Reproducibility Podcast Episode 2: Why is there a crisis?

Aaron Carroll:

Welcome back to the Healthcare Triage Podcast. This is episode two of our series on science culture and its effects on reproducibility. In our first episode, we zeroed in on reproducibility itself. We spoke to experts from various fields about what exactly the replication crisis is and how it came to be recognized in the mainstream. In this episode, we'll zero in on the why. Why might scientists p-hack, falsify, or otherwise commit research misconduct that ultimately prevents them from uncovering real and lasting answers? We ask the experts, why is this happening?

Speaker 2:

Bad incentives.

Ivan Oransky:

Science is competitive even for people who don't commit... In fact, maybe more competitive for people who don't commit misconduct. You could argue.

Speaker 4:

We are all kind of falling into this trap of publish or perish.

Simine Vazire:

I think we're incentivizing the feeling of discovery.

Ed Yong:

I think all the incentives on academia are just wrong and perverse. As a result, a lot of scientists are incentivized to produce work that is splashy and intention grabbing instead of work that's actually rigorous and true.

Brian Nosek:

The big challenge as we see it is that the values that we have for how we think science should operate are not aligned with the culture, the incentives, and the policy landscape for driving how science does operate.

Speaker 8:

And even Dr. Yewdell, who you heard from in the first episode and who does not think that replication is a major issue in science made a nod to these pressures.

Jon Yewdell:

I think people have to play the game. And I think that's a problem that the incentives are there. And I think it may affect people at a weak moment like they're submitting a grant and they have a questionable finding, but they really need to get the money. And so they'll ignore, right?

Speaker 8:

So we're getting the impression here that incentives baked into our academic culture are an issue.

Aaron Carroll:

They're a huge issue. And when the incentives include your whole career, you might go pretty far to achieve them. And beyond outright fraud, there are ways to achieve these incentives that really hurt science. And two that came up over and over in our interviews were problematic methodology and bad statistics.

Mike Dougherty:

Part of this comes from just some of our sort of methodological weaknesses within the sciences.

Steven Goodman:

A lot of these issues around research design, the use of statistics is inferential, tools and not just as technologies to reveal patterns.

Andrew Gelman:

Currently statistics is sometimes sold as being a kind of machine to convert uncertainty into certainty. So you do an experiment and if the p-value is statistically significant, then you've made a discovery. And if the p-value is not statistically significant, then you say that the effect is zero.

Speaker 8:

So we've got methodological problems and fixing those will go some way to fix things, but why are these problems so easily perpetuated when we know that they stymie our ability to make true scientific progress? All signs point to culture, a culture that functions more like a business than a true scientific enterprise.

Aaron Carroll:

If you want success in academia, you get it with publications and funding. For both of those, you need exciting data. And although the media portrayal of science makes it seem like it's a series of Eureka moments, it's just not. More often than not, analyses come back with no results or at least not very splashy results. And you can't count on keeping your job and paying your mortgage with less than splashy. We heard versions of this from many of our experts.

Josh Nicholson:

My name is Josh Nicholson, Co-founder and CEO of scite. Scite is a platform for researchers to discover and evaluate scientific literature. There's a game in science, right? If you want to do well in science, publish in these highly selective journals, go to these highly selective schools, get funding and you'll do well, right? And in order to do those things, you have to come up with things that are surprising, sexy, flashy work, as opposed to careful, messy, and maybe kind of boring in some instances. But a lot of, if you look at the history of science, a lot of really impactful things came out of boring work. It's just hard to judge it until years later.

Paula Stephan:

My name's Paula Stephan, and I'm a Professor Emerita at Georgia State University, and I'm a Research Associate at the National Bureau of Economic Research. And for at least the last 35 years of my career, I focus pretty much on the economics of science and the careers of scientists. Well, my research is really

focused on several pieces of this. Economics is about incentives and cost. And I've been very interested on how incentives and costs affect the practice of science. I think looking at short term bibliometrics is a huge problem. I really think that we need to be worried about just flashes in the pan. Well, I mean, of course my biggest concern is that people don't read articles, that they just look at where they're published.

And I think that's very true of search committees. I mean, at my university, when we put up people for promotion in economics, the provost always wanted to say: Wow! How many citation classics does this person have? How many... And furthermore, they didn't understand how much fields varied. But all they wanted was bibliometrics. I'm really interested in how places avoid risk. And I think bibliometric is a way that provost and committees and selection committees kind of outsourced the risk. They say: Oh, well, science decided this person's research was good enough and they published it. So we're going to hire them.

Aaron Carroll:

We'll hear more from Dr. Stephan in a minute. First though, when we're speaking of incentives, we must also nod to the absence of certain incentives. We talked to one of our experts about this.

