Transcript

Title: Prestigious journals struggle to reach even average reliability
Speaker: Professor Björn Brembs, Regensburg University

Introduction

Good afternoon, everyone. Thank you so much for coming. Welcome to the Jülich Open Science Speaker Series. Today, we will have Professor Björn Brembs. He will be talking about issues in publishing and impact factors and the metrics we use and the issues that are involved.

Professor Brems is a professor of neurogenetics at Regensburg University, and he is known for complaining a great deal about the status quo in science and for being an avid supporter of initiatives to change the way we practice and communicate science.

So, since this talk is not about his other scientific achievements, I will leave the introduction at that, and you have the floor.

Lecture

Thank you very much for having me. This talk is going to be on a very narrow topic, and that‘s only indirectly related to Open Science. And the indirect link comes through the centripetal forces that want to keep things and the inertia that wants to keep things

where they are and that tend to counteract any desires to reform science towards more Open Science. And that is about where we need to publish to get a job. And almost all, except for the first couple of few introductory slides, almost all of the slides will be data slides. Now, I‘m very comfortable and very used to looking at data all the time, but I gather you come from all kinds of different fields and so may not be as comfortable in quickly just looking at a graph and figuring out what the data means. So, I‘ll try to describe all the X- and Y-axis.

But while this may be technically kind of difficult, but try and interrupt me if I‘m going too fast, and then I can try and explain it again. But except for the first couple of slides, everything‘s going to be data. So, be aware of that, that it‘s not going to then in the end be again without data. It‘s all going to be data except for the first couple of slides.

(speaker starts sharing screen)

All of the data I‘m going to, almost all, nearly all of the data I‘m going to show, except for some data that came in after the paper was published, are in a paper with the same title, „Prestigious Journal Struggle to Reach Even Average Reliability“. And I put the link to that paper in the chat box just before the talk. And so, most of what I‘m going to show today in the central part is going to be in this paper. There‘s a couple of slides of more recent data and a couple of slides of what consequences it actually has that reliability is barely average in the prestigious journals.

But what is the whole point about prestigious journals and about prestige? And for that, we may, if we're classically minded, we may go back to the 15th century to a painting of a very famous Flemish painting of the 15th century, the „Adoration of the Lamb", which looks like this. It's a very typical of the time, very religiously themed painting, obviously. And if we fast forward to the 21st century, we have the „Adoration of the Glam", which you may think of like that, is that people who publish in Cell, Nature and Science and other glam journals are adored and are the ones that are getting promoted, funded, and promoted and hired.

Now, this is a very classic way of introducing. If you want to be more modern, you can pick Banksy, a Banksy painting here or a graffiti of people adoring the current commercialism, „The sale ends today" and everybody's looking at it and go, no, no, only today. And of course, you can do the same thing. You get the idea that this would be the more modern „Adoration of the glam journals" using a different style of introduction.

But the main idea here is that we have a sort of ranking in our journals, in the scholarly journals that we publish in. And one of the ways in which this rank is established, there's many different ways in which it is established, but one of them is „Clarivate Analytics Impact Factor", or „IF". And the IF has lots of issues, but what has to keep in mind, despite these issues that we're going to talk about right away, despite these issues, this impact factor number, and we'll see in a minute what that is and how it's supposedly calculated, it correlates very well with subjective journal prestige.

So, there's several studies over many years where people go to different fields and ask the experts, okay, here are 20 journals in your field, 50 journals in your field, can you rank them according to prestige? So, people who really know their fields, who know the literature, who should know where the good papers are published, they all say that the good papers are published in the journals that have the most prestige, and this prestige correlates very, very well with impact factor.

So, for the sake of argument now, I'm going to treat impact factor and prestige as near synonymous, such that usually, that doesn't hold all the time, but usually, by and large, the higher the impact factor, the higher the prestige of the journal, and the higher the expectation of the quality of the work being published there.

Now, there are three main problems with how the impact factor is calculated, and those three I'll share with you in a moment, the data behind those three issues, but the first issue is that the impact factor is, well, superficially calculated, but in fact, it's actually negotiated between whoever issues the IF right now is Clarivate Analytics, and the publishers that publish the journals.

