Reproducibility Podcast Episode 1: What is the Reproducibility Crisis?

Aaron Carroll:

Welcome back to the Healthcare Triage Podcast. Thanks to support from the National Institutes of Health, we've created a series on a special topic, science culture and reproducibility. Most people agree that there's a major problem with reproducibility of scientific studies. In fact, it's been estimated that as much as half of scientific studies are producing results that are false. We want to know what about the culture of science is contributing to the problem and how we can fix it.

We've created this eight-episode series to address all that. We talk to funders, journalists, and scientists from various backgrounds and at various career stages to try and outline what the issues are. We spend an episode digging into the incentives built into this system of academia and the pressure on scientists to produce big splashy, positive results. We touch on the often troubled ways in which we use statistics.

We zero in on grant writing and funding practices and the ways in which universities depend on their faculty getting grants. We pick apart journals, peer review, and publishing practices and expectations as a whole, highlighting our problematic obsession with bibliometrics. We examine the role the media has to play in science culture and in holding science accountable. And we ask questions about the way we mentor young scientists, including the pressures they face and the career options for which they're trained.

And then after all that, we ask our experts, "What can we do about it?" I'm Aaron Carroll, a pediatrician and health services researcher at the Indiana University School of Medicine.

Tiffany Doherty:

And I'm Tiffany Doherty, a neuroscientist and science communicator also at the Indiana University School of Medicine. And we'll be your host throughout this series.

Aaron Carroll:

To begin. We need to talk about what the problem is and if everyone thinks it really is a problem. So today in the first of these episodes, we talk to people about the replication crisis. What is it? And when, and how do we come to recognize it? And does everyone recognize it or recognize it as the same thing? Let's kick things off by asking people, is there a replication crisis?

Speaker 3:

Yes, and I have no problem calling it a crisis.

Brian Nosek:

The persistent finding is that it's not as credible, the published literature as we, we might think it is, or as we might hope it would be.

John Yewdell:

But reproducibility per se, I and many of my colleagues don't perceive as an issue.

Speaker 6:

Yes, I do. And I think that for several reasons,

Speaker 7:

I don't like the word crisis. I'm happy to use the word crisis when it's appropriate. I think when people started talking about the reproducibility crisis, it put a lot of people on a defensive. And it also presumes that this was a new problem, when in fact we just don't know. And, and now we're starting to see some evidence of issues in the past.

Speaker 8:

I think people like to call it a reproducibility crisis, but I think crisis is a very, very strong word.

Stephanie Lee:

Scientists have been discovering problems that have been there for a long time and now are developing tools and a vocabulary and a consciousness about how to address those problems.

Aaron Carroll:

So there are a range of opinions, but the majority of experts we interviewed see a problem. So what brought that problem to our attention? It turns out there's been some awareness of it since the 1980s, which we'll talk about. But a few major events seem to have brought it into the limelight more recently, so much so that it's caught the attention of major journalists, including Stephanie Lee of BuzzFeed News and Ed Yong of the Atlantic, both of whom agreed to participate in this podcast.

Essentially, we got a sense that things started coming to head in the last 10 to 15 years. Here's Stephanie.

Stephanie Lee:

Yeah. I don't know if I can time stamp it. I would say that awareness of this, to me, it seems to have really started in the mid-2000s when John Ioannidis wrote his famous paper, Why Most Published Findings Are False.

Aaron Carroll:

Just a quick note here about this paper, it was published in PLOS Medicine in 2005, and with the use of simulations reported that for most study designs and settings, the majority of research claims are more likely to be false than true. You'll probably hear our interviewees reference it once or twice throughout this series.

Stephanie Lee:

And then in the early 2010s, there were just a series of very high profile retractions and other issues that I think galvanized the social science and psychology fields in particular into looking at this. As well as some other fields like cancer biology.

Tiffany Doherty:

In a couple of interviews, we dug deeper into specific events that garnered major attention. One of them is a famous case of research misconduct, complete with lots of press, paper retractions and ultimately loss of a career. The other is a researcher who is digging into one major problem discovered over and over again throughout the literature, so much so that her career is so be dedicated to this problem now.

But let's start with that first event, the infamous case of Brian Wansink. We sat down with Dr. James Heathers, a major voice in the meta science field to talk about his work in general, and then more specifically his work on the Wansink case.

James Heathers:

Good morning, good afternoon, good evening, science people of the world. My name is James Heathers. I live in Boston, Massachusetts, United States of America. I am a funny combination of two jobs. I'm a physiologist and I'm a meta scientist. The second one's always what people ask about, so why don't we explain that. A meta scientist is someone who studies science using scientific techniques. And my little domain of that broader topic is something called error detection, which is the, I suppose, social and mathematical pursuit of mistakes in published scientific literature, how do we define them? How do we fix them?

Aaron Carroll:

So can you tell me, give me some good examples of case that have come across your field division and where you've looked into it and seen stuff that's not right?

James Heathers:

The famous one everyone's familiar with now is the Brian Wansink case.

Aaron Carroll:

I'm familiar with this and I'm horrified before you, but you should tell the story.

James Heathers:

Right. He wrote a blog post on his own blog that was essentially, "Here is the academic guide getting ahead." And the advice that was given was basically, it was a combination of nevermind that pesky work-life balance stuff. And why not take account a big wrench and hit a dataset until something falls off the side of it? Pretend you planned it and then go and publish that wherever the hell you'd like. So it was to me, at least, and to a lot of other people whose opinions I would take seriously, it seemed to be an amazing example of either something that was desperately cynical or powerfully ignorant.