Kelli Barr:

My name is Kelli Barr. I studied philosophy. I'm currently a postdoc at the University of Pittsburgh in History and Philosophy of Science Department. So if one was putting together the picture of the landscape, like the motivational landscape and the economic landscape and the political landscape and the social landscape of what it meant to do scientific research in the early 21st century at research universities or comprehensive universities, I don't think you'd be also that surprised that the replication crisis is happening. What were the structural incentives that were built in, in order encourage people to replicate work? Has replication really been that strong or consistent of a norm of scientific practice?

Speaker 8:

So we talk and think about replication like it's a major tenet of science and it should be, but those thoughts and ideas aren't often translated into practice. The system simply isn't structured to value and uphold replication.

Brian Nosek:

It's much easier for me to publish a novel finding than publish a finding saying: Oh yeah, I found that you were right. So why would I conduct a replication if journals don't care about replications, if funders won't fund replications? Innovation is the engine of scientific advancement, but the incentive models have gotten so extreme that its innovation at the expense of verification.

Aaron Carroll:

That was Dr. Brian Nosek from our first episode, Professor of Psychology at the University of Virginia, Executive Director of the Center for Open Science, and a primary investigator on the reproducibility project psychology. Dr. Sanjay Srivastava, a Psychology Professor at the University of Oregon, who you heard on the first episode of this series felt similarly.

Sanjay Srivastava:

We are structured as a field to reward certain kinds of outcomes. And there's this idea that science is self-correcting, but we've got certain incentives to produce certain kinds of results. We don't have

incentives to check each other's work. If you're asking me for the psychology of it, it's the intersection of incentives and dispositions. Scientists are human beings. We have good motives and we have bad motives. Everybody has them. Or not bad, but just mixed. We want to discover exciting new things and we want to gain status and we want to make money and all these other things. And different things in our environment activate these motives and they give us opportunities to express them. The more opportunities there are to activate our desire for status and money and fame and all that, we're going to be human. And those are going to push us in those directions even while we also sincerely want to make important discoveries, help people, and so forth.

Speaker 8:

And now let's turn to our interview with Dr. Simine Vazire to expand on something she said that you heard a snippet of at the beginning of this episode.

Simine Vazire:

My name is Simine Vazire. I'm a psychology researcher at the University of Melbourne. I always kind of studied research methods. And now I've shifted towards doing that more and doing metascience and studying kind of the replication crisis itself and potential solutions and how to make psychology research better. I think we're incentivizing the feeling of discovery. So that means incentivizing novelty, but also not actual novelty, but just like if you can get the reader to feel like you made some big, important discovery, even if it turns out it was a false positive or it's exaggerated or whatever, I feel like what gets rewarded is just meeting that bar of someone's impressed and thinks that you've made an important discovery, but if they scratch a bit below the surface, it might not hold up. So then you can do that a lot faster and quicker and be more productive than if you're trying to hold yourself to a higher standard of not publishing something until you're sure about it and publishing it even if it's a null result.

Aaron Carroll:

You also heard from Dr. Lauren Maggio in the first episode. And we're going back now to our conversation with her to touch on the relationship between science culture and reproducibility and to hear more on her work looking at QRPs or questionable research practices.

Lauren Maggio:

I think our culture breeds unfortunately many negative things that impact reproducibility part of it. And this has been interesting. I've been thinking a lot about promotion in tenure lately. I just put together my package for promotion in tenure and have been thinking about, even as I constructed my package, how the documents had to be laid out, how I had to set up my CV, how I had to put numbers next to all of my publications, whereas I didn't put numbers next to the courses I teach or next to the number of learners that I mentor. I think we're very much into quantification. There's that age old kind of joke that deans all know how to count, P&T committees all know how to count. We can add up publications pretty easily. We can add up grants dollars pretty easily instead of having to do a more deep content analysis of someone's CV.

Aaron Carroll:

And so you think we're looking at quantity, not quality.

Lauren Maggio:

Yes. Another thing that we looked at that I thought was quite interesting. We had our survey participants also fill in a publication pressure survey. We've seen that publication pressure is a large indicator of QRPs. And we saw that play out in our study as well, where that was the largest factor in whether or not someone was going to commit a questionable research practice, which kind of gets to the root at what we've been talking about today. People are wanting to get publications. They're feeling extremely pressured to do so.

Mike Dougherty:

Well, think about it this way. When people look at academic records, the first thing that pops out is the number of publications, right? We look at the sheer volume of work that people produce, right? I'm Mike Dougherty. I'm a Professor of Psychology at the University of Maryland. Well, if I were to have you sit down and I say: Okay, Aaron, your job is to make for me as many paper airplanes as you can in the next 15 minutes, right? And I'm going to pay you and I'm going to reward you for making as many of those paper airplanes as possible. What are you going to do? Well, you're going to maximize output. But when you maximize output like that, what happens? It's likely that you're going to make mistakes. Not intentionally, right? You're going to be maybe a little bit sloppy. Your airplanes aren't going to fly as optimally as they could. The work isn't going to be as good as it could be if I said, "Build me the best paper airplane that you can possibly build." And if you can make multiple of those, that's great.

