BURROUGHS
WELLCOME
FUND

EXPERIMENTAL QUALITY

INVESTING IN THE CAREERS OF YOUNG SCIENTISTS

THE BURROUGHS WELLCOME FUND is an independent private foundation dedicated to advancing the biomedical sciences by supporting research and other scientific and educational activities. Within this broad mission, BWF seeks to accomplish two primary goals—to help scientists early in their careers develop as independent investigators, and to advance fields in the basic biomedical sciences that are undervalued or in need of particular encouragement.

BWF's financial support is channeled primarily through competitive peer-reviewed award programs to degree-granting institutions in the U.S. and Canada on behalf of individual researchers. To complement these competitive award programs, BWF also makes grants to nonprofit organizations conducting activities intended to improve the general environment for science.

Governed by a Board of Directors composed of distinguished scientists and business leaders, BWF was founded in 1955 as the corporate foundation of the pharmaceutical firm Burroughs Wellcome Co. In 1993, a generous gift from the Wellcome Trust, enabled BWF to become fully independent from the company, which was acquired by Glaxo in 1995.

EXPERIMENTAL QUALITY

BURROUGHS
WELLCOME
FUND

21 T. W. Alexander Drive
P. O. Box 13901
Research Triangle Park, NC 27709-3901
info@bwfund.org
www.bwfund.org

INTRODUCTION:
EXPERIMENTAL QUALITY

What is research? In his book on experimental design,[1] David Glass of the Novartis Institutes for Biomedical Research in Cambridge, Massachusetts, puts it this way: ***Scientific research is the process of determining some property Y about some thing X, to a degree of accuracy sufficient for another person to confirm this property Y.***

In other words: If it needs to be accurate enough for someone else to confirm it, it needs to be reproducible. Therefore, what distinguishes research from playful observation is reproducibility.

But reproducibility seems to be in a crisis. The oncology research team at Amgen failed to reproduce 47 of 53 published preclinical cancer studies;[2] none of the effects of more than 100 compounds initially reported to lengthen life span in a mouse model of Amyotrophic Lateral Sclerosis (ALS) could be reproduced, and none were successful in human trials,[3] and the number of retracted studies has been increasing.[4]

While most were retracted because of misconduct, 21% of the retractions are related to sloppy science—contamination of reagents, mistakes in statistical analyses, or the authors' inability to reproduce their own data, according to a 2014 study.[5] "A lot of the science is very poorly done," says Arturo Casadevall from the Johns Hopkins Bloomberg School of Public Health in Baltimore, one of the study's authors.

And these retractions are likely only the tip of the iceberg, he says. Case in point: Only 1.4% are due to contaminated cell lines, even though studies have shown that about 15% of cell lines are contaminated with other cells, and 10-15% of cell cultures are contaminated with mycoplasma, a tiny bacterium. Therefore, Casadevall says, the true number of papers that use unreliable cell lines is likely far higher than those 1.4%.

This suggests that erroneous studies are often not retracted. One reason is that researchers who can't reproduce someone else's data have almost no place to publish their negative findings, because most journals have little interest in publishing them. The University of Washington's Ferric Fang, who coauthored Casadevall's 2014 retraction study,[5] says if a study finds that some drug has some effect, and 19 groups later find the opposite, much of that will never be published or it might be published in less prominent journals. It will appear, therefore, that the initial finding has never been challenged. "One of my postdocs cynically remarked to me, 'it's better to be first than to be right,'" he says.

The pressure to be the first to publish an interesting positive finding in a prestigious journal, coupled with the need to come up with a hypothesis to write research grants, can lead to confirmation bias, Fang says: There are labs, he says, where the principal investigator tells the staff, "'this is the result you are supposed to get.' It's not [about] the truth anymore, but rather [about] generating lots of data that seem to support the pet idea of the principal investigator."

Meanwhile, biological experiments are becoming ever more complex. One high throughput experiment can involve processing thousands or millions of data points. This means that experimental design and statistical knowledge are more important than ever, while training in these areas is often inadequate.

> “Experimental design and statistical knowledge are more important than ever, while training in these areas is often inadequate.”

This handbook tries to address some of these gaps. It will discuss the major traps researchers can fall into and how to avoid them. These include confirmation bias; unreliable reagents; small sample sizes; lack of blinding and randomization; the importance of standards; multiple testing and false positives; and recording and reporting experimental procedures and results.

But first, let’s take a look at how we got here: Why do we do science the way we do it, and is this approach still adequate today?

References:

1 David J. Glass: Experimental Design for Biologists, 2nd edition. Cold Spring Harbor Laboratory Press (2014)

2 *Nature* **483**, 531 (2012)

3 *Nature* **507**, 423 (2014)

4 *Proc. Natl. Acad. Sci.* USA **109**, 17028 (2012)

5 *FASEB J.* **28**, 3847 (2014)

Further reading:

Begley CG, Buchan AM, Dirnagl U: Robust research: Institutions must do their part for reproducibility. *Nature* **525**, 25 (2015)

Begley CG, Ioannidis JPA: Reproducibility in Science: Improving the standard for basic and preclinical research. *Circ. Res.* **116**, 116 (2015)

THE PHILOSOPHICAL FRAMEWORK

According to the instructions on how to write a research grant application on the web site of the National Institute of Neurological Disorders and Stroke (NINDS), most reviewers "feel that a good grant application is driven by a strong hypothesis. **The hypothesis is the foundation of your application. Make sure it's solid.** It must be important to the field, and you must have a means of testing it."[1]

This hypothesis-driven approach to experimental science has a long history. Initially, experiments were quite rare, says Edward Hall, a philosopher at Harvard University, who studies the philosophy of science. Instead, there were appeals to authority. "People would claim that something was true because Aristotle said so," he says. They made unproven general statements, assumptions or premises about how the world works, and deduced from them more specific conclusions. Because this approach derives its conclusions from these initial assumptions, it's called deductive reasoning, Hall says.

But Francis Bacon (1561-1626) warned that preexisting beliefs can bias the conclusions. In a way, he was the first who warned about confirmation bias. To prevent such bias, Bacon suggested we need to start with a clean slate free from any assumptions or beliefs. Instead, we should gather experimental evidence first and only then generalize principles from that, an approach called inductive reasoning. "He was shouting, 'you need to do experiments, don't just read your Aristotle,'" Hall says. By replacing preexisting ideology with experimental data as the basis of knowledge, Bacon launched a revolution in European science, Glass adds.[2]

David Hume (1711-1776), however, was skeptical: We can never know for sure, he argued, that the conclusions from our experiments are true, because we can't know for sure that things will keep behaving according to the same rules they have behaved by in the past.

However, while we can never be sure that something is true, we *can* know for sure that something is *false*, because just one observation that contradicts our assumption proves that assumption to be wrong. That's why Karl Popper (1902-1994) proposed that scientists should focus on what can be proven *false*, because that's the only thing they can be certain about. He suggested researchers should first come up with a hypothesis, which they should then try to falsify (not verify!) with their experiments.

This is the hypothesis-falsification approach most funding agencies are using today. But not everyone believes that this is the best way to go about science. That's because by abandoning Bacon's experiments-first approach, Popper actually went back to a modified version of deductive reasoning: Instead of starting from scratch with experimental data, one first needs a hypothesis. Therefore, the confirmation bias Bacon had originally warned against becomes a concern again, Glass says: Researchers might prefer observations that are consistent with the hypothesis.

Glass explains the problem with an example[2]: Assume you want to find out the color of the sky. If you have to come up with a hypothesis, that hypothesis would likely be, "The sky is red." Because you'd then only do experiments to falsify this hypothesis, they would likely be overly narrow: All you'd do is measure red versus non-red. Eventually, you'd probably do some of your measurements during dawn or dusk, when the sky is indeed red, and conclude that you cannot falsify the hypothesis that the sky is red and therefore assume that the color of the sky is likely red.

“The hypothesis is a dangerous framework... if it is used only to feed the scientist’s hubris.”

David Glass
Novartis Institutes for Biomedical Research

The example illustrates two problems the hypothesis-falsification approach might create for reproducibility and experimental quality. First, it might indeed create a confirmation bias that will cause a researcher to value the “red” data more than the—more common—“non-red” data. Because the scientist has to come up with the hypothesis first, s/he will at least subconsciously prefer experimental results that fail to falsify that hypothesis—that the sky is red. Why? Because there is a strong desire in each of us that we want our predictions to turn out to be correct. The role of the experiment, then, “is to manifest the scientist’s brilliance,” Glass argues.[2] “The hypothesis is a dangerous framework, in this sense, if it is used only to feed the scientist’s hubris.”