Aaron Carroll:

Let's just stop here and explicitly say that this blog post specifically described p-hacking, which we'll talk about a bit later as the way to get ahead in science as opposed to an unquestionably inappropriate way to analyze data. It's basically taking a dataset that did not provide any significant results, then running different analyses until a positive result spits out that you can then try and publish. Handling data in this manner is highly likely to provide spirituous results.

James Heathers:

Some people thought it was Satire. Yeah, it was about a visiting graduate student that he had. He was so determined to be successful, that she plowed through all of these datasets that had been previously rejected by someone for not having any results consistent with their original analysis plan. Otherwise, no one has a negative result, Which in itself should not be a problem if your experiment is well constructed, which it wasn't, but let's not go down the rabbit hole. We had the very first observation of what would later be kind of a motto for me with this.

Everyone knows, "Where there's smoke, there's fire." But in cases like this, because of the rather high threshold, you have to cross to make errors noticeable, I'm more on team where there's smoke, there's an arsonist. Something must be obvious. So I didn't do any of the initial work here. Colleagues of mine went and looked at the papers corresponding to what was described as great work in this blog post and they were blown up. There were just shovelful of figures that couldn't exist. There were internally contradicting parts of different analyses that couldn't possibly fit together.

It was shambles. It was obviously... They didn't even p-hack it right. It was so very bad that obviously done it one section at a time, how do we answer this particular question? And they'd done someone else analyses and excluded some data and hit it with a stick and turned it upside down and put it in a blender. And the resulting things that came out were both inaccurate and didn't agree with each other. And they wrote a paper that was just on the potential inaccuracy of that set of papers.

And there was some interest in that, and we all started to get more and more and more interested in the broader problems. And then started systematically working through the back catalog of these papers, because they were really alarmingly problematic. We went back as far as 1993-

Aaron Carroll:

Hey. Eish.

James Heathers:

... and when we finished, there were 50 annotated papers. Now, that was the driving force behind everything that happened. That was a phenomenal amount of work done in the middle of the night by all of us.

Tiffany Doherty:

We should mention here that a large majority of this work did not result in papers for Dr. Heathers and his colleagues or in anything that would contribute toward advancing their scientific careers. The amount of work you have to do to back up the very serious claim that someone has manipulated their data in this way is serious. These guys all spent an incredible amount of time and effort to uncover this at very little tangible benefit to themselves.

James Heathers:

The whole thing broke wide open was the correction, retraction, correction, retraction over and over again.

Aaron Carroll:

And we should be clear he had achieved the phenomenal level of success.

James Heathers:

Oh yeah.

Aaron Carroll:

Both in the general public as well as in the literature.

James Heathers:

He was extremely popular

Aaron Carroll:

Ted talks, books.

James Heathers:

Popular press books, yep.

Aaron Carroll:

Huge food lab. It's horrifying because of course everybody's like, "Well that has to be a one off story. But then we can keep hearing stories like this again and again.

Tiffany Doherty:

We'll be hearing from James a lot more throughout this entire podcast, but we want to switch now to Stephanie Lee, the science reporter at BuzzFeed News, who you heard toward the beginning of this episode and who broke the stories on Wansink.

Aaron Carroll:

This took off where the general public started to get it, when you started talking about all the issues with Brian Wansink. How did that start?

Stephanie Lee:

Well, it started in November 2016. That's not when I started covering it, but in November 2016, Brian Wansink who was the professor at Cornell, we should say. He ran this very high profile group called the Cornell Food and Brand Lab. And for year they had been studying all kinds of things related to food, marketing, consumer behavior, eating, dieting, questions that a lot of people are interested in like does the size of your bowl or plate affect how much you serve yourself? Does the shape of your wine glass affect how much you pour? Does leaving snacks on your kitchen countertop affect how much you weigh? Is the lighting and music in a restaurant, does that have a link with how much you eat?

And so his group was studying these questions about diet, nutrition and psychology for years and years. Because his work was so... It seemed to speak to what so many people wanted to actually know and he was on all the major talk shows, he got tons of media coverage. He had lots of grants from the federal government. He was high up at one point in the USDA and oversaw dietary guidelines for Americans.

Aaron Carroll:

Yeah.

Stephanie Lee:

So he was an authority on healthy eating and his message boiled down to ways that you could nudge yourself into healthy eating short of doing rigorous exercise. But there were easy ways that you could lose weight and eat better without even really consciously trying, if you just adjusted your environment.

Tiffany Doherty:

Here, Stephanie brings up the infamous blog post.

Stephanie Lee:

In November 2016, he wrote a blog post about how he had had this post doc or grad student. And he had given her a dataset from a previously conducted experiment where that involved pizza, oh, you can eat buffets and studying the behavior of diners. They got four papers out of that dataset. The pizza papers as they, I guess, are now infamously known. He put this up because he clearly didn't see anything wrong with it. He was proud that this student had put in the work and found some interesting results. But to the rest of the scientific community, this was alarm bells.