But let's focus on getting me an airplane that is going to optimize its flight path. It's going to optimize flight distance. It's going to stay in the air the longest. By incentivizing quantity, there has to be this inherent trade off with quality, whatever that quality dimension is and however we define it. When I open up a job application, I'm looking at this material and I'm comparing it to the person next to me or the other people in our application pool. And again, looking at just the sheer volume of research that people put out, right? We tend to equate productivity with number of publications, right? Without taking into account what's in those publications, how that research was carried out, and so forth. And so when I look at those, right? And I'm comparing someone who has say three publications from graduate school versus someone who has 10, the immediate reaction is: Oh, this person who has more must be better, right? Now, that person then becomes the new norm.

Aaron Carroll:

So how does this happen? Has it always been this way? It doesn't sound like it. We gained some insight when we sat down with Dr. Paula Stephan, who you heard in the beginning of this episode.

Paula Stephan:

One of my favorite quotes came from Richard Catlow when he became the foreign secretary for the Royal Academy two years ago. I just love what he said about bibliometrics. He said, and I'll read it to you, "I was lucky. When I began my scientific career in the 1970s, I had no real sense of how my work was cited. My discipline, computational materials-chemistry was barely acknowledged by mainstream chemists. If I had been citation-driven, I might have abandoned a field that is now central to developing sophisticated materials, including porous catalysts, electronic ceramics and ionically conducive materials. By the 1990s, when citation data became prominent, I was already a full professor." I think that really summarizes a lot of these issues. And that's just completely changed with platforms such as Google Scholar, et cetera. And people are just incredibly aware of it. And look at most people's web pages, it comes up instantly, their number of citations, their h-index, et cetera. And I think that has really had a big effect on the practice of science.

Speaker 8:

Dr. Stephan touched on another insidious outcome of our obsession with short term bibliometrics, explaining that papers with really novel findings, so those most likely to move science forward, don't start picking up citations until a few years after publication, whereas non-novel papers do really well on citations in the first couple years. So this means that we are potentially rewarding people who do fast and sloppy research while those doing the careful work that might really move the field forward are left to struggle because according to the short term numbers, they aren't being productive.

Aaron Carroll:

Could I get you define bibliometric data a little more clearly just for people who are going to listen to this? What do you exactly mean by that?

Paula Stephan:

Well, originally bibliometric data is invented by Eugene Garfield was really looking at citations, looking at every article that's published and looking at the references in that article, and then adding up those references for each author so that you could see how many times your previous work was being cited. And then of course, we began to have embellishments of that, we could say. One embellishment is the journal impact factor that allows each journal to compute how often the articles in the journal are cited page by page, in a sense. And we know that for example, Nature has the highest journal impact factor. Science is not that far below.

And so these are very high journal impact factors, and they seem to mean a great deal in science. If you talk to postdocs in the United States, many postdocs will tell you that their get out of jail free card is publishing in one of those three top journals. So people really focus on that. And then of course, they're all kinds of variations, for example, the H factor that instantly comes up on Google Scholars, on... I mean, I looked you up today and just the first thing I looked up, it tells me your citations and it tells me your journal impact factor. Okay.

Aaron Carroll:

No, I know it's there, absolutely. We also sat down with Dr. Diana Hicks who works in bibliometrics as well. And we talked to her about how the rise has contributed to science becoming something like a system to be gamed, as opposed to a useful tool to measure quality.

Diana Hicks:

My name is Diana Hicks. I am a professor at the School of Public Policy at the Georgia Institute of Technology in Atlanta. And my research and teaching is in science policy, especially to do with bibliometrics.

Aaron Carroll:

Is science always been a game like this, or has it is something changed?

Diana Hicks:

When I worked in Europe, as I worked at University of Sussex for a number of years, I mean, they didn't even really have a tenure process. They didn't even have an annual review... There was like no evaluation. So nobody was playing any games. There was nothing to game. Then I came over here and

we've got our annual evaluations. We've got the tenure process. Professionals are smart. So it's very hard to devise some simple numerical system that is what somebody at a great distance who wants to oversee something needs, but when you're on the ground, you're being evaluated by a number and you got any intelligence at all. You can find a way to optimize things without necessarily optimizing your performance.

Aaron Carroll:

And we should note that these incentives start at the top, meaning they shaped the behavior of the university itself, which then spreads down to affect everyone else. Back to Dr. Stephan for more on this.

Paula Stephan:

I think that one of the things that has made the system even worse is the obsession of universities with their overall reputation. I mean, for example, the most prestigious thing a university in the United States or Canada can want is what? Membership in the AAU. Okay. So membership in the AAU, what's it based on? It's based on how many members you have of the national academy and it's based on your external funding. Okay. So you can have universities in the United States that want to create medical schools because they want NIH funding. And they think that will get them membership in the AAU. And if you flip to Europe, do you know about the Shanghai rankings?

Aaron Carroll:

No. Please tell us.

