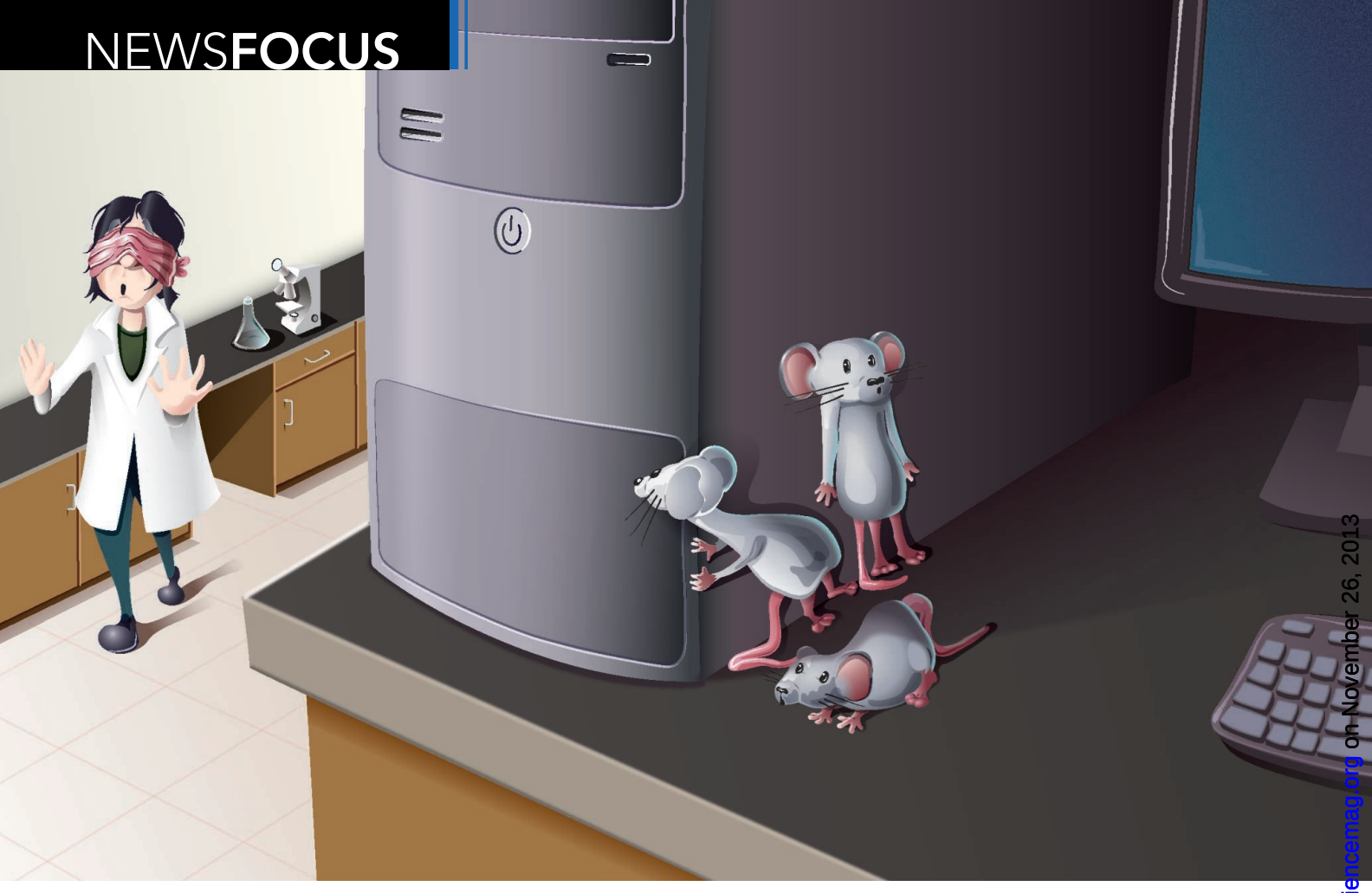



# When Mice Mislead

**Tackling a long-standing disconnect between animal and human studies, some charge that animal researchers need stricter safeguards and better statistics to ensure their science is solid**

**THREE MICE HAD VANISHED. AND ULRICH** Dirnagl had a hunch about where they'd ended up: in the metaphorical dustbin housing animals—and there are lots of them—that line up at an experiment's starting line but are discarded before the finish. The paper that Dirnagl, director of the Center for Stroke Research at Charité University Medicine Berlin, was reviewing described how a new drug protected a rodent's brain after a stroke. The authors used 20 mice, half of which got the therapy. But mysteriously, only seven of the 10 treated animals appeared in a graph analyzing the results.