And so I jumped into the fray that summer when I was just reading on Retraction Watch about all these papers continuing to get corrected or are retracted. And I was like, "This story's not over." Because at that point, the lab was not doing any media interviews and they'd gone from being very media-friendly to not really talking to the press very much at all. I was curious about what they were saying to each other behind the scenes. What is the conversation in the lab that's going through that kind of scrutiny.

I filed several public records requests to the lab and it's collaborators that I discovered there were so many other papers that were problematic and they were scrambling to try to find the data and address these issues. But they seemed to have been caught totally off guard. And then things just escalated from there.

Aaron Carroll:

When you find this kind of thing and you've pointed out. You didn't point it out and then Cornell was like, "Oh my God, you're right." And then Brian went and changed what he did, and the journals took back the papers and the world was improved. I imagine there was resistance to all of this and probably still is. Am I wrong?

Stephanie Lee:

No, that's a absolutely correct. So I started writing about the lab in August 2017. And at that point, Nick and James and that crew had been for months emailing journals, putting up their work online on Twitter and blog posts. And since its initial investigation, finding Cornell was just all but silent and they seemed to have hit a wall. I guess you would say the normal channels of trying to raise an issue about a scientific piece of work seemed to have broken down. So I started writing about the lab that late summer fall, and one story just led to the next.

That fall, after a couple months of, of writing about this Cornell said, "We are going to do a full investigation." Eventually the following spring, they found that he had committed academic misconduct and then he announced that he would retire.

Aaron Carroll:

And again, this would never pass in a court of law, but I'm asking you like, do you think he thinks he did anything wrong?

Stephanie Lee:

He has said that he does not think so.

Aaron Carroll:

And I think we see that from the beginning, like when he did that initial blog post I remember being like, oh my God, "He's saying the quiet part out loud."

Stephanie Lee:

When I was publishing correspondence between him and his collaborators, a very common reaction I heard was, "Wow, this is really extreme, really blatant, like really unacceptable. It's also an extreme version of something that goes on in a lot of labs. You run an experiment and maybe the results aren't quite what you wanted and that makes you rethink, "Oh, should we have looked at these variables this way?" And to be clear, exploratory research can and should happen. That is a huge part of science.

The huge flaw though was them not making that clear. And instead presenting it as like, "We always intended to study this." We ended up getting exactly what we were looking for.

Tiffany Doherty:

We want to stop here and highlight something Stephanie said. The Wansink case is an extreme version of something that goes on in a lot of labs. That story gets a lot of attention because it was blatant and it was big in so many ways. This level of manipulation probably isn't happening in most labs, but smaller forms of it likely are, and that's still a big problem. As Stephanie just said, exploratory research can and should happen, but how many people are actually p-hacking while calling it exploratory?

Stephanie Lee:

So the peer viewers didn't catch things. The media kept covering the studies and didn't see problems with them. The institution apparently loved it because they put out press releases and gave them a lab of a lot of staff. Funders enjoyed the results in part because they were so practical. A lot of his work ended up inspiring the smarter lunchroom movement, which was this nationwide program about how to design your cafeterias so kids will want to eat more vegetables and more fruits?

Aaron Carroll:

So beyond making it difficult for researchers to build on existing work, not addressing issues of reproducibility can cost the public as well in terms of new programs and policies built on unreliable data. So obviously there were major consequences from Wansink's inappropriate research behaviors, but it took a lot of work to bring them to light. And it may surprise you to know that getting mistakes and/or fraud appropriately addressed can actually be really difficult.

More often than not there are no consequences for bad research behavior. We sat down with Dr. Elisabeth Bik, who is very well known for her work on image duplication in papers, and we were shocked by a lot of things, including the response to the duplicated work she is dedicated to uncovering.

Elisabeth Bik:

My name is Elizabeth Bik. I did my PhD in microbiology in the Netherlands and in 2001, I moved to the U.S. and worked at Stanford for 15 years, doing microbiome research. And in around 2013, I got very much interested in science integrity work, mainly plagiarism at that time and I started doing that as a hobby. That turned into a full-time job about a year ago, I quit my up, I was working in industry back then. And I am now a full-time science integrity consultant, which means I am available to hire if you have a science integrity question.

But I'm mainly searching the literature for duplicated images, which I enjoy doing. And I am looking for images that are duplicated as a sign of science misconduct.

Aaron Carroll:

So let's talk about some of the work that you do on image duplication. How did you get started in that?

Elisabeth Bik:

It's a long story, but... I was working on plagiarism and I was mainly checking papers for plagiarized image... Sorry for plagiarized text. And by chance I was flipping through a thesis that had a lot of duplicated text in the introduction, a lot of plagiarism. And as I was going through that thesis, it also had images west of western blots, which are protein blots, so they are photos and they're supposed to be unique. In scientific papers, you'll often have illustrations, photos that illustrates the findings of that paper.

And one particular type of image that I'm using a lot that I see a lot of duplications in are photos of western blot. It looks like a bunch black stripes on a white surface, but the thickness of these stripes is an indication for how much protein is in a particular sample. So the photo serves as sort of a data as proof of a particular statement. But in this particular thesis, I found a photo of a western blot that was reused to serve as proof of something else, of a completely different experiment.

It was the same blot, it was actually turned 180 degrees. It was rotated and it's suddenly served as a proof for a different experiment. So let's say in Figure 2A, it was a time series of something, and in Figure 3C, it was a concentration series of something else. And it was this same photo, but it was turned upside down. And so the bands were seemingly in a different order, but it was the exact same photo. And I could recognize that because it had a little smudge on it that was very characteristic.