Paula Stephan:

Shanghai rankings came out of a university in Shanghai about maybe 20 years ago in which they started ranking universities internationally. And their rankings were based on how many Nobel Prizes you have, publications in Science, Nature and Cell are same old three friends. This really set the ranking business, the international ranking business just off incredibly. And now of course, they're the rankings of the times higher education that play a very large role. So why do universities care? Well, they care I think for raising money in the US from their alumni. Prestige is associated with that, but they also care. And this is definitely true in Europe, but I think true here, although we don't talk about it very much. They get international undergraduates to come based on rankings. And international undergraduates and international master students are the cash cow of a lot of these programs. Okay. So once again, we've gotten ourselves into this ranking game and into kind of its economics is playing another role here.

Speaker 8:

In terms of science not always being this way, other people we interviewed seem to agree. And many of our interviewees lamented the fact that the idea of a scientist spending their entire career having never made a major discovery is just unacceptable in our current culture. As one of our interviewees describes it, and I'm paraphrasing, say that you spend your entire career exploring a blind alley and you come up with nothing. That's okay. It had to be explored, right? But in our current culture, such "failed exploration" almost guarantees that you'll be seen as a failure. Moving past what things used to look like, let's dig a little more into what it looks like now in terms of incentives and research quality.

Ed Yong:

I'm Ed Yong. I am a science writer at The Atlantic.

Speaker 8:

We sat down with Ed Yong because he has written extensively about the reproducibility crisis.

Aaron Carroll:

Why do you think research that has difficult time being replicated gets published?

Ed Yong:

I think all the incentives on academia are just wrong and perverse. Academia rewards scientists predominantly for the act of publishing a lot of papers in high profile journals like more than anything else that determines success in securing grants, in achieving tenure, in being asked to speaker conferences. It's a metric that has come to define professional success in academia. And I think as a result, a lot of scientists are incentivized to produce work that is splashy and intention grabbing instead of work that's actually rigorous and true. Now, let's be clear here, this isn't to say that all scientists are producing sloppy work. What I'm saying is that academia almost like by default pushes people towards sloppy work. It is almost an act of resistance to do careful, rigorous science when there's so much that rewards people for doing the opposite. There is certainly a proportion of this crisis that is to do with deliberate misconduct and fraud. I don't think that's the majority, far from it. I think a lot of people are not deliberately seeking to mislead the community, but I think they are doing shoddy work nonetheless.

Speaker 8:

We also spoke to Dr. Ivan Oransky, who had similar thoughts on misconduct and fraud.

Ivan Oransky:

I'm Ivan Oransky. I'm the Co-founder of Retraction Watch, also Vice President of Editorial at Medscape, and Distinguished Writer in Residence at New York University's Arthur Carter Journalism Institute. I just want to be clear that reproducibility is a much bigger problem than scientific fraud and misconduct. And I say that full on thinking about misconduct and fraud, mostly in terms of Retraction Watch. So that's just sort of in terms of table setting. But what they also are is, and Adam Marcus, my co-founder is fond of saying, issues in reproducibility and fraud and misconduct, they're all sort of born of the same mother. And that mother is publish or perish. And the sort of need to publish papers to get grants, to get tenure, to get promoted, to really advance and so, or to win prizes. But, there are ways to cut corners in pretty much anything you do.

There are scientists who have started to cut corners that again may not be considered "fraud or misconduct" but are definitely not good scientific practice. And then there's some that we just don't understand the system well enough. So it's not reproducing because we didn't think about it the right way. And that could be because we cherry-picked or we only looked at data or criteria that would help our sort of argument. And so we're looking not at the entire picture. But it's actually very similar. So I think that the publish or perish problem, which again has been studied and people have looked at fairly extensively, that's at the root of a lot of this.

Aaron Carroll:

Publish or perish is a major cornerstone. This came up in almost every conversation we had for this series.

Jeff Flier:

I'm Jeff Flier. I'm a physician, MD degree, got that in 1972 at the Mount Sinai School of Medicine. Dialing forward to the latter parts of my career, I ended up as the Dean of Harvard Medical School, which I was for nine years. There is a culture that definitely changed from the time that I started to today in which there is an excessive expectation that you will try to publish and publish in some of these fancy prestige journals that are the ones that for some purposes, you're just not going to get promoted or appointed if you're not publishing there. The journals they're looking for papers that make claims that will allow them to say: This is transformative. This is amazing. This is going to change everything. This will get clicks. Scientists being smart folks typically learn that they're looking for those things. So they orient their findings to fulfill that request.

They want to take their findings. So there's two things. One is the things they choose to work on are sometimes chosen because they think they can make that claim. And secondly, and this is the more insidious part, even if they're not legitimately making the claim, they try to construct a paper, sometimes using inappropriate methods so that they can make the claim, even if it isn't true. So it's a little game between the journals and the editors and to some degree the reviewers, but the final decisions are the editors and the scientists who are kind of trying to figure out at this leading edge of publishing how to get their papers into those journals, what do they need to do to do it, how do they change their approach, the way they sell the findings, and even the way they choose the data to put into the papers. So that's a very broad statement of a problem in the culture of science.