"I wrote to the editor and said, 'I cannot judge this paper, I need to know where the three mice went,'" Dirnagl recalls. For 6 months, radio silence. Then, the editor responded. He'd heard from the authors, he told Dirnagl. The three mice, suffering from massive strokes, had died, and the authors had simply left them out of the paper. Extra analysis of their stroke drug, however, revealed that those mice had an important message to bear: The therapy harmed the brain rather than helping it.

"This isn't fraud," says Dirnagl, who often works with mice. Dropping animals from a research study for any number of reasons, he explains, is an entrenched, accepted part of the culture. "You look at your data, there are no rules. … People exclude animals at their whim, they just do it and they don't report it." That bad habit, he believes, is one of several that plague animal studies.

For years, researchers, pharmaceutical companies, drug regulators, and even the general public have lamented how rarely therapies that cure animals do much of anything for humans. Much attention has focused on whether mice with different diseases accurately reflect what happens in sick people. But Dirnagl and some others suggest there's another equally acute problem. Many animal studies are poorly done, they say, and if conducted with greater rigor they'd be a much more reliable predictor of human biology.

It's hard to generalize, of course: Animal studies cut across a massive swath of biology, tracking everything from the activity of single molecules in a healthy organ to side effects of a new drug poised for human testing. And many who stake their careers on animal studies conduct them with care, judiciously weighing how to structure their experiments and chasing the science wherever their furry subjects take it.

That said, even animal research that has a big effect on human drug studies—like the work Dirnagl reviewed—is governed by far fewer standards than clinical trials in people. There, volunteers are randomly assigned by

CREDIT: V. ALTOUNIAN/*SCIENCE*



*Published by AAAS*

computer to get a new drug or a placebo. Those running a trial are often blinded to who's in what category, preventing clinicians invested in a therapy's success from imagining hints of efficacy in patients they know are getting a new drug. And look up any clinical trial seeking volunteers and you'll see a long list of "inclusion" and "exclusion" criteria governing who can participate. If you have high blood pressure or if your cancer is being treated with a certain drug, you might be out of luck.

Animal studies rarely follow these rules. For ethical and cost reasons, researchers try to use as few animals as possible, which can mean minuscule sample sizes. Unblinded, unrandomized studies are the norm. In Dirnagl's words, "the way we do our research with our animals is stone-age."

From various quarters, there's pressure to change that. High-profile studies showing that preclinical results often cannot be reproduced are driving funders and researchers to seek solutions—as much to mend their public image as to guarantee sound science.

### The roots of bias

Dirnagl's concerns were sparked around the same time as a friend and colleague's across the English Channel. A decade ago, Malcolm Macleod, a Scottish neurologist at the University of Edinburgh, went hunting for new stroke therapies. He wanted to find compounds that had looked good in animals but had stalled there and that might be worth testing in people.

Macleod and his colleagues identified 603 drugs tested in animals, 374 of which had helped heal the brain. Of those, 97 had been tried in humans—and only one had worked. And that one, Macleod is quick to point out, wasn't tested because of animal data at all, but because it had already benefitted patients with heart attacks.

Startled by this chasm separating experimental animals and people, Macleod turned his attention to what was going wrong. One possibility, he reasoned, was that the therapy wasn't tested properly in humans—say the dose was too low, or it was given too long after a stroke. Another was that human testing had been appropriate, but the animals were simply a poor model of human stroke. And the third was that the drug wasn't tested properly in animals to begin with.