It's whatever the numbers are, these numbers are not reproducible, if people try to do that, they get different numbers than those that are published, and even if it would be calculated, and even if it would be reproducible, then we would usually flunk fourth grade, or fourth term, second year science students, if they would be using that kind of calculation on that kind of data, so it's really very poor mathematics.

So, what's the evidence that the impact factor is negotiated? Actually, this is a well-known fact since the mid-1990s, and in order to find out how one can see that, one first has to have a rough idea of what Thomson Reuters, before the Clarivate Analytics bought the impact factor, Thomson Reuters was issuing it, how they calculated it, the calculation is the same after it was coined by Mr. Garfield in the 1960s, and then first produced in the 1970s, so the journal impact factor of, let's say, this year, when it comes out next year, the journal impact factor of this year would be that every single citation from the Thomson Reuters or Clarivate Index journals in 2023 to papers in the journal you want to look at, journal X, divided by the number of citable articles published in journal X in the two preceding years – sorry, there's a zero too many in here, when I fixed the dates – so this should, the idea is that this should give you the average number of citations for a journal, and so this is a visualization of this.

As you see here, it was introduced in the 1950s, but the first numbers were generated in the 1970s, took quite a while, this ISI is the „Institute for Scientific Information", and you take the number of citations, all of the citations to this journal, doesn't matter what they cite, and you divide it by the number of articles published in the two preceding years, and so if you're citing something that's only three years old, it's not counted, and then you take those citations, divide it by the number of articles, and then you get an average.

For example, if ISI counts 100 citations in year two, the journal published 60 articles, in year one it's 40 articles, you divide 100 by 100, an average citation of one, so this is how it's supposedly calculated, and now let's have a look why this calculation actually is rather superficial. The first, as I said, the first accounts of impact factors being negotiated come from the early 1990s, I was first made aware of this in this PLOS Medicine editorial from 2006, when PLOS Medicine went online, they also went to, back then it was still Thomson Reuters, and asked for their impact factor, and Thomson Reuters says, well, that depends on how many citable articles you have, and PLOS thought, okay, well, what is a citable article?

Well, it could be, you know, only research articles, or research articles and reviews, or news and views articles, editorials, or what, and depending on how you would count the denominator, PLOS Medicine would have started with an impact factor of two, or an effect factor of 11, and they settled on 8.4, and this is what they describe in this editorial, but not every journal is so forthcoming in describing how they arrive at their impact factor, and you see it's like a factor, you know, one order of magnitude, you know, from 2 to 11, is easy negotiable, so if you read any, if you read any impact factor of any journal, you know, add 5 or 10, or subtract 5 or 10, that's the usual range where you should be putting these journals.

But it's, it gets more interesting than that, so Current Biology, for instance, back then, when I checked this ages ago, this was still in the database, and I'm not sure whether the database today goes back that far, but you'll see some screenshots that I took when I checked out why Current Biology, the Cell Press journal that was bought by Elsevier in 2001, how that went in 2003, and remember, those are the two years that the impact factor is calculated over, right, it was bought in 2001, and then in 2003, the impact factor went up from 7 to 11, and so let's have a look how that happened.

So, this is the ISI, right, the ISI database entry for Current Biology for 2002 and 2003, and so, like this, right, and what we can see here is all kinds of information about Current Biology, and one of the things that we can see is the impact factor here went from 7 in 2002, and in 2003 to almost 12, so 11.91. How can that happen? Well, the number of items published dropped dramatically, while the number of citations doesn‘t, didn‘t really change all that much, so it‘s 7,000 here and 7,000 something here, but, and interestingly, this is for the overlapping year. In 2001, Current Biology once published 528 articles, and at the same time, in the same year, 300 articles.

That, for one, raises the question, how can you run a database where you can have the different numbers of articles published in the same year, but apart from that, you see how the impact factor was raised. Presumably, Elsevier went to Thomson Reuters, the ISI, and said, „Hey look, why don‘t you take away some of our articles that make our impact factor look bigger?“, and Thomson Reuters, at the time, because we don‘t know what the arguments were, said, „Sure, of course, we‘ll do that for you!“, and so the impact factor went from 7 to 12. That‘s how easy that works.