Second, the hypothesis-falsification approach can cause scientists to miss the most important data (in this case, that the sky is in fact blue and black most of the time, and not red). That’s because it only allows for two outcomes—falsification or no falsification—and as a result creates a data filter that’s likely too narrow compared with what’s really going on.

What’s more, for many systems biology projects, where biologists “screen” entire biological systems like genomes or proteomes, the hypothesis-falsification approach doesn’t work, Glass says. Take the sequencing of the human genome: There is no reasonable hypothesis, Glass says, that could possibly be falsified by having the genome sequence in hand. For example, the hypothesis that there are 10 genes at least 50% homologous in sequence to the insulin gene, he says, is too narrow to justify sequencing the entire human genome to try to falsify it. Broader hypotheses like “understanding the genome would speed the search for cures to human disease” aren’t falsifiable either, he argues, because we can’t go back in time to see whether cures would have appeared as quickly without having the genome sequence.[2]

Because of these issues, some have started to call systems biology experiments “hypothesis generating.” To Glass, that’s just another way of saying that the hypothesis-falsification framework just doesn’t work for such experiments.

Therefore, Glass argues, we need to come up with a new scientific framework that avoids these problems and shortcomings. He suggests starting with a question instead of a hypothesis. This, he says, not only sets the right frame for systems biology experiments (“What is the sequence of the human genome?”), but also avoids confirmation bias. That’s because a question puts equal weight on both possible answers, while a declarative sentence that states a hypothesis (“The sky is red”) puts the emphasis on answers that are consistent with the hypothesis.[2]

Once we have experimental data that answer the initial question, Glass says, we can use them to formulate a model. Then the researcher can ask further questions to test the validity of the model and refine it if necessary. For example, Glass says, one could start with a question like, “does gene X resemble any proteins of known function?” To find out, we would obviously sequence the gene. If the sequence tells us that it resembles, say, ubiquitin ligases, we can ask another, more detailed question: “Does gene X function as a ubiquitin ligase?”[2]

There’s another reason why coming up with a hypothesis can be a problem in biomedical sciences today: Researchers often use their intuition to come up with hypotheses that make sense to them, Hall says, but intuition can be misleading, especially as the systems they deal with are becoming more complex. Empirical science like biomedical research, he says, is at the same point mathematics was at in the mid-19th century: Until then, there were no clearly defined standards of what counted as an acceptable method of proof, which is why intuition would often play a role in considering a mathematical proof acceptable.

But by the 19th century, intuition started to conflict with the evidence. For example, the fact that one can have two infinities with different sizes is not at all intuitive. So mathematicians came up with standard formal logic to replace intuition. "What is so striking is that we have nothing remotely analogous to that in the case of empirical science," Hall says. At the same time, many biomedical researchers are probably unaware that their intuition is far from reliable—another reason why asking questions might be better than coming up with a hypothesis.

Now that we've discussed how today's philosophical framework of science can affect reproducibility, let's discuss the more practical aspects. We'll start with the things you'll need to do to prepare first to be ready for your experiments: Your reagents.

References:

1 http://www.ninds.nih.gov/funding/write_grant_doc.htm

2 David J. Glass: Experimental Design for Biologists, 2nd edition. Cold Spring Harbor Laboratory Press (2014)

Further reading:

Glass DJ, Hall N: A Brief History of the Hypothesis. *Cell* **134**, 378 (2008)

Glass DJ: A Critique of the Hypothesis, and a Defense of the Question, as a Framework for Experimentation. *Clin. Chem.* **56**, 1080 (2010)

CHECK YOUR REAGENTS

"Young Blood May Hold Key to Reversing Aging," the *New York Times* headlined in May 2014,[1] after Amy Wagers of Harvard University and her colleagues reported new findings on GDF11, a protein they had previously found to decrease with age in the blood of mice: The protein seemed to be one of the factors responsible for age-reversal effects such as muscle regeneration when blood of young mice was infused into older mice.[2]

A year later, however, Glass and colleagues reported the opposite: GDF11 levels didn't decrease with age in the blood of mice, and the protein seemed to *inhibit* muscle regeneration.[3] One reason for the discrepancy, they reported, was that the antibody used in Wagers' original study didn't just recognize GDF11, but also the closely related GDF8 (also known as myostatin).

In a more recent paper, Wagers and her colleagues confirmed that the antibody recognizes both GDF8 and 11; however, they found that it also recognizes immunoglobulins, whose levels increase with age in the blood of mice.[4] This, Wagers says, means that the GDF11 increase Glass and colleagues reported is in fact due to the increase in immunoglobulin levels and not GDF11.

The case shows that researchers need to be aware that the reagents they use in their experiments might not be as reliable as they believe them to be. While antibody results are mostly reliable, Wagers says, it's always possible an antibody recognizes some as unknown protein targets—unless you exclude that possibility by testing it "against every other protein and protein conformation in the entire biological universe." Because that's virtually impossible to do in practice, it's important to complement results from antibody experiments with different approaches, says Wagers, who is currently developing a mass spectrometry-based assay and other assays to test her GDF11 observations.

Unreliable antibodies are just one of many problems; in their study last year, Casadevall and colleagues identified contaminated reagents as one of the most common reasons why papers were retracted due to scientific error.[5] In other cases, researchers failed to check whether the mice they were using really only lacked the one gene they were studying, or didn't test a script of computer code with known data before using it with the data they were studying.

The solution, Casadevall says, is obvious: Check and validate your reagents, cell lines, animals or computer code—before you start an experiment. One reason those things are often not done is a lack of communication in the lab, something Keith Micoli, a former postdoc who now directs the New York University School of Medicine Postdoctoral Program, learned the hard way: "Five people independently bought the same antibody from the same company to do the same test, and all found that it didn't work," he says. "This never came up at lab meeting, because [we didn't] talk about experiments that don't work. So our lab wasted thousands of dollars."

So to address problems as soon as they arise, make sure you hold regular lab meetings where people don't just give polished Power Point presentations, but can discuss primary data—without being afraid to speak up about things that failed. And having a big lab isn't an excuse not to do so, Micoli says, because senior people in the lab could hold the meetings instead of the principal investigator.

Next, let's take a more detailed look at some common reagent-related problems, and solutions for preventing them. This list is by no means complete, but serves to illustrate the kind of measures you can take to prevent problems.

> “Check and validate your reagents, cell lines, animals or computer code—before you start an experiment.”
>
> **Arturo Casadevall**
> **Johns Hopkins Bloomberg School of Public Health**

Unreliable Antibodies Are Actually Quite Common

A 2009 analysis of the quality of 49 antibodies that were supposed to be specific for certain G protein coupled receptors found that most of them bound to more than one receptor,[6] and a 2011 analysis found that about a quarter of 246 antibodies used in epigenetic studies bound to more than one target.[7]

There are many possible reasons why antibodies are unreliable; in Micoli’s case, the problem started once the company making the antibody he’d been using sold the rights to another company. “It no longer worked nearly as well,” he says. “We spent months trying to figure out what had gone wrong.” Eventually, after talking to people in other labs, he found that all of the people who had problems had purchased the same antibody from the same company within a few months of one another.

Solutions: Make sure you check and validate the antibodies you are working with before doing the experiment, and test them with positive and negative controls. Also, look for validated antibodies when you buy them. For example, antibodies-online.com, the world’s largest antibody market place, is now adding a green seal of approval to antibodies that have been validated. The seals are the result of an effort by the company Science Exchange to validate thousands of antibodies on the site.[8]

RNAi

Knockdown experiments with RNAi are notorious for off-target effects and variability in the level of the knockdown, says Harvard Medical School cell biologist Randall King. The authors of most RNAi screens, he says, don't sufficiently validate their results, suggesting that many studies may contain false positive results.[9] "I think it's just an example of a technique that's inherently not [as] robust" as some other approaches like genetic knockouts, he says.

Solution: Make sure you do an independent validation, King says, by reproducing a similar effect as the one you're reporting with a completely different RNAi that's designed to target the same gene in a different, independent way.