Macleod dug deeper. What he found alarmed him. Only 36% of the animal studies described randomly assigning animals to stroke treatment or placebo. Only 29% reported blinding. What's more, studies that didn't report randomizing and blinding—which was most of them—"gave substantially and significantly higher estimates of how good these drugs were," Macleod says. In one case, the effectiveness of a stroke drug was twice as high in the studies that didn't report randomizing as in those that did.

Macleod then turned to other neurological ailments: Alzheimer's, multiple sclerosis, Huntington's disease, Parkinson's, and pain. In animal studies of potential treatments, the situation was, if anything, worse than in stroke, the measures that might dampen bias applied even less often.

Many of these authors likely didn't recognize what Macleod perceived as lack of rigor in their studies because their mentors, and their mentors' mentors, had followed this same approach. "I was trained as an animal researcher," says Lisa Bero, now a health policy expert at the University of California, San Francisco. "Their idea of randomization is, you stick your hand in the cage and whichever one comes up to you, you grab. That is not a random way to select an animal." Some animals might be fearful, or biters, or they might just be curled up in the corner, asleep. None will be chosen. And there, bias begins.

Macleod's work, published in a series of papers beginning in 2004, is complemented by other strands of evidence. A 2008 paper in the journal *Amyotrophic Lateral Sclerosis* described efforts by the nonprofit ALS Therapy Development Institute to retest more than 70 compounds that had eased symptoms in a mouse model of the disease. Not a single one panned out.

It was what the ALS authors did next that was particularly interesting. At the end of their dismal replication efforts, they were left with a treasure trove of data on 2241 control animals—mice that hadn't gotten any active drug. The researchers randomly assigned mice to two groups, matched for sex, litter size, and other variables. Then they looked for differences in mean life expectancy—something they shouldn't see, because the two groups were essentially the same.

What they found was telling. If the two groups contained just four animals each, there was a 30% chance that an illusory life expectancy gap would show up. With 10 animals per group, the risk dropped to 10%. "You can imagine 10 labs doing this experiment," says Shai Silberberg, a program director at the National Institute of Neurological Disorders and Stroke (NINDS) in Bethesda, Maryland. "One gets an effect, and they publish it." The other nine are much less likely to submit a paper. Suddenly, the literature is skewed.

The numbers of animals that the ALS researchers used may sound small, but they're grounded in reality. A survey of 76 influential animal studies found that half used five or fewer animals per group.

Bero recently examined animal research of statins for heart disease. At the International Congress on Peer Review and Biomedical Publication in September, she reported that work funded by industry was less likely to endorse the drug in question than work from another funding source, maybe because companies don't want to pour millions of dollars into testing a treatment in people that's unlikely to help them.

### Status quo revisited

In Bethesda, Silberberg sits in a position of power, part of a committee advising the NINDS director on which of the most costly studies should be considered. About 3 years ago, Silberberg, who trained as a biophysicist in Israel and later the United States, grew more and more worried that the institute was greenlighting some projects that weren't based on solid science. He decided to do something about that.

There were lots of avenues Silberberg could have followed, and he settled on animals. In part, he was responding to data like Macleod's, with its startling evidence of what he saw as entrenched biases in animal research. A slice of NINDS's budget is funneled to translating animal studies to people. Among other things, Silberberg worried about "poor patients [who] are exposed to things they shouldn't be."

After lots of agitating and conversation within the National Institutes of Health (NIH), in the summer of 2012 Silberberg and some allies went outside it, convening a workshop in downtown Washington,

> *"The way we do our research with our animals is stone-age."*
>
> —**ULRICH DIRNAGL**, CHARITÉ UNIVERSITY MEDICINE BERLIN



Published by AAAS

D.C. Among the attendees were journal editors, whom he considers critical to raising standards of animal research. "Initially there was a lot of finger-pointing," he says. "The editors are responsible, the reviewers are responsible, funding agencies are responsible. At the end of the day we said, 'Look, it's everyone's responsibility, can we agree on some core set of issues that need to be reported' " in animal research?