And I thought, "Hey, I've seen that smudge before.: So you could imagine that maybe all protein blots, these western blots also look similar to each other. But in this case it was rotated 180 degrees, so that seemed a little bit less likely to be an honest error. And I then flipped more through that thesis and found other examples. There was another western blot, another one of those horizontal stripy things that I found to be duplicated in another chapter of that thesis. And that is when I realized I have apparently a very strange talent for remembering these blots and-

Aaron Carroll:

What else have you seen?

Elisabeth Bik:

So I mainly see western blots, but I also see overlapping microscopy images.

Aaron Carroll:

How many times has this happened?

Elisabeth Bik:

I had the same question. I decided to do a more systematic search, so I went online. I looked at a lot of scientific papers. I actually screened 20,000 papers in the evening hours in the weekend. So that cost me a couple of two years of my life. And I wanted to know how many of these papers contained duplicated images. And I found about 4% of papers that have photographic images. And that's a key thing, not all scientific papers have photos. But the ones that did have a certain type of photo, I found about 4% of these papers to contain duplicated photos and duplicated as in inappropriately duplicated, meaning that they were different experiments.

Aaron Carroll:

Let's say you find one of these. What happens next?

Elisabeth Bik:

In the beginning when I just started doing this, around 2014, I reported all of these to the journals. This is the professional way of doing it, so you write an email to the journal and you say, "Hey, in that paper, I found a duplicated image in Figure 2A and 3C. Can you check if you agree with that and maybe contact the authors?" So then the appropriate followup of that is that the journal editor should contact the authors, should ask for original blots or an explanation. And then they could either correct a paper or they could decide that maybe these two just looked alike, but were actually very different or they could even retract the paper.

But after reporting around 800 of these papers and fast forwarding five years later, so when 2020, these 800 papers that I reported in 2014 and 15 ,only and third of them have been either corrected or retracted. The rest of these papers have not been addressed at all, so the journals decided to either not contact the authors or maybe they did contact the authors, but they didn't get a reply, could be many, many possibilities here. But two thirds of these papers are still out there with their duplicated images, which means that these papers are probably not very trustworthy anymore, at least in my opinion.

Aaron Carroll:

No.

Elisabeth Bik:

But the journals did not act upon these findings. So now I'm taking things to Twitter and I sometimes post them on pop peer, which is a site where you can post comments on papers. So I'm now deliberately choosing a different route in how to report these papers, because I'm frustrated with the lack of response from institutes and journals.

Tiffany Doherty:

Dr. Bik can be found on Twitter under the handle MicrobiomDigest, where she has deservedly amassed quite a following. She even has a Patreon account where several people donate to support her efforts in this area.

Aaron Carroll:

I understand if they checked and the author said, "It's not a duplication." And they said, "What can we do?" They're swearing, there's no response. That doesn't seem possible.

Elisabeth Bik:

It is. I've written about big clusters of papers to institutes. And then the Institute will just say, "Oh..." At best they might respond because usually they don't respond anymore. If they respond, they might say, "Oh yeah, we did a little investigation and the author said these were all honest errors. And so we'll slap on a correction here, but nothing happened here." So institutes as well have a lack of response and tend, in my opinion, to sweep things under the rug.

Aaron Carroll:

When you say institutes, what do you mean? Are you talking about like academic universities or?

Elisabeth Bik:

Academic universities, yeah. They're mainly universities.

Tiffany Doherty:

And in at least one instance, Dr. BiK got a response from a university, but it wasn't to address the issues that she'd raised with a paper. It was to tell her that, "They, and I'm quoting, "had better things to do than to respond to the defamations of a failed scientist." Way out of a line. Dr. Bik does not generally contact the authors themselves and she tends to remove identifiers when posting specific work. But authors will sometimes respond when she posts their work. And in general, they deny that there's a duplication, which of course in this context just becomes one person's word against the others.

Aaron Carroll:

Do you find that the problems tend to cluster? Do they cluster among people? Do they cluster-

Elisabeth Bik:

Yes.

Aaron Carroll:

... among institutions or countries or?

Elisabeth Bik:

Not among institutions, although there's one exception to that rule. They do tend to cluster among certain labs. So if I now find a paper that I think has deeply images, I will check the other papers by that same research group. And in some cases I find other papers with similar problems. And the first authors, which are usually at least in my field of biology, the first authors are the junior students and the last authors are the senior students, so professors. In many of these cases, you actually see that it's associated to a lab, not to the first person.

So that might suggest that in a particular research group, there is a atmosphere in which people are either encouraged to do this or will fear for their job if they don't do misconduct,

Aaron Carroll:

This is something we address in a later episode on authority and mentorship in science. It turns out there's an amount of bullying in science that can't be ignored. It helps to propagate poor methods by intimidation of trainees. But anyway, back to Dr. Bik.

Elisabeth Bik:

I found one particular case where I found 100 papers from the same group. And I reported that I think a year ago and nothing has happened. And it's frustrating.

Aaron Carroll:

Wait, wait. 100 papers, which all showed that there were some kind of duplication of images?

Elisabeth Bik:

Yes.

Aaron Carroll:

And they did nothing?

Elisabeth Bik:

Yeah. I'll occasionally write in and they say, "We cannot say anything, but I think this person by now has retired." But the papers are all still out there.