Now, you could think that, well, this is just one journal, and a biochemist checked a whole bunch of biochemistry, cellular biology, biochemistry journals, and looked for those missing papers, and the way you do that is you go to PubMed and check how many papers were published, and then you go to Thomson Reuters, or Institute for Scientific Information, and now Clarivate, and check what counts as what are the cited articles in the denominator, and what you see is that there‘s a lot of missing articles.

Only some journals have zero missing articles, and those are the ones, then, if you plot the calculated impact factor and the one that‘s published, that lie on this expected diagonal, whereas all the other ones that have missing articles, they all have increased impact factors compared to how you would calculate if all articles that are published would be counted as citable articles.

Now, you could say that, well, that‘s because these articles never get cited. Well, so this is one of those articles from, I think, 2000, where is it, 2014, right? These slides have been used for quite some years now, and this is a policy forum article that does not count for the denominator in the, for the impact factor, and then a few months later, I checked back 19 months later, I checked how many citations did it get, get 81 citations. So, those citations are not omitted from the impact factor calculations, they count, all citations count in the numerator, and so those 81, and in 19 months, count towards the numerator, but they‘re not divided by this article.

So, that‘s how you inflate your impact factor, simply by publishing all kinds of articles that get a lot of citations, but they don‘t count in the numerator. Now, well, this is how the publishers inflate their impact factor, how they negotiate the impact factor, but you may think that, well, at least once it‘s published, it‘s a reliable measure, you have a number, whatever that number is, it may be inflated, but you have a number.

Now, in 2007, the Rockefeller University Press asked Thomson Reuters to get their data, they bought the data, and they calculated, it's not a very complicated mathematics, right, it's a division, and they got wrong results, up to 90 percent different from the published impact factors. So, they contacted Thomson Reuters and said, hey, how can that be, and they said, oh, it's because you have your research database. Here is the database that we publish, or that we use the published impact factors on, and so they sent them a second database, and don't ask me why they have two databases, but they sent the second database, but even that didn't match. So, essentially, people are hired in academia and funded for a number that's not really calculated, and that number isn't even reproducible.

But now, let's assume that they weren't negotiated, that there were actually calculations being done, and that these calculations would be reproducible. Now, the third aspect is that these numbers are not mathematically sound, and what does that mean? As I explained before, it's a division, and you get the arithmetic mean. The arithmetic mean you actually only use when you have normal distributed data, such that you have a bell -shaped curve, where you have the mean in the middle, and then you have roughly, you know, the same count, the same number of data on the left and on the right side of it. Now, let's have a look if the citation data actually look that way.

So, what's plotted here on the left side is the citation rate. So, this is more than 30 citations per year, 14 to 15 citations per year, less than one citation per year, and the curves that you see are three different biochemical journals, and the number of articles that are cited, and how often these articles are cited, how many number of articles, and how often they are cited. So, there are many articles in these journals, 550, 350, 200 and something, that are cited less than one time per year, and very, very few articles are cited more than 30 times a year.

So, the average is dominated by the large values, and so it's also inflated. So, not only is the number inflated, the number itself, the mean itself is inflated, as opposed to a measure that one would teach fourth term, second year undergraduate science students, in how you would actually represent a left skewed distribution like that, namely with, for instance, a media, and on the right side is just then a consequence, one of the consequences of that, that's four different authors, and plotted for those four different authors is simply the impact factor of the journal, and how many articles, how many citations the articles actually then perceived.

And what you find is that it's essentially arbitrary, there is some correlation for some authors, and for others there's no correlation at all, which is no surprise essentially, and if you look at that in larger numbers, and not just four, what you find is that impact factor is not really predictive of how many citations you get.

So, just because you publish it in a high -impact journal, doesn't mean you will get many citations, simply because the relationship between citations and journal is really really poor, because every journal's impact factor is just dominated by a few highly cited papers, and not by the majority, by the 75 -80 percent of articles that are hardly cited.

Now, you may think this is just these three biochemistry journals here, so I have a range of journals that all show this left skewed distribution, and of course the glam mags are no different, they also have highly skewed, highly left skewed citations, you can see that here, that, you know, there's 20 articles that are cited very frequently, there's even a few articles that are cited extremely frequently, but a lot of articles that are cited hardly at all, and the same is true for whatever journal you look at. So, that means that where you publish does not tell you how many citations you're going to get.