> “An estimated 10–15% of cell cultures are thought to be infected with mycoplasma.”

Mycoplasma Contamination of Cell Lines

An estimated 10-15% of cell cultures are thought to be infected with this tiny bacterium that’s too small to see under the light microscope and not sensitive to antibiotics that can be used with cultured cells because it lacks cell walls, says James Deatherage, chief of the cell biology branch at the National Institute of General Medical Sciences (NIGMS), referring to a recent study of over 9,000 RNA sequence data from cultured mammalian cells that found mycoplasma sequences in 11%.[10]

Solution: Check for mycoplasma sequences by polymerase chain reaction (PCR). If you find contamination, the only certain way to get rid of mycoplasmas is by throwing away the infected cells and starting over, Deatherage says. And make sure you report that you’ve done the mycoplasma check in your paper, he adds: In a recent analysis of 101 randomly selected NIH-funded papers, Zhongzhen Nie of NIGMS found that none of them reported whether they checked their cells for mycoplasma. If you don’t report it, the reader has no way to know whether the test was done at all, Deatherage says.

Cell Lines Contaminated With Other Cell Lines

This is also a widespread problem: A 1999 study showed 18% of 252 tumor cell lines to be contaminated with different cell lines,[11] and a 2003 study found 14.9% of 550 leukemia cell lines contaminated.[12] Often the contaminating cells are the famous HeLa cells, Micoli says, because they are especially aggressive in the way they grow. “One eventually grows out to dominate the other,” he says, “and now you are using a cell line that’s not what you think it is.”

This type of contamination is especially problematic when studying processes that likely differ between cell types, like the regulation of induced cell death (apoptosis), Deatherage says: In painstakingly detailed experiments, he says, Peter Sorger at Harvard Medical School has shown that differences in the level of one or two proteins in the pathways that trigger apoptosis typically observed in different cell types can lead to opposite results.[13] As a result, cancer drug candidates that work by inducing apoptosis might have different effects in a misidentified cell line, for example.

Solutions:

- Check if you are working with a cell line that has been reported by others as contaminated or misidentified. The International Cell Line Authentication Committee (ICLAC) maintains a list of such false cell lines on their web site.[14] If there is an authentic stock of a cell line on the list, make sure you get it from a trusted vendor such as ATCC, Coriell or DSMZ, Deatherage says. This may cost a couple of hundred dollars, Casadevall says, "but if you work [for] a year on the wrong thing, you're talking about [wasting] thousands and thousands of dollars." Otherwise, don't use false cell lines: If no authentic stock exists, Deatherage adds, you should not use it at all.
- If you are working with a human cell line, you can (and should) check the identity of your cell lines, for example by PCR amplification of a set of short tandem repeat loci on the DNA.[15] Doing so is affordable now, Deatherage says: "You can buy a kit for $1000 and do a cell line for $25 to $40," he says. "There is no excuse anymore for not authenticating your cells." You can then compare the result with the profiles suppliers like ATCC have on their web sites.[16] For non-human cell lines, however, such profile databases aren't available yet, Deatherage says.
- Make sure you report in your publication where you got your cell lines from and that you authenticated them. That will go a long way enabling others to trust your results and reproduce them. In an analysis of 101 randomly chosen NIH-funded papers, Zhongzhen Nie found that 43% either didn't indicate where they got their cell lines or they got them from another lab instead of a trusted repository; none of the labs that got the cells from another lab reported that they authenticated them.

"Every time someone starts a series of experiments, get them a fresh cell line from a trusted vendor."

High or unknown passage number of cell lines.

The number of passages a cell line went through can make a difference as to how the cell line behaves biologically. "I've seen real differences," Micoli says, between cells passaged less than 10 times compared with more than 15 times. The resulting changes can be quite dramatic. They include rearrangement or loss of chromosomes, Micoli says, all of which can lead to changes in gene or protein expression. In fact, Deatherage adds, studies have shown that an increase from 18 to 40 passages can affect mRNA expression, secretion, adhesion and proliferation. As a result, he says, "you wouldn't be able to rely on high passage cells for any kind of physiological conclusions like drug responses."

Solution: Every time someone starts a series of experiments, get them a fresh cell line from a trusted vendor. Make sure you expand and freeze them after very few passages immediately after they arrive and then distribute aliquots, keeping a record of the number of passages the cells went through. And make sure you report the passage number in your publication.

References:

1 http://nyti.ms/1hpGiUH

2 *Science* **344**, 649 (2014)

3 *Cell Metab.* **22**, 164 (2015)

4 *Circ. Res.* 2015, doi: 10.1161/CIRCRESAHA.115.307521

5 *FASEB J.* **28**, 3847 (2014)

6 *Naunyn Schmiedebergs Arch. Pharmacol.* **379**, 385 (2009)

7 *Nat. Struct. Mol. Biol.* **18**, 91 (2011)

8 http://www.bwfund.org/newsroom/newsletter-articles/special-report-biomedical-research-are-all-results-correct

9 *ACS Chem. Biol.* **6**, 47 (2011)

10 *Nucleic Acids Res.* **43**, 2535 (2015)

11 *Int. J. Cancer* **83**, 555 (1999)

12 *Leukemia* **17**, 416 (2003)

13 *Mol. Syst. Biol.* **7**, 553 (2011)

14 http://iclac.org/databases/cross-contaminations/

15 http://iclac.org/wp-content/uploads/Authentication-SOP_09-Jan-2014.pdf

16 http://www.atcc.org/en/STR_Database.aspx

Further reading:

Baker M: Reproducibility crisis: Blame it on the antibodies. *Nature* **521**, 274 (2015)

Sigoillot FD, King RW: Vigilance and validation: Keys to success in RNAi screening. *ACS Chem. Biol.* **6**, 47 (2011)

Callaway E: Contamination hits cell work. *Nature* **511**, 518 (2014)

Chatterjee R: Cell biology. Cases of mistaken identity. *Science* **315**, 928 (2007)

SET UP YOUR EXPERIMENTAL SYSTEM

Once you've validated your reagents, you need to set up the experimental system.

First, there is sample size: Just how many animals or samples do you need per experimental group to get meaningful results? This isn't so important when you do exploratory research just to see what's there like an initial screen where there are no controls. But in any experiment where you compare a specific experimental group with a control to test a specific question, sample size is crucial.

Even so, sample sizes in published studies often seem to be inadequate: When Shai Silberberg at the NINDS checked 76 high impact preclinical animal studies that had been cited more than 500 times (compiled in ref.1), he found that half used five animals or less per experimental group, and five didn't report the sample size at all. "You'll hardly ever find anyone telling you how they estimated the sample size; it's [often] just pulled out of a hat," says Silberberg, who helped organize an NINDS workshop in June 2012 that, among other things, called for appropriate sample size estimation in preclinical research[2] and who has been involved in the development of training videos that discuss reproducibility issues.[3]

Sample sizes this small can lead to misleading results, Silberberg says, referring to a 2008 study,[4] where researchers with the non-profit ALS Therapy Development Institute (ALS TDI) in Cambridge, Massachusetts, compiled experimental data from more than 2,000 mice overexpressing a mutated human *SOD1* gene (which simulates ALS) that had never been treated with any experimental drug. They randomly assigned animals from this database to two different groups and checked for differences in life expectancy between the groups—a readout for experiments that test ALS candidate drugs.

Because none of the mice had been treated in any way, any statistically significant life expectancy differences between the two groups could only be due to chance. However, they found that groups of four animals had a statistically significant difference in life expectancy in 30% of cases, and even groups of ten animals still differed in 10% of cases.[4]

"The numbers are startling," Silberberg says: If ALS studies typically use 10 or fewer animals per experimental group, then at least one in 10 ALS candidate drug studies with the *SOD1* mouse model probably discovered a life expectancy effect that is just the result of chance. Combine that with the fact that negative studies usually go unpublished, he adds, and "you can easily imagine that nine groups tried to do the experiment [and] didn't see anything. And one group by chance got a significant difference, and that's what makes its way into the literature."