Aaron Carroll:

Right.

Elisabeth Bik:

They're not corrected or retracted. There is no official statement. And it's frustrating for me because I see these things in minutes. And I understand as an academic institute, you need to do a careful investigation, but it's so obvious that if I can see it in minutes, it's hard for me to understand why these things take five years to investigate.

Aaron Carroll:

Do you think they're even doing the experiments? Are they doing it not getting the result they want and then just fudging at the end? Or are they just literally making it all up?

Elisabeth Bik:

It's probably at the end because there's many ways to fake results by doing it in a clever way in the lab. I could show that a particular protein is expressed less in a sample by just loading less of that sample. And I don't want to give any ideas, but I can fake results producing unique bends. You would never know that I faked experiments if I had done that that way, so.

Aaron Carroll:

What scares me is if 4% are willing to do this, there's got to be way more that are maybe doing it smarter.

Elisabeth Bik:

Yes. And that's what I fear. I've been fighting this fight for six, seven years and have not really been taken seriously. And I want to make people more aware that this is a problem by posting things on Twitter. But occasionally, I'm so mad about a paper or it's such a high impact paper, like a nature or a science paper from a well respected university. I'm so certain, this is a big problem, I want to tell this immediately to the world that this very important paper actually seems to have some image duplication or manipulation going on.

So I have posted a couple of those papers online and expected the journal to respond immediately. And in most of these cases, they do, they'll say, "Oh, thank you. We'll look into this." But it's

months and months later, and these papers have not even been corrected or retracted. They don't even have an expression of concern.

Aaron Carroll:

Did they not worry that someone would figure this out?

Elisabeth Bik:

Well, this is where I'm suspecting that they care more about the citations than people buying these papers than that they care about real science.

Aaron Carroll:

In a few minutes, we'll talk to someone from the NIH about what happens when fraud is reported to them. But first we want to take a look at work that takes an even wider look at research misconduct and reproducibility in general, and we wanted to talk to the people doing that work. Before we get there, though, we want to acknowledge that not everyone thinks that reproducibility is a widespread issue.

John Yewdell:

John Yewdell, I work at the National Institute of Allergy Infectious Diseases at the National Institutes of Health in Bethesda, Maryland. I am the principal investigator of a basic research laboratory. And we study the interaction of hosts, including people, I guess, in the end with viruses.

Aaron Carroll:

Dr. Yewdell was one of the people you heard in the very beginning of this episode.

John Yewdell:

Reproducibility per se, I and many of my colleagues don't perceive as an issue. I don't think the problems that affect the social sciences at all apply to the biological sciences that are experiment-driven. There are some issues that are similar, but to conflict it to, I think is a big mistake.

Tiffany Doherty:

I think we should be careful here about what is meant by a field being experiment-driven versus not. Social scientists are experiment-driven. He may mean that much stricter controls can be instituted from the bench, then can be instituted working with people, which is valid. But it's also important here to point out that replication issues plague several fields outside of the social sciences that are bench-based. People have pointed to issues of replication in fields like antibody research, cancer biology, et cetera.

Aaron Carroll:

When you say in your area, it's not an issue, are you talking about bench research or specifically experimental biology or both?

John Yewdell:

I can't speak to other fields like chemistry or physics, but in bench biology, what we do is complicated that the basic approach when you're studying anything is to take more than one approach. And so we're doing studies to address, I wouldn't say the same question, but the range of questions that are related to

each other. And we do multiple, multiple experiments over years and years typically to arrive at conclusions, which we know are conditional.

In any one experiment they do, we can come up with p-values, but what senior scientists learn is to trust trends from multiple experiments prefer over a fairly long period of time.

Aaron Carroll:

Do you think that that's the way that basic science or bench science in general works? Or do you think that that's your lab?

John Yewdell:

Well, I think that's the way it works in general.

Aaron Carroll:

I've got to be honest, you're one of the first people we've spoken to that is not totally pessimistic about all of this.

John Yewdell:

Reproducibility is a simple term for something that's really complicated.

Tiffany Doherty:

Dr. Yewdell explained here that we talk about reproducibility as binary when it's really not. Something may not reproduce for several reasons, but still be a solid finding. He explains one example where a paper from his lab reported that a very high fraction of newly made proteins are degraded, which is a really important finding. But following this publication, two postdocs tried to replicate these findings after the original postdoc had left the lab, but they couldn't get it to replicate.

Then a couple of years later, someone emailed Dr. Yewdell and basically said, "Hey, cool finding, but you didn't discover this." And they attached a paper published 20 years prior with basically the same results. So Dr. Yewdell's point here is like, if these were false findings, how are they replicating 20 years later?

John Yewdell:

I don't disagree with the Stanford group of John Ioannidis. I doubt that Ioannidis is right about this huge traction studies being completely wrong.

Aaron Carroll:

This is the paper we referred to in the beginning of the episode, where a group from Stanford estimated that around 50% of research findings are likely false.

John Yewdell:

By putting layers and layers of extra work you have to do it, doesn't help things. And the core problem in terms of fraud, if that's an issue, which is an issue is the incentives that people have in their career. So if you want to fix the problem, you need to fix the career. And just what you were getting at Aaron about the incentives for people to get grants and money. Okay, let's make it more reasonable.