Now, if these numbers are essentially made up, then why do we have this weird impression that all the top work is published in the top journals? Well, one explanation may be that this is an effect like astrology, right, you read your horoscope in the morning and you find you're the smartest person on the planet, you'll meet the love of your life today, and your workday is just going to be fantastic, and then when you get home you'll have the fantastic evening of your life, and then you read that and you go, well, this matches me perfectly, my days are all like that.

But what you don't do is you don't read the other 11 horoscopes, because what you find is that probably half of them, at least they perfectly fit you as well, and so what one needs to do is not just read the things that one is concerned with, and that one wants to read and what finds interesting, what one needs to do is apply the scientific method, look at as much data as possible, and then see does that data that I can look at, does that have anything to do with the journal rankings that we're using to hire and fire people, and this is what we're going to look at.

I'm going to start with my favorite diagram, and that is from a field that's between chemistry and biology, structural biology, it's essentially every little dot that you see on this plot is the publication of a computer model of a biochemically important molecule, and it's essentially the structure, right, you can look at where are the different atoms located, what is the distance between those atoms, which bonds are sitting where, so what does, how does that molecule look like. On the X-axis, we have different journals, and those journals are ordered according to the average of what is plotted here, so those black lines are the average, and what is plotted here, that's the Y-axis, is a quality measure.

Lots of factors go into this, because those are all computer models, this is something that you can calculate with an algorithm, and one of the things that goes into this algorithm is, for instance, if the distance between atoms matches the one that we know that it ought to be, so for instance, a carbon -carbon bond has to have a certain distance, and if you do the work properly to generate these computer models, then you should be very close to this distance, and many other factors go into this, and so you can calculate what would be the perfect model, and every model deviates from this perfect model, which is why the lowest number is the lowest deviation from perfect, and so lower is better in this quality measure, and what is done here is that all of these data are being averaged together, such that the average quality of the whole data set is zero, so this is what published here, so everything that's below zero is better than average, and everything that's above zero is better, is worse than average, and that's what is significantly, statistically significantly better than average, is published, is dyed in blue, and has an asterisk, and everything that's worse than average, significantly has an asterisk, and is red.

What you find on the right side, on the worst side, is Cell, Science, Nature, PNAS, somewhere use PNAS, also significantly worse, so what you find is that the higher the reputation, the higher the prestige of the journal, the worse the quality of this biochemical work.

Now, you might say, well, that's biochemistry, and so who cares about biochemistry? They may have reasons for why this is the case, maybe they do sloppy work, but on important molecules, and so you publish something that's really bad, but it's an important molecule, and you know, a rough structure is better than no structure, and so who cares about this?

So, let's have a look at other fields. This is a field on biological models, for biological disease models, so one of them, let's say, is on liver cancer, and what you do is you try to treat the animals, so you take a rat or a mouse with a liver, induced liver cancer, you divide them up in a treatment group and in a control group, and then you want to find whether your treatment that you've come up with is actually effective.

There's two important methodological aspects that you need to do when you do this kind of research. One is that you don't pick already the ones with the less severe liver cancer for the treatment group, and the ones that have really, really bad liver cancer for the control group, but you do this randomly, and the other thing is that the person that scores the liver after the treatment doesn't know whether that liver comes from the treatment branch or from the control branch, and so this is what is plotted here, so on the right, on the left side, is a measure of how much randomization was performed, so people that did this work, McLeod and colleagues, looked at the method section and searched for a mention of whether the two groups were randomized, or the assignment of the animals to the groups was randomized, and here it was checked whether the assessment in the end was done in blind.

What you find is, for one, that shockingly only 30% of the papers really adhere to this standard. One would expect this should be about 100%, and at least for the randomization effect, there's also a negative correlation such that the higher impact, here on the x -axis, the higher impact journals report less randomization than the lower impact journals. Now, you might say that this is just another field, and in this case, it's just method sections. Who cares? They're probably all randomized to the same extent. It's just that, you know, they were busy in the high -impact journals. They just failed to mention the method section, whereas, in fact, it was really done. It's not very easy to find out if that is the case, so let's go and check another field and see if it looks any different there.