It's not too much of a surprise, then, that the track record of ALS candidate drugs identified in the *SOD1* mouse model is far from good: When ALS TDI researchers tried to reproduce the results of the more than 100 drug candidates previously reported to slow down ALS in the *SOD1* mouse model, none of them had an effect.[5] One of them, minocycline, had initially been reported to have a big effect in mice,[6;7] but in a phase III randomized clinical trial with 412 patients that was funded by NINDS, patients who took the drug actually deteriorated faster.[8;9] So "with what we know now," Silberberg says, NINDS might have chosen to first reproduce the preclinical studies with larger groups of animals before funding a clinical trial.

These examples should make it clear that it's very important to determine what sample size you need to get meaningful results. To do so, you'll first need to get a feel for how your measurements—the effect size—compare with the variability around the negative controls. That's essentially what the ALS TDI researchers determined when they checked the variability in their untreated *SOD1* mice. Once you've done the same for the positive controls, you should be able to tell if there is enough of a range between the positive and negative controls to measure a meaningful effect.

For high throughput screens, King says, you should run a series of positive and negative controls to then calculate the so-called z prime factor. That's a number between 0 and 1, King says, that indicates the difference between positive and negative controls and the variability or standard deviation around them. A number between 0.5 and 1 usually means that the assay quality is acceptable, i.e. that the effect size compared with variability is big enough to give you a measurable signal.

Next, you'll use the information on variability and effect size to determine the sample size, like the number of animals per experimental group you'll need to get meaningful results. If the variability around positive and negative controls is big, or the effect size small, Silberberg says, "you obviously need more animals." Also, make sure that you later report the sample size and how you determined it in your paper.

This is a good time to consult a statistician, says Micoli, who learned the hard way what happens if you don't: After completing a study of how a drug affects the skeleton of mice, he found out that he didn't have enough mice for statistically robust conclusions. "We consulted with a biostatistician, who told us very clearly at least what range of the number of animals we needed to use," Micoli says. "We [eventually] repeated [the study with] the right number of mice. If we had just added 20 more mice to our initial study, we would have not wasted that much time."

As long as your sample size is big enough to get meaningful results, consider replicating an animal experiment rather than doing a single experiment with all animals at once, Fang says. This will allow you to make sure that you didn't just get the result because there was something unique about the group of animals you tested (like a certain kind of stress, or an infection). "It could be just a one-off thing that happened in that particular batch of mice," he says.

For *in vitro* studies, the equivalent to sample size is how often you need to replicate the experiment. Keep in mind that these replicates need to be truly independent biological replicates and not just technical replicates, where you split the same experiment into several groups. For example, say you are testing the effect of a compound on a set of three tissue culture dishes at the same time. If you do this at three different times, you've really only done three and not nine biological replicates. That's because each time you do the three-dish experiment, you do it with the same solution and at the same time, which is the same as if you split the experiment into three groups or technical replicates. Similarly, if you give a compound to a pregnant mouse that gives birth to ten pups and study the effect on the pups, you've really only done one experiment and not ten because all newborns come from the same mother who got the drug, Silberberg says.

> “As long as your sample size is big enough to get meaningful results, consider replicating an animal experiment rather than doing a single experiment with all animals at once.”
>
> **Ferric Fang**
> **University of Washington**

Also, make sure you plan to include controls each time you do an experiment, even if you’re just repeating it. “[One] shortcut people take is eliminating their controls in each experiment, after they’ve done it enough times [to make them] feel like they know the system is working,” says Micoli. “That’s something you should never do. Those should always be part of your experiment.”

References:

1 *JAMA* **296**, 1731 (2006)
2 *Nature* **490**, 187 (2012)
3 http://www.nih.gov/science/reproducibility/training.htm
4 *Amyotroph. Lateral Scler.* **9**, 4 (2008)
5 *Nature* **507**, 423 (2014)
6 *Nature* **417**, 74 (2002)
7 *Neuroreport* **13**, 1067 (2002)
8 *Lancet Neurol.* **6**, 1045 (2007)
9 *J. R. Coll. Physicians Edinb.* **38**, 35 (2008)

Further information:

NIH video module 3 (“Biological and Technical Replicates”) and 4 (“Sample Size, Outliers, and Exclusion Criteria”) (http://www.nih.gov/science/reproducibility/training.htm)

Lazic SE: The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci.* **11**, 5 (2010)

MAKE A PLAN AND STICK TO IT

As we'll explain in detail in the chapter on multiple testing, **it is very important that you make a plan** of exactly what you'll measure in the experiment and how often you'll measure it and then not only stick to that plan but also report it (and any deviations from it) in your paper.

The reason is that often researchers measure many more things than they later use in their results, thinking that it's OK to only use the "best" readings, or they later remove what they believe to be "outliers," because these disagree with what they believe the results should look like.

That, however, carries the risk that your results will only amplify your confirmation bias. Another risk is that you might take a look, statistically speaking, at your data more often than you think you are, which increases the risk of false positives and spurious results (this is called multiple testing). Making a research plan first and sticking to that, in the same way clinical trial researchers have to register their trial design beforehand, can prevent these problems.

BLIND AND RANDOMIZE

In October 2006, the drug maker Astra Zeneca announced that a clinical trial called SAINT II of the antioxidant compound NXY-059 for the treatment of stroke had failed, and **abandoned its efforts to seek FDA approval** for the compound.[1]

How could this happen? After all, the preclinical studies the trial was based on had clearly reported an effect. However, a 2008 analysis of preclinical animal studies of NXY-059 showed that the biggest effects were observed in the very preclinical studies that didn't report randomization and blinding,[2] suggesting that lack of blinding and randomization might have inflated the results. The same seems to be true for preclinical animal studies in pain and multiple sclerosis research, Silberberg says: The largest effects come from studies that aren't randomized or blinded.

Lack of blinding also seems to make cancer studies unreliable, says C. Glenn Begley, chief scientific officer at TetraLogic Pharmaceuticals: When Begley, who previously led the Amgen team that failed to reproduce 47 of 53 preclinical cancer studies, asked some of the original authors to reproduce their own experiments in their own labs—with the only difference being that the experiments were now blinded—most were unable to do so.[3;4]

This raises the question: How can lack of blinding and randomization lead to chance findings that are difficult to reproduce? Aren't animals, for example, already randomly mixed in their cages?

Often they're not, Silberberg says. "I'll give you an anecdote someone told me in one of my talks," he says. "They said that they never used to randomize the animals because the animals arrive at the institution and then they get transferred to cages. So they said that's randomized already. But it turns out that when the people transfer them to cages, they typically first pick up the more docile animals because it is easier to catch them. So those will be in one cage, and the other cage will be the ones that are more active. Or you've got a cage which is closer to the door and therefore affected by draft and another cage is further in the room and in a quieter place. Or you've got one aggressive animal in the cage which affects all the others and puts them at stress. So there are many many ways where you can think that it doesn't matter and it does."

For *in vitro* studies, Silberberg says, randomization is just as important. Imagine, he says, you're doing a dose response curve of a drug effect on ion channels: If you just gradually go from low to high doses, you might get desensitization, something randomization would avoid.

So how do you randomize? With animals, that's often quite easy, says Fang: Just toss a coin—or let the computer generate a random number—to determine which animal goes into which group.

Blinding is also not very hard: Just remove the labels. When Fang scores samples from mice treated in different ways for pathology, "we'll just send them to our pathologist with no information, and they'll score them," he says. "I think that those kinds of tricks from clinical research are very useful for wet labs."

“Whenever you’re comparing an experimental and a control group with a specific question or model in mind, it’s better to do so in a randomized and blinded way.”

Shai Silberberg
NINDS

Blinding can also help you avoid another common problem, says Martina Bremer, a statistician at San Jose State University. Researchers tend to clean up the data for example by removing outliers if they believe they already know what the result of an experiment should look like. That’s unlikely to happen if you remove the column headings and ask someone else to analyze the data.

So make sure you blind and randomize your samples. This is less important in exploratory research like sequencing a genome or an initial screen where you don’t have controls, says Silberberg. But whenever you’re comparing an experimental and a control group with a specific question or model in mind, it’s better to do so in a randomized and blinded way.

References:

1 *Science News* **172**, 26 (2007)

2 *Stroke* **39**, 2824 (2008)

3 *Nature* **483**, 531 (2012)

4 http://www.bwfund.org/newsroom/newsletter-articles/special-report-biomedical-research-are-all-results-correct

Further information:

Landis SC et al.: A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* **490**, 187 (2012)

NIH video module 2: “Blinding and Randomization” (http://www.nih.gov/science/reproducibility/training.htm).