Aaron Carroll:

We'll definitely get to career incentives in the next episode, but let's explore the question of whether fraud is a core problem. From what we've gathered, inappropriate research behaviors aren't the only problem, but they're definitely happening.

Lauren Maggio:

My name is Lauren Maggio. I'm an associate professor of medicine at the Uniformed Services University in Bethesda, Maryland. I'm also the associate director for Scholarly Communications in our Center for Health Professions Education. I've been fortunate to work with a really wonderful team around many different topics. One of the areas where we first started to explore was around questionable research practices in my specific field, which is health professions education. We had no sense in our field unlike some other fields more broadly have looked at what we call QRPs.

We were stunned, we put a survey into the field, a self report survey. It was based on some instruments that had been already assembled and used in the Netherlands. And we assumed we would see no one reporting bad behaviors. What we found is over 90% of respondents did indicate that they had engaged in some questionable research practice.

Aaron Carroll:

The top QRP they found had to do with guest authorship or feeling pressured to give someone authorship who didn't really meet the qualifications for it. And while there are connections amongst all of these dishonest behaviors, we were most interested in those more direct relationships with reproducibility.

Lauren Maggio:

So we had a little over 6% of the respondents telling us that they had changed data in qualitative transcripts. They had changed people's words. We saw 5% admitted to plagiarism, 3.4% deleted data upon pressure from their advisor or a co-author. And then this blew my mind, I didn't think we would see it at all, but 2.4% admitted outright to fabricating data. Yeah, we were surprised. We had close to 600 researchers reporting here.

Tiffany Doherty:

Okay. So 2.4% of 600 is about 14 people admitting to outright fabricating data. Applied to the larger picture of science, this could really mean disastrous things. And given the sensitive nature of these questions, i.e. asking someone if their work has been fraudulent in some way, even if anonymous, this could be a conservative estimate. And so Dr. Maggio's work isn't an anomaly. A quick literature search tells us that several self-report studies are finding similar results. For example, a 2009 study reported that almost 2% of scientists admitted to fabricating, falsifying or modifying their data or results at least one time, and almost 34% admitted to other questionable research practices.

Interestingly, when asked about these same behaviors in their colleagues, these numbers jumped to over 14% for falsification, and up to 72% for other misconduct.

Aaron Carroll:

Most of the misconduct we'd like to spotlight includes things like p-hacking, which is also called data dredging or a few other names, and essentially refers to the practice of analyzing data such that it will yield the results you want.

Sanjay Srivastava:

My name is Sanjay Srivastava. I'm a professor at the University of Oregon. I'm in the Department of Psychology and my main line of research, I'm a personality and social psychologist. But about a decade ago, I started a blog not really knowing what I was going to blog about. And not very long after I started that, psychology had a number of events happen that led to what people have started calling the Replication Crisis, or people call it the Credibility Revolution. And so I started blogging in the out it and then becoming active in other ways and the conversation around that.

Aaron Carroll:

Let's talk about p-hacking. I want to get your thoughts on it.

Sanjay Srivastava:

This article, it's famous in psychology circles, False Positive Psychology. It was really what introduced this... I think it might have been the first time the term p-hacking was used in upon published article. And it talked about these practices and demonstrated how they can lead to huge distortions. And this kind of started the replication crisis, and psychologists, they're asked years later to write a retrospective. And they had this really interesting expression, they said, "When we started this work, we thought of p-hacking, we thought it was wrong like jaywalking is wrong. And when we ran these simulations, we realized it was wrong like robbing a bank is wrong."

And I think that to me, really captures a lot of the pre 2010 mindset of myself and a lot of people I know, having the back of your mind, okay, real research is messy. We've got to make some concessions to the messiness. We can't do everything textbook perfect. And so you would do things, some of these practices that we now have come to call p-hacking like trying an analysis a couple different ways, because you don't know what's the one or you think you don't know what's the one best way to do it. And so you're like, "Okay, there's a couple legitimate ways to do it. I'm going to try both, see what happens.

I'm going to try... I collected 50 subjects. I got a result that's close to significant, so I'll collect 10 more and see if I can get my sample up." Sounds intuitively totally reasonable. It's not, it blows up the statistics in ways that some people... I actually just by luck happened to have run across us that the idea that this wasn't a good thing to do, so I wasn't doing this. So there are all these things that people would do, and we now know that through variety of converging evidence, through simulations, through meta studies, et cetera, that they lead to very large distortions. And they really mean that the resulting work is not credible.

And so I think in 2010, 2011, if you asked me how I feel about it, I'd say, "It's something that a lot of people were doing." We collectively were encouraging people under the table to do it because we didn't realize it was a problem. We shouldn't be blaming individuals. We should be trying to change collectively and viewing this as like we all need to get better together. I think a decade later, it's harder to say that because it's really hard to say, at least in my field that you're a scientist in the year 2020 and you don't know yet that these things are wrong.

It's not to the point where I'd say it's wrong like fraud is wrong, but it's moved in the direction of, I think scientists should be held more responsible for it.

Tiffany Doherty:

So in terms of behaviors that underlie replication issues, data fabrication and p-hacking are two big ones. So we've talked about that a little and we've talked about some of the events that brought this to everyone's attention in a big way. And now we want to talk about evidence beyond those individual stories. And to do that, you have to talk to Dr. Brian Nosek. You heard a clip from him in the beginning of this episode, and now we'll hear more about the major work he does in terms of replication and open science.