And then of course, it extends more broadly. It's not just the scientists doing it because they think this will be a fact facing them. It's also, they know when they're being counseled by their department chairs and by their faculty affairs departments that: Gee, do you have a paper in Nature or Cell or Science or New England Journal or whatever? No. Oh, you don't. Well, maybe you need three more papers. Whereas if you had one in this journal, maybe that'll do the job. And I know this because I chaired, I just figured it out. I chaired over nine years about 90 professorial promotions committees at Harvard Medical School. I can't tell you how many times I've heard people say: Oh, to go from assistant to associate, you need 25 papers. Well, that's not true. We have no criteria about the number of papers, but that lore is still out there. Why is it out there? It's out there because in practice, that's what ends up happening. But there's these ways of assessing quality that are not themselves quality.

Aaron Carroll:

So it's that getting into like the more prestigious journal as opposed to better science?

Jeff Flier:

Correct. So I wrote an article in Nature a year and a half, two years ago, which said that promotions should pay greater attention to quality and reproducibility. When we're assessing people's promotions, we don't spend enough time explicitly talking about the quality of the data.

Speaker 8:

Here's Dr. Nosek again.

Brian Nosek:

There are lots of ways we could talk about the dysfunctional culture and incentives, but for us, it comes down to a core issue, which is that the incentives for my success as a practicing researcher in academic research are focused on me getting it published, not on me getting it right. And there are different behaviors that I can do that make my research more publishable at the expense of its credibility. And even though I want my research to be credible, even though I'm trying to do research as accurately and

as impactfully as I can, the fact that I also want a job and I want to keep my job and I want to advance in my career introduces potential for bias in the decisions that I make in how I analyze my data, how I report my findings, that create a conflict of interest. What's good for me versus what's good for the science.

Aaron Carroll:

How do you think we got to a place in science where the incentives were misaligned from what we would like them to be?

Brian Nosek:

Of course, there are lots of likely factors in that. And I can only sort of give a just so story for how it is we got here with any kind of modest precision, but I think there are some obvious elements that likely contribute. One is that it is a competitive landscape. There are only so many jobs for so many people that would like to have those jobs, particularly in academic careers. And so once there is competition in the system, of course we need ways to distinguish who's doing better and who's doing worse. Just that fact is something that many high performance disciplines deal with, but there are some unique things that we have done in science that create some dysfunctional incentives in how we evaluate who is good at their job. And one of those is that publication is the currency of advancement. Getting my name on a paper is an important part of me showing that I do productive work, that I am a productive scientist.

Being the first author or the last author on the paper is even more important for that. That is of course, well yeah, you need to have some sort of outcome system to be able to count whether people are contributing to the literature, but that then is married with this fact that there are different places that I could publish my work. And some of those places are better for me than others. There's a hierarchy of the journals. And if I can get my work into Science or Nature or Cell, that's a big outcome for me compared to some journal with seven words in the name. Okay. As competition, there's limited space for the most prestigious outlets. And there's all kinds of things that we can do to try to make it so that our work is more likely to achieve those prestigious heights. That of course just is an incentive landscape that is difficult for resisting exploiting flexibility in how it is we do our research and report our research in order to get those better rewards.

And rewards stack upon each other, right? You get more publications, you get more publications. You get more publications, you get more grants. You get more grants, you get more publications. And so all of these things sort of work in concert with each other and arrive at this sort of simple counting heuristic. How many publications do they have? How prestigious are the outlets? How much money did they bring in? Oh, they must be a good scientist. Hire them. So it's no surprise that in that context, institutions have defaulted to the lowest common denominators, which is like: Count how many citations they have. Calculate their h-index. What else are we going to do? It's complicated. And simultaneously of course, institutions realize and many do excellent jobs of doing deep dives, especially for promotion, right? Hiring is hard. I have 200 applications. We're going to hire one. I can't read everybody's work. Promotion should be easy. It's this one person. And we're going to spend like the whole year deciding whether we want to keep them or not.

So we do have time to read. So we can implement practices of actually going into people's work and evaluating the quality of their work in those committees. But what we're not doing I think, as well as we need to do, is providing the messages to the researcher, him or herself, to help guide them on what it means to be successful. I'll give an example of this in my own experience just to illustrate. I went up for full professor a few years ago. And the administrator in our department said, "Please print out all of your

papers and submit them to the committee." And the printing out first of all, that's ridiculous. It's the 21st century.

Aaron Carroll:

Dark ages [crosstalk 00:33:57].

Brian Nosek:

Right. You got to be kidding. But print out all of my papers for full professor promotion. I had a hundred papers. Why do you want me to give them all to the committee? So that's the criteria is you have to submit all of your papers so the committee can review them. So that-

Aaron Carroll:

You need to kill a tree for that.