Begley, CG: Six red flags for suspect work. *Nature* **497**, 433 (2013)

Steward O & Balice-Gordon R: Rigor or Mortis: Best Practices for Preclinical Research in Neuroscience. *Neuron* **84**, 572 (2014)

TEST AND REPLICATE

A few years ago, a postdoc in Fang's lab had a problem. He was getting inconsistent results in biochemical experiments with Nitrosoglutathione, a compound that releases nitric oxide (NO). Eventually, he reasoned that **something about the way he was doing his experiments wasn't consistent.**

“People unwittingly were just coming into the warm room flipping the lights on and off randomly, not realizing [this] was an important variable,” remembers Fang. His postdoc tested that by sticking an electrode in his solution and flipping on the light, and indeed, Fang remembers, “you could see this big spike of NO release. So he started shielding all of his reaction vessels and doing the experiments without the lights on and everything settled down.”

This is why Fang tells people to initially try to do their experiments in a way that’s as consistent as possible—like setting up or harvesting a bacterial or cell culture at the same hour of the day. “You never know if there is a hidden variable that might cause a problem with reproducibility,” Fang says. So initially, make sure you get consistent results, by trying to *avoid any variation* in your experimental conditions.

But once you've got your initial consistent results, you'll need to do exactly the opposite: Test and replicate if these results hold up under *variable* conditions. That's because you'll need to make sure the result is robust enough to survive the variability and change in conditions that will undoubtedly occur if someone else in a different lab tries to repeat the result. This is especially important if your initial results are just barely statistically significant, Fang says. A robust finding, he adds, is "not only seen on March 14th when you have a full moon, but really can be done under lots of circumstances in lots of different locations where you change [variables]."

Introducing variation will at first be difficult, because most researchers are trained to minimize it so they can get significant results even with small effect sizes. Can't reproduce a result with different batches of an enzyme? Chances are your PI will tell you to stick with the same batch.

Instead, you need to realize that other lab that will try to reproduce your result will likely do so with a different batch. So don't be afraid to introduce variability. Variable repeats of the experiment will ensure that your effect size is bigger than any of these kinds of sources of variability, King says.

So what should you vary? First, there are the kinds of experimental conditions you know might vary, like using a different assay, a reagent from a different supplier, different batches of a reagent, a cell line with a different number of passages or cell lines grown in a different batch of serum.

“You’ll need to make sure the result is robust enough to survive the variability and change in conditions that will undoubtedly occur if someone else in a different lab tries to repeat the result.”

The serum variation is especially important, Deatherage says. A common practice is that labs test different serum batches for their cell line experiments and then buy big amounts of the one batch where they see an optimal phenotype or response. But that’s a problem for reproducibility, he says, because other labs likely won’t have access to that same batch, won’t be able to reproduce the initial result, and won’t know why. So in a way, by creating optimal but narrow conditions for your experiment and then not disclosing this serum batch dependence in the publication, you’ll make it much more difficult for others to reproduce it. “People [may] fine-tune the conditions so that their experiments are giving a good strong signal,” Deatherage says. “It could take people working from the published papers years to figure out how to get a similarly strong response. And that’s just to get started to repeat the experiments.” So repeat your experiment with a range of different serum batches, and then report the entire range of results you got with the different batches in your paper, or at least disclose that the experiments are sensitive to serum batch.

But it gets trickier, because there are also many unknown factors that might affect the results: There could be differences in the water in different labs or subtle differences in the way people do experiments. What’s more, studies have shown that mice handled by men seem to have a lower pain response than mice handled by women,[1] and that mice exposed to Salmonella during the day (when they rest) are more susceptible to infection than mice exposed at night (when they are active).[2] Learning about the Salmonella study “was really quite frightening to me,” says Fang, who, after all, studies Salmonella infections himself. “This is not a variable that I would normally take into account.”

So how could you ever anticipate variation in such unknown factors? The best way to simulate this kind of variation is to have someone else in another lab repeat the experiment. The second best way is to have a different person in the same lab do it. That's how King came to realize that the way you dilute small molecules can sometimes cause them to precipitate. "It sort of controls for those variables that you might not have thought about when the first person was doing the experiment," he says.

Asking someone else to repeat the experiment will also force you to communicate all necessary details. "By having somebody else do the experiment who is maybe not as trained in the technique, not quite as expert, you realize what you really have to communicate [so they can] perform your experiment reproducibly," says King. When it's time to write the paper, make sure you include all these details in your materials and methods section.

References:

1 *Nat. Methods* **11**, 629 (2014)

2 *Proc. Natl. Acad. Sci.* USA **110**, 9897 (2013)

“Asking someone else to repeat the experiment will also force you to communicate all necessary details.”

Randall King
Harvard Medical School

LEARN STATISTICS (AND CONSULT A STATISTICIAN)

Weak statistics knowledge is another factor that can make studies unreliable. “I think we could really improve our formal statistical training for young scientists,” says Fang. Silberberg puts it more bluntly: **“Most people don’t have a clue how to use statistics.”** Teaching statistics is beyond the focus of this book, but let’s discuss a few examples that illustrate the problem.

One common issue seems to be a lack of understanding of whether data are correlated or independent from each other. Fang, editor-in-chief of the journal *Infection and Immunity*, recently commissioned an analysis of 110 papers that had appeared in April and May 2013 in the journal.[1] “Probably the most common major pitfall,” the authors wrote, was treating data as independent from one another when they weren’t: Measurements from the same animal taken at different times, for example, aren’t independent from each other. As a result, researchers used the wrong statistical tests: The paired t test, for example, is only appropriate for independent data; otherwise, the unpaired t test is a better choice.

But even the t test is only appropriate in some cases: when comparing the means of data that are distributed “normally,” like a bell curve—the way height in the human population is distributed, for example. Distributions different from that require different tests. For example, life times are distributed exponentially (with many small and few large numbers), which means that tests like Mann-Whitney or Wilcoxon are more appropriate.

One reason researchers use inappropriate tests is that they run their data through different tests until one gives them significant results, Micoli says: “As computer tools got better, I’ve known people who push their data through Excel and just do every possible statistical test in the hopes that one of them will give them a p-value less than 0.05.”

Confidence intervals are also often confusing. A 2005 study found that fewer than 20% of researchers correctly understand that partially overlapping confidence intervals (or standard deviations) don’t necessarily mean that two results aren’t significantly different from each other.[2-4]

And often researchers seem to assume that as long as they have a p value below 0.05, they have meaningful results. But effect size is just as important, says Glass. Take, for example, a 2012 meta-analysis of over 22,000 people, half of whom took aspirin for five years while the other half didn’t.[5] The aspirin-takers had a reduction in the risk of heart attacks that was highly statistically significant: The p value was smaller than 0.00001. But their actual risk—the effect size in this case—was only reduced by 0.77%, which is likely smaller than the risk of side effects.[6] “People just don’t appreciate what a p value means versus an important effect,” Glass says.

> “As computer tools got better, I’ve known people who push their data through Excel and just do every possible statistical test in the hopes that one of them will give them a p-value less than 0.05.”

Keith Micoli
New York University School of Medicine

References:

1 *Infect. Immun.* **82**, 916 (2014)

2 *Psychol. Methods* **10**, 389 (2005)

3 https://www.cscu.cornell.edu/news/statnews/stnews73.pdf

4 *J. Insect Sci.* **3**, 34 (2003)

5 *Am. J. Cardiol.* **107**, 1796 (2011)

6 *J. Grad. Med. Educ.* **4**, 279 (2012)

Further reading:

Gerald van Belle: Statistical Rules of Thumb. 2nd edition. *Wiley* (2008)

Martina Bremer, Rebecca W. Doerge: Using R at the Bench: Step-by-Step Data Analytics for Biologists. Cold Spring Harbor Laboratory Press (2015)

Joseph K Blitzstein, Jessica Hwang: Introduction to Probability. CRC Press (2014)

Andrew Vickers: What is a p-value anyway? 34 Stories To Help You Actually Understand Statistics. Pearson. (2009)

David S. Moore, George P. McCabe, Bruce A. Craig: Introduction to the Practice of Statistics. 8th edition. W. H. Freeman (2014)

BEWARE OF MULTIPLE TESTING

In 2006, Canadian researchers found that **Sagittarians are 38% more likely to break a leg** than people of other astrological signs, after checking the reasons why residents of Ontario province had unplanned stays in the hospital.[1]

But they didn't believe their own results, because they designed their study with a flaw: They checked so many different health problems for associations with certain astrological signs that some of them, just by chance, would show up as statistically significant. They wanted to make a point: That researchers often look at their data many times and in many ways, and once they find one unusual result that's statistically significant, they report it.