Brian Nosek:

My name is Brian Nosek. I'm a professor of psychology at the University of Virginia and executive director of the Center for Open Science.

Aaron Carroll:

How do we know this is a problem? How do we know that it's turning out research, which isn't holding up?

Brian Nosek:

Yeah, that's a great question. And across the various disciplines that have self-examined them to see how credible is our research, the persistent finding is that it's not as credible, the published literature as we might think it is, or as we might hope it would be. And one way of assessing credibility that is the most common way is to say, "Well, if the research literature is credible, then if I try to repeat a finding that is reported in the literature, redo their methodology, try to obtain similar kinds of data and see if I find the same thing, well, then I should."

Because a hallmark of credibility of a scientific claim is that it's reproducible. That you can either reuse the same data and find the same result, that's a low bar. Or you can get entirely new data and find that that same concept still holds when you look at it again in a different way. And a variety of different replication projects that have looked at different samples of the research literature and cancer biology, and psychology and economics, all have a persistent finding, which is when we try to repeat this sample of studies, a substantial subset of them fail to be replicated.

And that doesn't necessarily mean that the original study is wrong because the replication might have been screwed up or there might be important moderating factors between the original study and the final study. But what it does introduce is some uncertainty of, "Wait a second. We thought that this was credible results. We thought that if we do a good faith attempt to trying to get those findings again, that we would." So the fact that we don't now changes our mental orientation of, "Well, maybe we're not as certain about that as we thought we might have been. Maybe we need to look a little closer at how it is we're doing the work.

Aaron Carroll:

So can you walk us through some of your studies that are specifically looked at this? How do we actually conduct those? What are the results? What do we find?

Brian Nosek:

Yeah. The one that we've done that has received the most attention is called The Reproducibility Project Psychology. And the idea for this project was to take a sample of studies, in this case, it was from three journals of research that they published in 2008. Then try to select studies from those journals in that time period to have a defined set, so we're not just cherry-picking findings to try to replicate. Replicate

as many of them as we could. And we were ultimately able to conduct replications of 100 of these studies.

So we had this massive project, 270 ultimate co-authors and other 85 people contributing. Each of them following this protocol of selecting a paper for replication, identifying a specific finding that's the target of replication. And then developing a protocol by requesting the original materials, designing how it is the study would be run. Making sure that it is high powered to be able to detect the finding if it's there. Getting the original author's feedback on that design.

And then running it after preregistering all of those plans to try to get as credible a result as we can on those replications. And in that particular example, we did 100 replications as a group and we were successfully replicated less than 40 of them.

Aaron Carroll:

And these papers were chosen from premier journals in each subspecialty because they wanted to target the work that we is most likely driving the field at that time.

Brian Nosek:

In terms of research, we have continued to start poke at these various explanations for why things might be replicable or not replicable. So just as one example, or maybe I'll give two examples on two polls. One is in part of our labs series where we take the same protocol and we run it in multiple laboratories to see if there's variation in the results. One of the many labs projects that we're just finishing up now is called Many Labs 4. And the idea of Many Labs 4 was to take 10 of the findings from the Reproducibility Project in psychology, where the original authors had said before we ran the replication, we're not quite sure about your design.

So without knowing what the results were, they had expressed some skepticism about whether we would be able to attain their results. And only one of those 10 had successfully replicated in the Reproducibility Project. So an obvious possibility we screwed something up. Those are 10 where the original authors saw a problem. We weren't able to address the problem in our designs, and so no surprise, we failed. So in Many Labs 4, we took those 10 and we said, "Let's run the replication design that we did and Reproducibility Project again, but now with many labs to really increase the power. And let's also run a version that goes through formal peer review until the expert reviewer say, 'Okay, this is an acceptable design.'

And then maybe we'll be able to increase the replicability of those to show that if you don't attend to those tacit knowledge or whatever other factors that experts provide in critique, that you can increase the reproducibility result. So we haven't released the paper yet, but the preview finding is that that didn't make a whole lot of difference.

Tiffany Doherty:

This is a big deal, considering that a major risk response to a study not being replicated is that the people replicating just didn't have the expertise or knowledge that the original authors had. I think at the time, this podcast is out, that paper may be published and perhaps other work done. So if we've got any listeners who are particularly interested in this, go check it out.

Aaron Carroll:

We wanted to know if there were factors that might predict whether specific work was likely to replicate or not. And according to Dr. Nosek, many factors have been proposed, but there are two that are most

consistently observed. The first is that results that are less intuitive are less likely to successfully replicate. And the second has to do with p-values, which in very basic terms is a measure of probability. A standard benchmark in science studies is a p-value of 0.05.

If you run your analysis and get a p-value of 0.05 or less, you call that a significant result. So a consistent observation has been that studies with p-values closer to 0.05 are less likely to replicate than studies with much lower p-values.

Tiffany Doherty:

And I think it's important to note that in many cases, replication issues are not due to anything nefarious. Sometimes things just don't replicate. And that is a result of healthy risk-taking science. Sometimes things don't replicate for a host of reasons that we can't precisely pin down, but we know it's not anything dishonest on the part of the researcher. But we wanted to get back to repercussions for the cases where it does seem that something nefarious is going on, when images and other forms of data are being manipulated.

So far, it seems like getting anything done about it is next to impossible. So as we mentioned, we spoke to someone from the NIH about what happens when fraud is reported to them.