Brian Nosek:

Yeah. The tree is going to die. And what the committee is going to do is bring out a scale and weigh the papers. Oh, that's a tall stack. Promote him, right? And they're not going to read a hundred papers. So what they have communicated to me, it's barely even implicit, is volume is what matters.

Speaker 8:

And now we should touch on something that might feel uncomfortable. At what point do we start pointing fingers for the kind of behaviors that this inspires? Or should we point fingers at all? We're all laboring under the same system of incentives. And it appears that a pretty large number of people are engaging in these behaviors that more often than not corrupt the science. Generally, when we asked our experts if the real problem was bad actors or bad incentives, most of them pointed at the incentives without hesitation, as in these are good people who want to do the right thing, but they also have to pay their mortgages and feed their families.

Aaron Carroll:

So we're talking about researchers who know what the right thing to do is, but don't do it because of these incentive constraints or they unknowingly engaging in behaviors that are destructive to science because this is simply how they were taught to succeed. Based on conversations with some of our experts, we think this could be part of the explanation that we'll talk more in a bit about just how much of it. For now, we want to touch on training, specifically training and statistics and the appropriate way to view and handle data. While we touched on statistics with all of our experts, we took a deeper dive with two of them, Dr. Andrew Gelman and Dr. Steven Goodman.

Andrew Gelman:

My name's Andrew Gelman, and I'm a statistician, and I also am a political scientist.

Steven Goodman:

My name is Steve Goodman. I'm an Associate Dean and Professor of Epidemiology and Population Health and of Medicine at Stanford University. So a lot of these issues around research design, the use of statistics is inferential, tools and not just as technologies to reveal patterns is much newer, particularly the emphasis on design and the marriage of question design and statistics. There's a sense, I don't want

to paint with too broad a brush, but among many who've not studied statistics except in a purely instrumental way, they have a set of data and you just throw statistics at it and it just shows you the truth or shows you the signal. They're far less accustomed to thinking in terms of large amounts of uncertainty. They tend to think mechanistically, which is also, I want to say why they achieved a lot. They want to know how things work. The idea is not that there's going to be this huge uncertainty around various parameters. They literally want to know if this causes this and this affects that. They think in a far more deterministic way.

Aaron Carroll:

Dr. Gelman agreed.

Andrew Gelman:

Currently statistics is sometimes sold as being a kind of machine to convert uncertainty into certainty. So you do an experiment and if the p-value is statistically significant, then you've made a discovery. And if the p-value is not statistically significant, then you say that the effect is zero.

Speaker 8:

Here's Dr. Goodman again.

Aaron Carroll:

Let's talk about p-values. So much of our focus unfortunately is on P is less than 0.05.

Steven Goodman:

Correct.

Aaron Carroll:

And I'm going to assume you think that's a huge mistake.

Steven Goodman:

First of all, it is not dissimilar from the reduction of all complex evaluations and whether it's your suitability, whether you're going to be a good doctor on the basis of your med CAT scores. The reduction of a complex inference to a single number is automatically going to be a problem from the get go. The second is that it has a cutoff. So any reduction of the scientific process, or again, of any complex evaluation process to a single number that then has to pass a magic threshold in order to be viewed as reliable or not is going to be a big problem. So that is the second problem. Again, it's not specific to p-values, but p-values have become sort of the target.

The third, and this is where some of the focus now is, or has been, is that it doesn't mean what anybody thinks it means. They think it has something to do with the reliability of your conclusion, that is you're 95% certain to be right in whatever claim you make. And of course, it has only a very indirect and tortured relationship to that. The biggest problem with p-values and the whole system of statistics is that they've robbed us of our ability to talk about uncertainty in any quantitative way. The metaphor that's actually very apt is the temperature. How do you know what 75 degrees means? It's because you've felt it a million times. So when you say 75 degrees, we both know what that feels like. When you say 85, we know what it feels like.

We're extremely practiced. But imagine if the way we reported the temperature was: If it was over 90, we would just say hot. And under 90, we would say not hot. That we weren't allowed to say anything else, hot, not hot. The frequentist approach, which is embodied by the use of p-values and hypothesis testing. The most dangerous product of it is that it has robbed scientists of the ability to talk about gradations of uncertainty and to know how it feels. We think in binary ways, and we've robbed ourselves of a hundred years of experience of knowing what a partially confirmed result is, and being able to communicate that in words, where that can be a shared experience and we can have reasoned debate.

Aaron Carroll:

So is the bigger issue a lack of real statistical understanding among many scientists or bending the rules, thanks to the pressure of incentives? Here's Dr. Gelman again.

Andrew Gelman:

I feel all these people from the incompetence to the cheaters and everybody in between, they think they're doing good science. I think they think of statistics as being like paperwork, the hoops that they have to jump through. I've seen enough things, where people seem to think that an effect is real. It'll show up in every sample of data they'll possibly see. They think that. So they run something. Oh, P is 0.04. I can publish it. Or even P equals 0.08. You can publish that too if you're a good writer and you have good connections, right? If you have a friend in a journal, whatever. So they do that and, oh, this is real. And then, it doesn't replicate. Well, there's a reason for this. Blah, blah, blah. Oh, I do this. And now the p-value is no good. Oh, this is a pain because I know the exact same result if the p-value were 0.04, I could get it published. So I'll do what it takes.