So, the next field we're looking at is my own field. It's neuroscience. Every dot that you see here is an experiment, a neuroscience experiment. On the Y-axis, you see plotted the impact factor. On the X-axis, you see the impact factor plotted, excuse me, and on the Y-axis, you see statistical power, and to an approximation, statistical power essentially tells you whether you have the appropriate sample size in your experiment for your expected effect sizes and the variation in your data.

The general convention is that you design your experiments such that your sample size is so big that you reach a statistical power of about 80 percent, so if peer review would be working the way it should, they would reject peer reviewers, would reject everything that's not 80 percent power.

Now, as you can see, actually, the literature or this literature in neuroscience spans the entire range from almost zero statistical power to all the way to 100%, and what you can see here is that, you know, those people that published these articles clearly have a higher chance of getting a permanent job in science than the people who published this paper.

The other thing is that there's not really much of a correlation at all, so you have four high-impact publications in here, and then others of still pretty high impact, and what you see of those four, three do not reach the criterion for power, one does, so three quarters don't do, don't pass the criterion, and 25% do, and it doesn't really look much better in this range, so no, not really that the big name journals perform any better than any of the other journals.

Let's have a look at yet another field, this is molecular psychiatry. Every symbol that you see in this plot is a study that associates a single gene with a psychiatric disorder, so let's say gene X with smoking or gene Y with obesity. The size of the circle is the sample size, and what is on the X-axis is the impact factor is plotted, and on the Y-axis it's the odds ratio, the logarithmic of the odds ratio, and it's, again, it's normalized to zero.

The odds ratio here is essentially how big was the effect size of this single study compared to the actual effect size when one takes all the available studies together, and zero means that the study hit the effect size of this association between gene and disorder perfectly. What you find is, and everything above zero is an overestimation, means that the effect is, the reported effect in the study is larger than the actual effect of the meta -analysis, and what you find is two correlations.

One is plotted here, the higher the impact factor, the more you overestimate the effect size, and what you don't see is a correlation that the effect size, the sample size, decreases with impact factor, so essentially what this plot says is that in this field, if I may formulate it a little bit drastically, we test five people, we find gene X is associated with smoking to 25%, you cancel your study, so, oh, we have to send this to Nature right away, who would have thought that gene X is associated with smoking by 25%, and Nature says, wow, this is really surprising, they publish it.

Then later people study more and say "Oh, I can't believe that, let's have a look at this", and they use larger sample sizes, and then they find eventually it's only five percent association of this gene, which would still be a lot, but this is just a hypothetical example, right, so this is what this here shows, is that high -impact studies use a too small sample size, which leads to an overestimation of the effect size.

Now, again, could be something different, but this is an effect on ocean acidification, or this is a study that looks at different, again here, similar plots with impact factor and effect size, but from a different field, the impact of ocean acidification on fish behavior, and you should see similar things, the higher the impact factor, the higher the effect size, which, of course, would be very strange indeed.

So, let's stay with methodology, like how well was the work done, because this is, after all, this is what counts, right, if something is published in these high -impact journals that isn't true, then it doesn't matter how fantastic it would be, right, if I cure cancer in science, then, and I can't cure cancer in reality, it's not really worth anything.

So, the important aspect, of course, is how reliable is that work, and one of the ways in which you ensure that your work is reliable is that you disambiguate your resources, so which precisely, which antibody did I use, which cell line, which genetic construct, which fly line did I use, and for almost all of it, for almost all of these reagents and resources, there are numbers now that one can use, and one can put them, so they're called resource identifiers, and you can put them in your method section to make sure that people who want to reproduce your work know exactly which strain or which variant or which batch you used

What's plotted here in the upper left is, again, the impact factor, and then how often different resources were used, and what you would expect is that if the impact factor, if the impact factor had anything to do with quality and reproducibility, then you would have a lot of dots up here and very few dots down here, but what you find is it's all over the place, from 0 to 100%, and this is now all of them, this is antibodies, cell lines, constructs, and all of these are then plotted here, again, to show that not one of these things, not one of these things really correlates with impact factors. Also here, the prestigious journals don't seem to do anything different than any other journal.