This is referred to as the multiple testing problem: It's as if you roll a pair of dice until you get snake eyes—and then pretend you rolled it once, says S. Stanley Young, former director of bioinformatics at the National Institute of Statistical Sciences. Assuming the usual p value for statistical significance of 0.05, checking 100 such associations will, on average, result in five significant ones, just by chance.

Multiple testing can be a problem with epidemiological studies, where researchers compare a group of healthy people with another group of people with a certain disease. They often check dozens, if not hundreds of different life style factors for significant associations with the disease, and only report the ones that show a significant association but don't correct for the fact that they rolled the dice that many times. This is part of the reason why one day, we are told that coffee increases our cancer risk, and the next day it doesn't.[2]

Often, researchers aren't even aware that they are checking their data many times, says Donald Berry, a biostatistician at MD Anderson Cancer Center in Houston. In an article on this problem,[3] he explains it this way: Say you're traveling in a foreign land, and "pass through the town of Oz. The inhabitants seem unusually tall. You retrace your route to ask the heights of the people you had seen. The 25 adults queried averaged 4 inches taller than the mean height of their compatriots, after adjusting for sex and age. Accounting for sampling variability you [...] find the observation to be highly statistically significant (P < .001)."

However, the finding is most likely a false positive, Berry says. That's because you probably noticed many unusual things in the inhabitants of Oz in addition to height—like the color of people's clothes, whether they wore glasses, had freckles etc.—without being aware that you were checking them for all those features. In a similar way, Berry says, it's a problem if researchers do an experiment and get something that seems really unusual, then redo it, but only report what they believe is the better reading, not necessarily the first reading. "They are not taught to report everything," he says. "If you are going to redo it because it looked unusual to you, you still have to give the first reading. They don't do that."

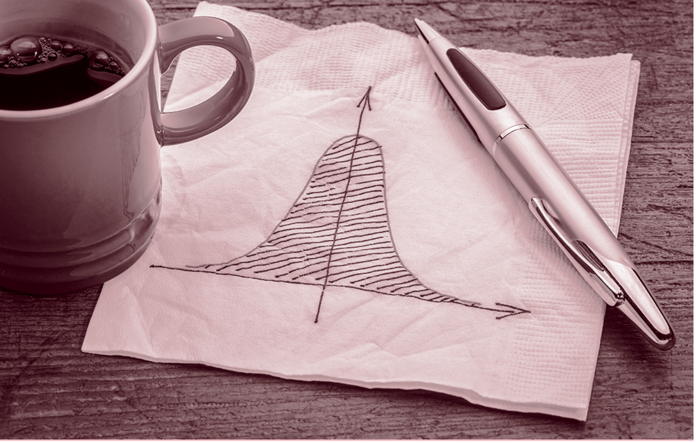
> “It’s a problem if researchers do an experiment and get something that seems really unusual, then redo it, but only report what they believe is the better reading, not necessarily the first reading.”
>
> **Donald Berry**
> **MD Anderson Cancer Center**

A related problem is when researchers keep adding experimental samples until they get a significant result, Silberberg says. They believe that they eventually got statistical significance because they had included enough samples to detect a true effect. Not so, he says: What they really did was increase the likelihood of getting a false positive result each time they analyzed the data after adding more samples. Uri Simonsohn, a behavioral scientist at the University of Pennsylvania, showed that—especially with small data sets—chances are that adding more samples might initially render results statistically significant but then insignificant again, even though the sample size keeps increasing.[4]

In the cases mentioned so far, researchers sometimes roll the dice without being aware of it. But in systems biology studies, analyzing the same data set many times is a clear feature of the study, and the multiple testing problem therefore much more obvious. Examples are microarray analyses that compare the expression of thousands of genes in mice treated with a drug and untreated mice; or a genome-wide association study that checks which among thousands of genetic markers are significantly more common in people with a disease than in people without it.

So how can you address the multiple testing problem in your research?

1. First, you'll need to do what researchers conducting randomized clinical trials are already required to do: Keep track of how often you actually look at the data. To do so, you'll need to decide on the experimental design and when to start and stop data collection before the experiment and then stick to that plan all the way through. You'll eventually also need to report, in your paper, the experimental plan and whether and why you deviated from it. This also means that you can't just "clean up" your data after you're done with the experiment, for example by removing outliers, because that, too, would mean deviating from your initial research plan as to which data to collect and include. One web site that provides help and guidance with the preregistration of your experimental plan is the Open Science Framework,[5] run by the non-profit Center for Open Science.[6]
2. Once it's clear how often you roll the dice, you'll need to get together with a statistician to check if and how you should statistically adjust for that, for example by using more stringent p values.
3. Another way to minimize false positive results in systems biology experiments that analyze a lot of data is independent replication of the experiment. Say you're checking which of 10,000 genes significantly change their expression level in a group of mice with a disease compared with healthy mice. With a p value of 0.05 for significance, you'd likely get 500 (5%) false positives just by chance. Most should disappear if you repeat the experiment with a new set of mice.



References:

1 *J. Clin. Epidemiol.* **59**, 964 (2006)

2 “Numbers can lie,” *Los Angeles Times* (9-17-2007)

3 *J. Natl. Cancer Inst.* **104**, 1124 (2012)

4 *Psychol. Sci.* **22**, 1359 (2011)

5 http://osf.io

6 http://cos.io

Further reading:

Landis SC et al.: A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* **490**, 187 (2012)

Kimmelman J & Anderson JA: Should preclinical studies be registered? *Nat. Biotechnol.* **30**, 488 (2012)

LEARN TO THINK BAYESIAN

Researchers are usually most familiar with the traditional, frequentist approach to statistics and one of its most commonly used statistical tests, which comes up with a p value—the probability that an experimental result or something more extreme could have occurred just by chance. **Usually, results are considered statistically significant if this probability is less than 5%.**

But many problems with reproducibility, like the multiple testing issue, or the fact that many negative results are never published, can be spotted more easily with the Bayesian approach to statistics. That's because this approach doesn't just look at the experimental results themselves, but also takes into account any other information—including information before the experiment took place—that might be relevant to the outcome.

To explain this approach, Hall gives the following example: Suppose you get a positive test result for some terrible rare disease you've heard about. The doctor tells you the test is very reliable: The false positive rate is 0.01%—the test will mistakenly diagnose only 1 of 10,000 healthy people as having the disease. That will likely cause you to think that you probably have the disease.

But in Bayesian statistics, there is another factor you need to consider: Just how likely is it that you get the disease in the first place? This is what Bayesian statisticians call the prior probability that you have the disease. In this example, that's the prevalence of the disease.

Assume the prevalence in the US is 1/100 million: out of 100 million people, one has the disease. Now because the test you took gets 1 in 10,000 results wrong, and because 100 million people contain 10,000 of such wrong readings if all 100 million people took the test, 10,000 healthy people of the 100 million will get a false positive test result, in addition to the single person who really has the disease.

Once you know that, you'll probably relax: Your positive test result only means that the chance that you have the disease is actually only 1:10,001. This is what Bayesians call posterior probability—the probability of the result you are interested in, after considering all available information including the prior probability.

This so-called base rate fallacy shows how the Bayesian approach takes prior probabilities into account. To explain how this matters in research, Hall gives another example: A company does a randomized controlled trial of a cancer drug. In the traditional, frequentist approach, the trial is taken to support the claim that the cancer drug works if fewer drug takers get cancer than placebo takers—at a p value of less than 0.05. This means that the difference is statistically significant: There is a less than 5% probability that you'd have seen the same or a bigger effect just by chance.

Now assume the company does 100 such trials. With a p value of 0.05 for statistical significance, 5 of them (5%) will show that the drug worked—just by chance. If the company only published the 5, everyone would be impressed and say, "'oh wow, a great new drug to treat cancer,'" Hall says.