So sometimes I think even the cheating, I doubt that even the cheaters think they're cheating in a deep way. I think that they think that they know what's going on. People have this phrase, statistical significance does not imply practical significance. But in addition, my problem is not so much that, but from more of the theoretical grounds that if you really found it was 1%, 0.1 percentage point, you really believed it, you really had ironclad proof, I'd still say so what? Not just because it's a small amount, but because I have no reason to think it would be the same sign or the same magnitude the next year in a different place. There's no reason to think it's a universal. Well, they have a story, which is if it works, it really works. If it doesn't work, it doesn't work. So that's a mental model of the world and I think it's wrong. And that's how we talk, that's how we write statistics books. So one of my themes is that statisticians, it's our fault, a lot of this.

Aaron Carroll:

Well, I think everybody's at fault.

Andrew Gelman:

But that's how people are taught, right? If it's statistically significant, people are taught that there's a 95% chance you found something real.

Speaker 8:

We also talked about this with one of our science media experts, Christie Aschwanden.

Christie Aschwanden:

I'm Christie Aschwanden and I'm a science journalist.

Aaron Carroll:

I think we could argue that there are lots of scientists that don't fully understand statistics.

Christie Aschwanden:

Oh yeah, that's absolutely correct. Most people in the media and the general public don't understand statistics. Many times scientists think of statistics as this program you run on your data at the end to figure out what you can publish. And that's not what it is at all. And statistics are something you should be thinking about before you even do the study. Which data are we going to collect? How will we analyze it? But that's not always the way that it's thought of. And so then when those scientists go to explain their research to the public or to the media, they use things like: Well, this is statistically significant. And yeah, we know now. I mean, I've written very extensively about p-values and p-hacking and the problems with these approaches. When the public hears significant, they hear important or real. And these terms have different meanings within statistics than they do in the general public.

Speaker 8:

So are we not training people well enough? Are we training them wrong? Here's Dr. Goodman again.

Steven Goodman:

We actually did a survey of the graduate students here to figure out what statistics courses they were taking. And to my non surprise, they take a huge number. They're very smart, huge number of quite advanced courses. I mean, just mind blowing technical [inaudible 00:43:36] force type courses. But what we ask them, what confidence do you have in knowing what a confidence interval is, or a p-value, or the purpose of randomization? Guess what? Very low confidence. So a lot of them are building skyscrapers from the second floor up in the quantitative field, particularly in the domains of machine learning and all these new techniques. They learn how to apply them, but they're not really taught the foundations. What makes for a secure foundation of knowledge? What do these basic units of uncertainty mean quantitatively and qualitatively? That is not really emphasized. It's the technical stuff that's emphasized.

Aaron Carroll:

So we talk to people who range from thinking that most of this is like outright fraud, that it's people who know what they're doing is wrong and they're totally actively engaging it to those who think it's people doing what they think is right and this is just how they were taught and how it all works. And where do you come down in that spectrum?

Steven Goodman:

I feel very strongly about this. The problem is not fraud. It's not. There are certainly fraudsters out there. We already have structures around detecting fraud there, actually not that great, but the vast majority of scientists are in it for the right reasons. They're in it because they want to make a difference. They're in it because they want to discover truth. And the problems we have, we wouldn't have this depth of problems if it was just about personal integrity. They are using methods that they are in fact taught to use that produce unreliable results and they don't know it. Do they think they are producing virtually unusable results? No, they don't, but they are. And in the basic sciences, it's very easy to open the pages of Science or Nature and find experiments with an N of three, no joke, N of four, without a full recognition or representation of what that means.

Aaron Carroll:

Just breaking in here to note that N refers to the number of subjects in a study. So an N of three means the experiment had a total of three subjects. With few exceptions, this number of subjects is far too low for making any real conclusions.

Steven Goodman:

Is that fraud? Is it malfeasance? No, they don't know, selectively reporting experiments that work. They actually, if interviewed, they'll say, we didn't want to report the things that didn't work because we finally got all the parameters right and the reagents right. They don't recognize that's selective reporting is actually a profound problem. The fitting models with many terms and trying to refine all the terms so it fits just right. That's what they're taught as good science without recognizing that they're shrink wrapping their statistics to their data and they're not producing any knowledge of generalizable importance.

Speaker 8:

I just want to stop here and emphasize that line, shrink wrapping statistics to the data. I think this is an excellent description of one of the biggest problems that seems to be occurring pretty regularly without many checks.

Steven Goodman:

This is a huge problem, both in clinical science and in basic science. Anomalous results get thrown out because they're part of what didn't work. So these are not people who are fraudsters. These are not people, I'm going to say, who don't know how to do science, but in some sense, they don't know how to do these things. This is part of the fabric of science. When I say it has to start with training and retraining and a new culture, that's what I mean.