Errors. Errors is another thing that one can look at, well, maybe they have better peer review, so they check for errors more consistently, the highly prestigious journals, so maybe their method section is not better, maybe their protein structure quality isn't better, but at least they fix all the errors, and one of the things that people have quantified is errors in Excel sheets, so people, let's say, in bioinformatics or other, in omics, in general genomics, proteomics, metabolomics, they often get gene lists that are associated, A with B, some gene is associated with something.

And then to get published, they put their output of their bioinformatics algorithms, they put it, copy and paste it into an Excel sheet, save it, and submit it with the manuscript. Now, the problem with that is that Excel tends to interpret different, the gene names as dates, for instance, so let's say the gene DEC1 gets interpreted by Excel as December 1, 1666, and then once that is done, you can't get back to which gene that was, and so people have gone and looked at how common are these errors in gene names, and what you find is that the highest occurrence of these errors was in Nature, the overall average, and that was over 30%, so a third of these articles had these issues, the overall average was a fifth, so 20%, and if you take the journals that were, are up here, then their impact factor is higher than the journals that are down here.

Again, it seems the higher the impact factor, the higher the error rate, so there's not really anything to their error checking either, it seems. Now, this particular effect is quite amusing, because when this was published in 2016, one would think that people pay attention to their Excel sheets before they submit them, and maybe fix these things by hand, but if you look at the publications with Excel gene lists, they just go up, more and more people submit Excel lists of their genes.

And the affected publications, the number of affected publications that actually have issues, also goes up, so people don‘t seem to really check more, and then, which means the proportion is still, you know, remains around 20-30% or something like that, so even after this was published, people didn‘t really pay any attention to that, they just copy and paste it into Excel and submit it.

That has led to a really weird effect, namely that the scientists that are working on this, they‘re renaming the human genes such that you take, you invent genes that Excel cannot misrepresent. Now, I‘ve recently heard that, I think now, just a week ago, I didn‘t have time to look up the reference. A week ago, I think Excel started fixing this such that scientists now can stop changing human gene names to adjust for Excel errors.

o, other errors, maybe that‘s just Excel people, whoever looks into a submitted Excel sheet, right? Everybody just publishes and nobody looks at it, so this is really completely irrelevant, so let‘s look at errors that are actually in the main paper. This is now from the field of cognitive neuroscience and psychology, and what people have done here is they‘ve looked at the p-value, so the significance value that is in the text, and compared it to the one when you take the data that is submitted with the publication, and they checked whether this p-value is actually, that‘s in the text, is the same that one calculates by using the data, and what they find is that it‘s actually not the same.

It actually deviates sometimes, and you could think it‘s just a typo, right? You read it, you either do a copy and paste, or you read it in your statistics program, and then you type it into the paper, and then there‘s a typo. Now, the problem is that most of those errors made p smaller, meaning more significant, or maybe from insignificant to significant. Sorry, from non -significant to significant. So, but let‘s leave that out, that weird asymmetry that these errors are not symmetrical, and let‘s just have a look.

Maybe these errors are caught in the high-impact journals that really, you know, have the good reviewers, but looking at the impact factor, and the number of records that are erroneous, the higher the impact factor, the higher the number of records, or if you just look at the papers, the percentage of papers that are affected, and note that this is not zero. Also here, no matter what measure you look at, more p-value errors, the higher the prestige of the journal. So, you see, you‘re getting the picture that really it‘s hard to find something that‘s actually good about the high-prestige journals. Another paper about questionable research practices.

So, when you submit clinical studies, you have to submit all the data, your plan of what your outcomes are going to be, what you‘re going to test, and so you can compare that with the published paper that then cites the study, and what you can look at is how often were variables added or dropped, data added or dropped, covariates added or dropped into this statistical evaluation. All of those are questionable research practices.

The scale was changed, and what people found was that when they compared the top eight journals and the non-top eight journals, they found that in the non-top journals, adding and dropping, so let's put it this way, adding and dropping variables became more common, adding and dropping covariates became more common in the top percent research journals, and changing the scale, well, changing the scale was not that dramatic, but it also increased by 8.7%.

