioning any reliance on p-values and suggest a more radical stance: that researchers describe their data and report the magnitude of an effect and its precision, using confidence intervals and graphical representations[3]. But such a shift would only be possible if journals drop the pervasive requirement that researchers present a p-value of less than 0.05, the conventional bar for ‘statistical significance.’

“We get rewarded for publishing — that’s essentially our unit of currency,” says **Malcolm Macleod**, professor of neurology and translational neuroscience at the University of Edinburgh. “So we’ve become very good at making papers, but the papers we make often don’t completely reflect what happened in our labs.”

In 2012, the Nature journals **developed a checklist** to help their authors better report their methods. But the list doesn’t clearly spell out how to handle interrelated data points, such as those from a single animal, says **Emmeke Aarts**, assistant professor of statistics at Utrecht University in the Netherlands.

In a 2014 *Nature Neuroscience* paper, Aarts reported that about half of the 314 mouse studies she had reviewed had not used an appropriate statistical model to analyze findings. She proposed that scientists use a type of statistical analysis that accounts for dependency in data[4]. The paper generated some buzz and has since been cited in more than 80 studies.

“I’m optimistic,” Aarts says. “It takes some time for things like these to sink in.”

Data do-over:

Ironically, Kind was unable to publish his 2011 review because none of the journals he approached were interested. He mentioned the work to colleagues at conferences, but found that even talking about pseudoreplication was sensitive. “It certainly doesn’t win you friends,” he says.

He is preparing to analyze fragile X studies published in top journals since 2011 to see whether there’s been a decrease in pseudoreplication since his initial analysis.

“My suspicion is that it might have gotten a bit better,” he says, “but not much.”

As for the paper that propelled Kind’s dive into the debate in the first place, it did not ruin his

friendship with the lead investigator. When Kind alerted the scientist, the latter immediately sent Kind all of his data.

“He just happens to be a very conscientious researcher who wants to know what’s right and what’s wrong,” Kind says.

Kind applied a statistical model to the data, relating it to each animal instead of to individual cells. The results remained statistically significant. This wasn’t just luck, Kind says: Apart from the unwitting pseudoreplication, the study was done carefully.

REFERENCES:

1. Hurlbert S.H. *Ecological Monographs* **54**, 187-211 (1984) **Abstract**
2. Lazic S.E. *BMC Neurosci*. **11**, 5 (2010) **PubMed**
3. Halsey L.G. *et al*. *Nat. Methods* **12**, 179-185 (2015) **PubMed**
4. Aarts E. *et al*. *Nat. Neurosci.* **17**, 491-496 (2014) **PubMed**