Bayesians are less likely to make that mistake, because they consider all available information. "A Bayesian," Hall says, "would publish all of the data from all 100 studies. But if you just look at the studies through a frequentist lens, it's very easy to miss this. Very basic stuff like making sure that so-called negative results get published, which is just an obvious thing to do if you are Bayesian, is still [often] not done, [whereas] in physics, Bayesian techniques are standard by now."

"Bayesian thinking helps you to be aware of many experimental design problems that contribute to reproducibility issues."

Because the Bayesian approach involves looking at all observations including negative ones, it should also help avoid the multiple testing problem, which is, after all, the result of ignoring negative data. What's more, it forces you to check previously available evidence before starting an experiment. This is often not done in biomedical research, which is why there are unnecessary clinical trials—like one that enrolled 7,000 stroke patients to test if the calcium channel blocker nimodipine was effective to treat them, even though preclinical studies were available before the trial started that showed no effect.[1]

It's not always easy to do the computations in Bayesian statistics, and estimating prior probabilities can often be subjective. But Bayesian thinking helps you to be aware of many experimental design problems that contribute to reproducibility issues. So you may want to take a closer look.

References:

1 http://www.bwfund.org/newsroom/newsletter-articles/special-report-biomedical-research-are-all-results-correct

Further reading:

Tversky A & Kahneman D: Evidential impact of base rates (on the base rate fallacy), pages 153-160 in: Daniel Kahneman, Paul Slovic, Amos Tversky (editors): Judgment under uncertainty: Heuristics and biases. Cambridge University Press (1982)

Ian Hacking: An Introduction to Probability and Inductive Logic. Cambridge University Press (2001)

Donald A. Berry: Statistics: A Bayesian Perspective. Duxbury Press (1995)

Probabilistic Programming and Bayesian Methods for Hackers. https://camdavidsonpilon.github.io/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers

USE STANDARDS

We already discussed multiple testing as a major statistical challenge of high throughput experiments. But there are **many other technical issues** that can make such experiments difficult to reproduce. Some of them **can be addressed by the use of standards.**

For example, to measure the mRNA expression levels of thousands of genes on microarray chips, you need to isolate RNAs from cells, attach fluorescent dyes to them and then hybridize the labeled RNAs with thousands of complementary DNAs that are on a microarray chip. Often you'll use two mRNAs from an experimental sample and a control, each of which you label with a different fluorescent dye.

The problem is that ozone can degrade one of the two commonly used dyes more than the other. As a result, microarray results on a smoggy day will differ from results on a not-so-smoggy day, says Ronald N. Germain at the National Institute of Allergy and Infectious Diseases (NIAID). Ozone scavengers used in many labs only partially solve the issue, he adds.

This is why it's important to make a standard RNA batch and include it in every experiment; ideally, labs should also share their standards with other labs to improve comparability of results between labs. But often researchers still aren't convinced this is necessary, says Germain, who has been developing such standards. "I have found enormous resistance from people," he says.

As a result, it's often difficult to compare microarray data from published studies because standards are often missing, Germain says, adding that NIH institutes like the NIAID are discussing ways to require standardizations at least whenever they fund collaborations between different labs.

There are also efforts to develop standards with other high throughput technologies. In the case of RNA-seq, which is slowly replacing microarray analysis and involves sequencing and counting all mRNAs of a cell directly, an "External RNA Control Consortium" has developed so-called spike-in RNA standards that contain predefined numbers of transcripts.

Standards are also becoming available for flow cytometry, where a laser helps counting or sorting cells by analyzing certain cell surface markers, typically using fluorescently labeled antibodies to these markers. For the major cell markers, Germain says, researchers have come up with a standard antibody panel researchers can use in their flow cytometry assays.[1]

“It’s often difficult to compare microarray data from published studies because standards are often missing.”

Ronald N. Germain
National Institute of Allergy and Infectious Diseases

But as the case studies in the next chapter show, not all reproducibility problems with high throughput experiments can be solved by using standards. The good news is that in principle, it’s possible to eventually solve most problems. Doing so, however, can be difficult and time consuming and often requires collaboration between the different labs whose results can’t be reconciled.

References:

1 *Nat. Rev. Immunol.* **12**, 191 (2012)

CASE STUDIES: WHAT TO DO WHEN YOU CAN'T REPRODUCE SOMEONE ELSE'S DATA

Case Study 1: FACS

It was a research collaboration like many others: The labs of Kornelia Polyak at the Dana-Farber Cancer Institute in Boston and Mina Bissell at Lawrence Berkeley National Laboratory in Berkeley were using a variation of flow cytometry, fluorescence activated cell sorting (FACS), to analyze healthy breast tissue cells for two markers, CD10 and CD44 (markers of two types of cells found in the mammary gland epithelium). But even though they were using the same source tissue, there was a consistent difference in the results from both labs: Bissell's lab found all cells positive for both markers, while in Polyak's lab at least some cells only expressed one.

What was going on? To find the culprit, they tried using the same FACS machine, the same antibodies, the same data processing software, and went through each step of the protocol to make sure both labs went through exactly the same steps. Still, the differences persisted. "[It] was more than a year and it was very frustrating," Polyak says. "We didn't understand: Why can't we follow a protocol?"

Finally, when Polyak's postdoc went to Bissell's lab to perform the experiment side by side with Bissell's postdoc, they found that the outcome had to do with the enzymatic digestion step that breaks the tissue into single cells.[1] And it wasn't the batch of enzyme used; it was the way they stirred the cells. While Polyak's postdoc used a magnetic stir bar for eight hours, Bissell's used a more gentle shaker for 12 hours. This made sense: The faster stirring in Polyak's lab might have disrupted the markers in some of the cells, explaining why not all of the cells expressed both of the markers. While Bissell's lab was doing "some very gentle shaking," Polyak says, "we [had] a stir bar—and that [seemed] to be perturbing some of the antigens on the surface."

The case, Polyak says, taught her that sometimes even the most detailed protocol might not contain the very detail that makes the difference: "Who thinks [the] speed of stirring tissue [or] shaking would make a difference? But it does!" That's why Polyak has started to videotape some of the procedures done in her lab. With a video in hand, they would have found out much earlier, she says. What's more, she says, "you really have to contact people if you have a disagreement. If you don't, it's almost not possible to figure out what's wrong or why you cannot reproduce [something]."

So when she recently couldn't reproduce a FACS profile of leukocytes published by the lab of Lisa Coussens at Oregon Health & Science University, she didn't hesitate to contact the other lab. "It turned out we had to change our FACS machine and the filters," Polyak says, adding that again, it was also crucial to not digest the tissue for too long to prevent disruption of the surface markers. "We could not have done that if we had not talked directly to the other lab. And even then, it took us six months."

> "Who thinks [the] speed of stirring tissue [or] shaking would make a difference? But it does!"

Kornelia Polyak
Dana-Farber Cancer Institute

Case Study 2: Proteomics

When proteomics researchers want to find out which proteins interact with each other, they allow them to bind together, then enzymatically cut the protein clump into many pieces, electrically charge the pieces, smash them against a gas cloud to generate even smaller fragments, and measure the masses and charges of the resulting peptide pieces in a mass spectrometer (MS). Finally, a computer program cobbles together all the pieces into the original peptides. It's as if "you were to take a wine glass, throw it against a wall, [and] glue the pieces together," says Ruedi Aebersold, a proteomics researcher at ETH Zürich.

Given the complexity of the process, it's perhaps not surprising that it is sometimes poorly reproducible: When a group of researchers mixed equal amounts of 20 proteins and sent the mix to 27 labs, only seven labs got all proteins right.[2] "The results were stunning," says Aebersold. But when Aebersold analyzed all of the data the same way in a single place—in his old lab in Seattle—the discrepancies mostly disappeared. This, Aebersold says, shows that "by far the largest contribution to the variability was the data analysis software."

In another version of the procedure called affinity purification MS, researchers let cells express a protein with a handle on it, lyse the cells and then use that handle to isolate only that protein and the other proteins bound to it; then, they analyze which proteins are involved by MS. When Aebersold found that only 30% of his results overlapped with the results from another lab led by Giulio Superti-Furga in Vienna, the two labs decided to collaborate to find out why.

“Repeat an analysis, even if that’s expensive... you learn a lot about the quality of the data.”

Ruedi Aebersold
ETH Zürich

Again, it turned out that the data analysis was the culprit:[3; 4] Among other things, one lab’s analysis software allowed for more false positive (and fewer false negative) data points than the other.