Speaker 8:

Most of the people we spoke to agreed that the problem isn't necessarily with bad actors. Most agree that the majority of scientists are being genuine and doing what they think is best. But there is an argument to be made that some of these issues have come far enough into the light that claiming ignorance of them is no longer acceptable. Here's Dr. Vazire.

Simine Vazire:

I think it's getting harder and harder for people to claim ignorance unless they're brand new to the field. So what used to be inexcusable like: Oh, you probably didn't know, but actually this has bad effects. It's getting harder and harder to bring up like: Oh, but you did this practice that we now know increases the risk of false positives by a lot, because of course they know that. So when you point it out now, you're embarrassing them or something like that. Whereas before, everyone could believe and it was often true that the author didn't know that and wouldn't have done it if they knew. I think it's a substantial amount, not just ignorance, a substantial amount of it is not having the strength, the will or something to change it. And it's not that we actively want to do bad things. It's that it's really hard to do the right thing.

Speaker 8:

Dr. Gelman said something along the same lines.

Aaron Carroll:

So why do you think there's resistance from so many people to trying to make significant changes to this?

Andrew Gelman:

I think its hard work, right? I mean, it's kind of annoying to have to do something different.

Aaron Carroll:

And so again, we ask, how did we get here with statistics? Was it always this bad? Or have we somehow warped it as we went along? Dr. Goodman had some insights.

Steven Goodman:

Well, there's a long history. But if you look at the writings at the beginning of the people who in a sense created this system, they never intended it to look the way it is today. These tools began to be used by scientists in a way they were not intended to be, the logical foundations were not fully explored, and we created a monster is what I will say. And it's easy. It took a part of science, which is really the hardest part, which is how to bring together everything you know about the science and then the design and the measurement and all the things that go into creating scientific hypothesis and deciding what the strength of the evidence is, which is very difficult. And it made it easy. It made it automatic. And it's like, why do we use SATs if they're so unreliable as predictors? Why do we use any number of predictors in society? It's because they're easy and they become enshrined and they become targets themselves.

And this is the famous Goodhart's law, which is any measure that becomes a target ceases to have value as a measure. This is really critical to understand. You have to understand how these evolved historically because what I will say is that many things that now in the current context look very bad. We're better than what went before. So the evolution of the application of formal statistical methods and formal statistical testing was a great leap forward in its time, but it began to be used by users who didn't understand the tools. And they took on a life of their own. What's so interesting is if you read statistical methods in various books and different disciplines, almost unrecognizable. The way economists use statistics, very different way than many clinical scientists. Clinical science is different than basic sciences. Ecology, different than economics. So what we have is how different disciplines have used these tools and built cultures around them. So they become separated from their foundations.

Speaker 8:

And now we'll take a step back to look again at the wider picture of today's science culture. One of the reasons things have evolved this way may have something to do with how the scientific enterprise has come to function like a business. Dr. Yuri Lazebnik has written and presented on this idea. And we reached out to him for his thoughts.

Yuri Lazebnik:

I looked up and I realized that people start to using science to this word, workforce, right? And it didn't associate with me with something healthy, at least in regards to science. So I start to investigate where it's coming from. I mean, the environment basically is that the environment of what I can call it a pseudo market. Eventually this is what I realized by reading and following the... It's just what happened at some point, the end result is that science was assigned a market model, right? So the idea is that we are either sellers or buyers. Both, we can be both. We are both, but depending on the circumstances. And this is

why you can hear, how you should sell yourself, how you should sell a paper, how you need to sell an idea, and so on and so forth. And what's happening then, once the people in this situation, they start to play this game. And this game is linked to your survival or your parent survival, or you need to feed your family, even if you can kind of forget about yourself. Then people start just follow those artificial rules.

Speaker 8:

He discussed the differences in behavior between someone who finds purpose in science, so purpose to find answers, cures, or whatever it is they're trying to discover, and someone whose purpose has become selling their science in order to keep their job. As Dr. Lazebnik put it, you eventually shut down science to satisfy the model. You shut down science to survive whether you realize you're doing it or not. And unlike a real market, there's little feedback in this make believe business model upon which you can identify and change the things that aren't working. So where does all this leave us? Do we have an irredeemably broken system? We ask Christie Aschwanden because she's written on this very thing.

Aaron Carroll:

Do you think science is broken?

Christie Aschwanden:

Well, I wrote a whole story called Science Isn't Broken. I think the subhead of a something like that's a hell of a lot harder than we give it credit for. I don't think that science is broken. I think the culture around science and the way that science is practiced in many cases is broken.

Aaron Carroll:

So all is probably not lost. It's just hard work to weed out some of these major problems. In the final episode of this series, we'll talk about some potential solutions discussed in these many conversations with our experts. But first, we want to dedicate some time to particular hotspots, and we'll start with funding in our next episode. Hope to see you there.

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 Olsen. 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.