In the months since then, there's been measurable progress. The scrutiny of animal studies is one piece of an NIH effort to improve openness and reproducibility in all the science it funds. Several institutes are beginning to pilot new approaches to grant review. For an application based on animal results, this might mean requiring that the previous work describe whether blinding, randomization, and calculations about sample size were considered to minimize the risk of bias. "Sometimes the fundamentals get pushed aside—the basics of experimental design, the basics of statistics," says Lawrence Tabak, principal deputy director of NIH, who is coordinating these efforts.

Another of NIH's ventures is at the National Institute of Environmental Health Sciences in Research Triangle Park, North Carolina, where toxicologist Kristina Thayer is looking for a way to grade animal studies, in part to guide regulators making recommendations about particular chemicals. For work that examines the hazards of bisphenol A, a compound found in many plastics, Thayer is experimenting with 15 "risk of bias" questions. Among them: Did the researchers randomly allocate animals to treatment groups? Did they know which animals were exposed to chemicals? Were experimental conditions the same across different groups of animals? "When you're looking at bias, it's not just yes or no," she says. "There can be different shades of gray, and there can be scientific judgment in there." The Environmental Protection Agency is also reconsidering how it evaluates animal data.

Journals, too, are getting in on the act. In April, *Nature* released a checklist for authors and reviewers, requesting extra detail about scientific methods in life sciences papers. Among other things, the checklist asks whether the animals were randomized and the researchers blinded, and requests the criteria by which animals were dropped from the study—an effort to avoid the three missing mice Dirnagl encountered. *Science Translational Medicine* announced a similar initiative in June, and *Science* is considering the same.

> *"I am not pessimistic enough to believe that the entire scientific community is obfuscating results."*
> —**JOSEPH BASS**, NORTHWESTERN UNIVERSITY

Some in the field consider such requirements uncalled for. "I am not pessimistic enough to believe that the entire scientific community is obfuscating results, or that there's a systematic bias," says Joseph Bass, who studies mouse models of obesity and diabetes at Northwestern University in Chicago, Illinois. Although Bass agrees that mouse studies often aren't reproducible—a problem he takes seriously—he believes that's not primarily because of statistics. Rather, he suggests the reasons vary by field, even by experiment. For example, results in Bass's area, metabolism, can be affected by temperature, to which animals are acutely sensitive. They can also be skewed if a genetic manipulation causes a side effect late in life, and researchers try to use older mice to replicate an effect observed in young animals. Applying blanket requirements across all of animal research, he argues, isn't realistic.

Bass is just as concerned about the undercurrent that scientists aren't to be trusted. "A lot of what this argument is, is that there's this ethical flaw across the community, and we're going to correct it by mandating these laws," he says.

Dirnagl agrees with this last point, even though he believes that new standards are needed. "A lot of the academic researchers, they are being accused of producing crap, complete crap," he says. "I think this is overshooting it, and it's even dangerous. … We need to properly discuss these quality issues." More importantly, "we need to teach them to the next generation." He tries to present his case with optimism, so as not to discourage or alienate his colleagues.

Dirnagl also says he's cleaned up his own act, something that, for the most part, hasn't been particularly onerous. He marks the tails of all his animals with numbers and uses a number generator that spits out a list to help him randomly select mice. If during a surgery an animal's blood pressure drops below a certain level, Dirnagl excludes it, whether it's getting a new stroke treatment or not. He's starting to do what clinical trialists have done for years—run multicenter studies, where labs pool their animals to boost the experiment's reliability with greater numbers.

One open question is whether such adjustments will help animal experiments hold up to scrutiny. "It's almost certain that we're not completely right" about what's worth changing and what's not, Macleod says, and that will need to be gauged over time. Ultimately, though, he believes better research standards will lead to a renewed trust in mouse models of disease. "I wouldn't be wasting all my time" on this, he says, if he didn't have faith that the mice had it in them to be auspicious guides—if only we could figure out the best way to use them.

**–JENNIFER COUZIN-FRANKEL**

CREDIT: V. ALTOUNIAN/*SCIENCE*



Published by AAAS