One lesson from the experience, Aebersold says, is that “it’s important that people document exactly what [statistical] filters they use, at what level they filter [and] with what tools they filter.” This, he adds, applies to all high-throughput techniques including microarrays, sequencing and metabolomics. “People [should] describe in their papers what tool they used with what parameters and make the raw data accessible.”

What’s more, he says: Repeat an analysis, even if that’s expensive. “People say, ‘well, I can’t afford to do duplicate or triplicate analyses because it costs money and I don’t learn anything new.’ And that’s a notion I would strongly argue against. Yes, you will not learn anything new in the sense that you identify more genes or proteins, but you learn a lot about the quality of the data.”

References:

1 *Cell Rep.* **6**, 779 (2014)

2 *Nat. Methods* **6**, 423 (2009)

3 *Nat. Methods* **10**, 307 (2013)

4 *Nat. Methods* **10**, 301 (2013)

Further information:

NIH video module 1 “Lack of Transparency” discusses contacting another lab to reconcile discrepancies (http://www.nih.gov/science/reproducibility/training.htm)

JUST AS IMPORTANT: NOTETAKING AND REPORTING

When you have to ask one of your former postdocs to **get on a plane** to explain how they did an experiment to the person continuing their project, you know you might have a problem with notetaking.

Deatherage still remembers a case where this actually happened: A postdoc in a cell biology lab couldn't reproduce the results of another postdoc who had left the lab. When they flew in the previous postdoc to check what he had done differently, they found that he had been using a different assay kit to measure protein levels: While the new postdoc was using an assay called Lowry (which is more sensitive to tryptophanes and tyrosines), the previous postdoc had been using a Bradford assay (which is more sensitive to arginine residues). However, he had stopped mentioning this detail in his lab notes because he ran these assays so often that he stopped writing down which assay he was using, Deatherage says.

King had a similar experience in his lab: When two of his rotation students tried to dissolve a compound, it kept precipitating. What they didn't realize, King says, was that they had to add a small volume of the compound first, before adding the medium, as opposed to doing it the other way around. The reason they didn't know was that the postdoc who taught them the protocol had forgotten to include that detail.

The lesson here is clear: Even recording seemingly unimportant experimental details can be crucial, such as the number of passages a cell line went through, the exact source and clone ID of an antibody, or where you got a mouse strain from. Germain still remembers that when he was a grad student, he even used to record the batch of certain reagents like the fetal calf serum he used. "We can't always anticipate all of the information that we, in retrospect, should have collected," he says. "But it's surely useful to [think] hard about what all the variables can be."

It's also important to record the experimental results in their original form instead of a sanitized version, Fang says, even the "ugly western blots with the big stains on them and the lanes in the wrong order." But that's sometimes easier said than done, in part because different lab members use a different system to record their experiments. "Standardizing record keeping and data management is a huge lack right now," says Micoli. One possible solution: electronic notebooks.[1] "You cannot erase them, they are always there," Casadevall says.

Recording exactly what you did is especially challenging when using software. Polyak says she started to record videos of the exact steps people in her lab go through when they process RNA-seq or ChIP-seq data.

There are even tracking programs that enable users to record the exact keystrokes and commands they used when using software, Germain says. This allows others to later reanalyze a deposited dataset exactly the same way the original authors did.

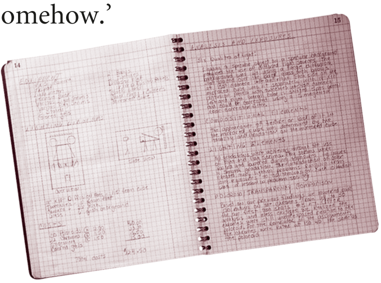
One tool that's increasingly used this way is the IPython notebook.[2] They record every step when users enter and execute code, and even let them include explanatory text, plots and other media—all in a single document. Readers can later download a notebook and use it as a basis for their own data analyses.

“Standardizing record keeping and data management is a huge lack right now.”

Keith Micoli
New York University School of Medicine

Another tool is GenePattern. Developed by the Broad Institute as a “platform for reproducible bioinformatics,” it allows users to capture and later share all of the steps they take when analyzing genomic data.[3] It remembers all parameters and software versions used for an analysis, which later enables others to reproduce it. Recently, it’s even become possible to run GenePattern from Python.[4]

Once you’ve captured all of the experimental details, you’ll still need to report them in your papers. Often researchers don’t do so, Bremer says. For example, for RNA-seq, there are many different normalization procedures to correct biases in the data towards certain types of sequences, but “often people do not even say how they normalized,” she says. “They say ‘we normalized the data somehow.’ Sometimes they don’t even mention which [software] program they analyzed the data with. That’s a big problem.”



So make sure you include enough information to enable others to reproduce the experiments in your paper. Here is a recap of the things mentioned before that you should report in the papers you publish:

- whether you checked cell lines for mycoplasma contamination;
- where you got your cell line(s) from, whether you authenticated them, and how many passages the cell lines you used went through;
- the sample size or number of independent replicates in your experiments and how you determined it;
- the experimental plan and design you made before starting your experiments and whether and why you deviated from it, including the results of all planned measurements, even the ones with a negative outcome. If you use cell lines, this includes, for example, the entire range of results you got with different serum batches used to grow the cell lines in;
- whether and how you did blinding and randomization, and if you didn't, state the reason why not.

> “The goal isn’t to get the most papers in *Science, Cell* and *Nature* or to win prizes. It’s to produce something that really advances society’s understanding of something.”
>
> **Ferric Fang**
> **University of Washington**

Whenever you’re wondering whether maximizing the reproducibility of your results is a waste of time that might keep you from publishing your results as quickly as possible, remember why you got into science in the first place, says Fang: “I understand that people have career anxieties and they are greater today than ever,” he says. “But the reason that you get into science in the first place is you feel that it’s a way to serve society. The goal isn’t to get the most papers in *Science, Cell* and *Nature* or to win prizes. It’s to produce something that really advances society’s understanding of something. And that is something that any scientist can really look back and be proud of—that they added something to this edifice of scientific understanding that someday is going to be useful to other people. That’s the satisfaction that comes from science and that can only come from doing your work really carefully.”

References:

1 *Nature* **481**, 430 (2012)

2 http://ipython.org/notebook.html

3 http://www.broadinstitute.org/cancer/software/genepattern

4 http://www.broadinstitute.org/cancer/software/genepattern programmers-guide#_Using_GenePattern_from_Python

Burroughs Wellcome Fund Career Development Guide Series

Advancing Careers: Articles from the Focus Newsletter
Advancing Careers is a collection of articles that were published in the Burroughs Wellcome Fund's FOCUS newsletter. Topics include:

- Considerations in Accepting a Faculty Position
- Managing Your Laboratory
- Communicating and Funding Your Science
- Balancing Work with the Rest of Your Life

Communicating Science: Giving Talks
Practical tips on presenting your work in a variety of circumstances—from the formal to the informal.

Intellectual Property: An Overview
Provides an overview on patents and copyrights, the biggest players in a broad classification of intellectual property and the lynchpins behind innovation and commercialization of biological inventions.

Moving On: Managing Career Transitions
Moving on is never easy and neither is recognizing it's time to do so. This guide is meant to help scientists gain some control over a process that can seem subjective and prone to idiosyncracies.

Staffing the Lab: Perspectives from Both Sides of the Bench
Are you looking to hire the perfect postdoc? Are you looking to be hired? This guide takes a look from both perspectives on creating a productive work environment.

Thriving in an Era of Team Science
How can you build a career in science when much of your work occurs in the context of team efforts? This book provides tips and advice on how to survive and thrive in collaborative science.

Working with Institutional Review Boards
This guide provides a general introduction and insight from experts on what an Institutional Review Board does and understanding its importance.

Email news@bwfund.org to order your free copies.

Acknowledgments

Series Developer: Victoria McGovern, Ph.D.
Author: Andreas von Bubnoff, Ph.D.
Editor: Russ Campbell
Designer: Liaison Design Group

Special thanks to Mindy McFeaters

BURROUGHS
WELLCOME
FUND

21 T. W. Alexander Drive
P. O. Box 13901
Research Triangle Park, NC 27709-3901
info@bwfund.org
www.bwfund.org