So, essentially, what that means is, no matter how you look at it, errors, quality, methodology, questionable research practices, the high-impact journals, the prestigious journals that we use to hire and fire people, they attract the most unreliable research. So now you don't need to be a biology major or an evolutionary biologist to understand that if for 30 years we have recruited and hired and promoted and funded people who publish in these journals, then we have tendentially hired and funded and recruited people who publish unreliable research.

Okay, so now let's have a look what the consequences towards the end, let's have a look what the consequences are of hiring people who produce unreliable research. How reliable is our research? Let's just have a look at reproducibility projects. The first one published now over 10 years ago -- we'll skip this for the sake of time, we'll skip also this for the sake of time -- the first one of these reproducibility projects was one in psychology where they picked the Open Science Collaboration, picked a hundred prominent psychology experiments and replicated them.

And what you see here is a plot of the p-value. Again, the p-value should be below 0.05 for it to be significant, and most of the published studies showed effects that were significant, except for those three, so 97 were significant. The replications, the large majority of them, was not significant anymore. And if you looked at the effect size, all of them had, in the original studies, had positive effect sizes, whereas here a quarter of them actually reversed the direction of the effect, and the overall effect sizes were much, much smaller.

So, that was quite damaging with, in total, if you collect, wait, no, this is a very subtle way of showing it, if you just classify it into replicated or not replicated, not even half, not even half of the psychology studies actually replicated. Now, it's tricky to know what is a good, what kind of replicability should we expect. This is hard to put a number on that. For me, personally, if only half of my work would be reproducible, I would really consider taking a different job, I think.

So, that being said, let's have a look at other fields. This is social sciences – oh, sorry, this is still in German – social sciences, and you see this is all science and nature articles that look at, and then less than one means the effect is probably not there, larger than one means, yes, probably the effect is there. Those are the different studies, in this case, a little bit more than half replicated successfully in these articles.

Let's look at the economics. There, also, about 60 % were replicated. You can see it like this, replicated with p .05. Roughly, it depends on what you want to see. Prediction market beliefs are lower, original within 95%. So, conservatively, a little bit over half in economics.

And now, for me, as a biologist, most closely, this is my last slide, most close to my heart is a reproducibility project that just concluded this year, the reproducibility project cancer research, where they designed the project to replicate 193 experiments. And they asked the authors to help. They asked them to clarify the protocol, asked them whether they could share reagents, could share codes, report more details than in the paper on the analysis, whether the data was shared. You can see here the percentages.

In the end, of those targeted 193 experiments, they only could start 87, right? So, not even half. So, half of cancer research can't, doesn't even make it all the way to the bench to try to replicate it because something's missing, something essential is missing. So, from that, not even half. They then narrowed it down to 50 experiments that they actually were able to complete, a very, very low number. And it gets even worse. So, from those 50 experiments, they come up with 23 papers. They looked at 158 effects, all outcomes, and there were some, some were not included for reasons I don't, I didn't really understand in the paper.

That's why I show you the plot and the citation and a quote from the paper. Essentially, what you see here, the replication value wasn't larger than 0.05 or smaller. Those are the ones that are plotted here. So, you see the original effect size on the X-axis with dots and the replication effect size. Of course, a perfect replication would be along those lines. Actually, you see it's very much below that line. You see that the p-value, a lot of the p-values were very, effect sizes were large, whereas in the replication, you see very small, very small effect sizes and a very narrow band of significance. So, and some of them, even you see here, have a negative effect size in the replication, mimicking the effects of the psychology, of the psychology replication project.

So, in total, if you take now there, this quote, if you take the success rate out of those 50 experiments, 46% were successful. And you scale this up to the experiments of those 193. So, in, if this is representative of cancer research, if you're trying to replicate 193 experiments that you find in the literature, only 12% will replicate. Now, I would say cancer research has some significant societal impact.

And if this were reproducible, and if this were representative, and maybe even worse, if this would be representative for medical research overall, then essentially, it would mean that more than 80% of the money, of the public money that's being spent on medical research is wasted, because nobody can ever replicate the results that come out of that. Now, this is, in my case, in my opinion, probably one of the most extreme outcomes.

But this is one of the outcomes of what happens when you say, oh, we have to hire this person, they published in one of these high journals. If this is what you want, if 12% replication rate is what you want, for your institution, then you should encourage your institution to continue to hire professors, according to journal rank.

Thank you very much.