Patricia Valdez:

My name is Patricia Valdez. I'm the NIH Extramural Research Integrity Officer. The majority of what I do at the NIH handling allegations of research misconduct, undue foreign influence, sexual harassment. And in addition, I'm the extramural person who's been in charge of implementing the NIH policy that was rolled out in 2016 on enhancing reproducibility through rigor and transparency.

Aaron Carroll:

And clearly a lot of these happen at the level of the journal on publication, which is outside of your domain. But how are allegations of research misconduct reported to you or the NIH?

Patricia Valdez:

We receive allegations through different areas. We see that during the peer review cycle, we do start to see an increase in allegations. So a lot of times their peer reviewers will come forward to the scientific review officer and bring their concerns to the scientific review officer if they think there may be research misconduct. And we train our staff to handle these appropriately, of course, so we make sure that the reviewers are told to report these things confidentially. We also get anonymous complaints and we also have disclosures from institutions.

Aaron Carroll:

So what happens when an allegation like this comes in? What's the process?

Patricia Valdez:

We try to handle all of these types of allegations similarly. So each institute or center, so that's NCI, NIAID, they have their own Research Integrity Officer for extramural research. So they will, wherever the allegation comes in, in the IC, they then send it to the IC RIO who then sends it to my office. Now, what we do is we do this assessment, so we really are looking for whether or not NIH funds are involved. Whether there's enough specific information there that was suggested that it's research misconduct.

Does it meet the definition of research misconduct? Is it credible? Those types of things. And so once we have that information, then we write a memo ORI, we make a referral to them.

Tiffany Doherty:

That's the Office of Research Integrity.

Aaron Carroll:

And then they take over and they do the actual investigation?

Patricia Valdez:

ORI will take our referral, and so ORI doesn't conduct investigations, they will oversee investigations. And so what they'll do is they'll contact the institution and then ask them to start either the inquiry or the investigation at that stage.

Tiffany Doherty:

We ask Dr. Valdez what some of the consequences might be for committing fraud. One is that if a professor switches jobs during an ongoing investigation, the NIH will contact the new institution and let them know there are potential issues with a grant coming in with their newly hired PI. Another is that specific award conditions might be imposed, basically requiring additional oversight on a grant.

Aaron Carroll:

Are we talking a small, small percentage of awards or is this... Look, if it's greater than 1%, I'm horrified, but is it like an infinite decimal number or a number?

Patricia Valdez:

It's a small number, I would say. But when you think about the effects of one problem and how the butterfly effect goes out, it affects other labs who are trying to reproduce that data and people try to follow up with the data. Their clinical trials, it might be based on the data. The Duke cases particularly were not great.

Aaron Carroll:

Dr. Valdez is referring here to accusations that Duke University submitted false data to win and maintain grants from the NIH and the EPA. Duke has since settled this lawsuit paying the U.S. government $112.5 million. The NIH also imposed restrictions on Duke that have since been lifted

Patricia Valdez:

With research misconduct, it's interesting because it's almost like you have to determine whether or not someone willingly did something versus just didn't know. They're both bad behaviors, right?

Aaron Carroll:

So we've got an overall theme here of inappropriate research behavior occurring, but some general difficult uncovering it and pinpointing where it might land on the scale of misconduct. And it seems that consequences have been limited to the major cases. But we'd like to end this episode on a more positive note. Yes, most people agree that we've got a problem, but we can think about that problem in perspective. Back to our conversation with Dr. Srivastava.

Sanjay Srivastava:

As far as why is it a problem? There are these various mega studies like the Reproducibility Project, they've tried to estimate a percent of replicability. I find that less interesting on a substantive level. It's useful because it gets people's attention to say, "X percent of studies don't replicate." To me it just the number is lower than it has to be. Of course, you shouldn't have a science that's 100% reproducible because that means you're not taking any risks. So there are good reasons for things not to replicate sometimes because you're pushing new boundaries and you have to have room to be wrong.

But there are bad reasons like you're making unforced errors, you're doing things that we could be doing better in a way that's not giving up risk-taking and curiosity and boundary-pushing. And I think it's a large enough gap that it needs people's attention, it needs a collective effort to get better.

Tiffany Doherty:

And we think that collective effort is growing stronger. Reproducibility has become part of the larger conversation now and that's a good thing. Okay. So that was a lot. The overwhelming opinion is that there is a problem. Some people think it's specific to certain fields, some people think it's recent, but it appears to be a widespread phenomenon and one that has been around a while, although seemingly made much worse by the incentives that exist in today's scientific culture. And that's a major theme of the series, how science culture contributes to the reproducibility crisis.

Why would scientists p-hack or falsify or otherwise commit misconduct that prevents us from finding real answers? Which is presumably what we all really want. In our next episode, we'll try and get to the heart of that question, really digging into the incentives experienced in today's scientific culture.

Aaron Carroll:

Thanks for listening to this special episode of the Healthcare Triage Podcast, part of an eight-episode series on science culture and reproducibility edited and produced by Stan Muller and Mark Olson. We would like to thank the National Institutes of Health for funding this series and a special thanks to our guests for lending their time and expertise. If you're interested in incorporating this series into your undergraduate or graduate courses, please visit www.healthcaretriage.info/reproducibility-podcast, where you'll find free lesson guides to accompany each episode.