The Health Report

on ABC Radio National

The validity of published research findings

28 July 2008

Norman Swan talks to Professor John Ioannidis from the University of Ioannina in Greece and Tufts University in Boston about his paper which has the title: 'Why Most Published Research Findings Are False'.

Hide Transcript

Transcript

Norman Swan: A few weeks ago, one of you, quite clearly irritated by one of my stories and the implications being drawn from it, sent me a paper with the reassuring title "Why most published research findings are false". Now tempted as I was to ignore it since such subversion could leave this humble little program with no material to broadcast and put me out onto the streets, I thought what the heck, let's look the devil in the eye.

The author of the paper was Professor John Ioannidis who is Chairman of the Department of Hygiene and Epidemiology at the University of Ioannina in Greece and he's also at Tufts University School of Medicine in Boston. So I got him on the line to explain this falsehood stuff.

John Ioannidis: This is probably what one would expect based on what we are struggling to do in scientific work even with the best of intentions. You know scientific research is very difficult and discoveries that really make a huge impact are probably not very common. In some fields probably we are more lucky and more efficient than others but on average I would say that most of what we think we have found originally turns out to be false eventually when we do more studies.

Norman Swan: So our starting point here is somebody has published a paper which says they found something, you know a drug works, essentially they've made a finding, their hypothesis to start with was proven by their research and your suggestion is that's quite likely not to be true. So let's go through the argument.

John Ioannidis: It varies on what kind of a hypothesis has been made up front and it varies a lot depending on the field. For example some types of research findings are very likely to be true, even 90% or 95% likely to be true.

Norman Swan: Such as?

John Ioannidis: Others for example if we do a very large randomised trial for a medication or some medical intervention and we already have lots of research that has been done at the basic research level, preclinical level, early clinical studies and we have some fairly good notion that this is something that is worthwhile testing at a large scale in patients and see whether it does really work. And all the evidence has been pretty positive until then so -

Norman Swan: So it's not an idea out of the blue. One of the early ideas might have been out of the blue but by the time it has got to that stage -

John Ioannidis: Exactly, so once we get to the stage of going through a very large randomised trial to prove that this is something that's worth moving into clinical practice and we do get a statistically significant result then this is very likely to be true. If it's a well designed, large well powered, well conducted, well presented study, maybe it's 85% to 90% likely to be true. On the other hand if this is something that either has no specified hypothesis and it's mostly data dredging, just shifting through data, lots of data, with no clear sense that something really should be there and just testing lots of different hypotheses at the same time then, if we do get a significant result that's not very likely to be true. And if the number of hypotheses is huge then it's probably even less than 1 in 1000.

Norman Swan: So let's just sit with that for a moment, so people call that data dredging or fishing, they've got this vast number of variables and they can pick one out, they pick one out and say here is a significant factor. Now there are quite a lot of respected studies out there that would fit that bill, for example you're an Adjunct Professor in Boston, just up the road at Harvard the School of Public Health they have a huge study called the Nurses Health Study where they have been following 120,000 nurses for many years and made findings about hormone replacement therapy and various other things. Now some people have said that's data dredging, they've looked at all these variables and they come up with something that is positive would you say then that things like the US Nurses Study, the US Health Professional Study, which again they've been following men and women for many years, even perhaps the Framingham Study where they followed 5000 people, more than 5000 now for many years where they got lots of variables that they could be wrong?

John Ioannidis: Well these are excellent studies, I would say that these are exemplary studies and they have been designed and conducted by exemplary scientists.

Norman Swan: But they didn't have any hypothesis. I mean the Framingham might have -

John Ioannidis: Exactly, sometimes they did have some hypotheses. I think it would be an over-generalisation to say that they were just searching in the dark for associations. I think that each time that I have seen a paper from these teams and other prestigious teams there seems to be some sort of hypothesis but there's also a lot of an exploratory component, you know searching for new associations. There's nothing wrong with searching for associations and doing exploratory research, data dredging as you mentioned, this is part of the discovery process. However if something really comes to the surface just out of that data dredging process it is an interesting observation but before I can say more about it I would like to see it replicated in several other teams and with robust study designs. So this is OK as a first step but if someone finds for example that hormone replacement therapy can cut mortality in half, or that vitamin E supplementation can decrease deaths by heart attacks by 40% or 50% I would say OK, that's one observation that came out of a very nice epidemiological study - how likely is it to be true? Probably in the range of 10%.

Norman Swan: 10%?

John Ioannidis: Yes, or maybe 20% at the most and I would like to see more studies trying to confirm this. If other studies confirm this then that 10% or 20% goes up to something like 30%, 40%, 60%, 80% as we accumulate more data and all studies show the same thing. And at some point it will reach a level of certainty that I would say that every single woman should be on hormone replacement therapy after a certain age, or every person on earth should get vitamin E supplements. However, if we do more studies and that was the case for hormone replacement therapy and vitamin E we did even larger studies and even randomised controlled studies where we tried to test that and we saw that vitamin E had no benefit, apparently if you combine all the data it seems to have even some slight increase in mortality and hormone replacement therapy again, it has some benefits but the proposed benefits from the epidemiological studies on what we thought would be most essential were not really validated. And again the balance is mostly against using that intervention.

Norman Swan: And how often, you talk about bias and statistical bias in the reporting, to what extent do you see the statistics manipulated in order to get a positive result?

John Ioannidis: Well one does not necessarily need any manipulation. Let's say that someone does the perfect study, the perfect epidemiological study, the perfect exploratory analysis hunting for associations. However there are ten other teams that do equally perfect studies and only one is lucky just because of chance to find some particular association with some exposure or intervention of interest. Now if we had the benefit of reporting the results of all ten, or eleven investigations with equal weight and equally soon and in equal detail then we would not be misled, we would see that here are ten studies that find nothing, and there's one that's found something but if you pull them together you see that there's absolutely no effect, nothing to be seen, so it's just statistical rules that say if you run too many studies and too many analyses a few of them will show something that is just chance.

However in the current publication environment researchers are really urged to report that they have made discoveries, competition is very fierce, they have to say that we have found something and they probably don't have much time or even willingness to report and comment on what 'negative results', even though these studies may be just as important and as well conducted. So what we end up seeing many times is just the tip of the significant results that appear due to chance.

Norman Swan: Now you have several rules, you call them corollaries of this which you say drive this bias that positive results can in fact be false. Shall we just quickly run through these because size is important here?

John Ioannidis: One aspect is size, how big the study is and if we have a scientific field that has very small studies that are being conducted then if research findings emerge in that field they are less likely to be true compared to scientific fields that usually have much larger samples in doing their research.

Norman Swan: How do you know whether something is a small sample or not because you could do a study of 2000 people, is sounds big, but it's still not enough because you're not going to see the effect on 2000 people - you need 5000?

John Ioannidis: That's true, there are many effects that are very, very subtle and they would need many tens of thousands, if not millions of people to discover -

Norman Swan: Could you give me an example?

John Ioannidis: For example in genetic epidemiology in the last couple of years we have started making lots of very interesting discoveries by testing hundreds of thousands of genetic effects and most of the genetic effects that have come up, we are talking about relative risks in the range of 1.1 to 1.4.

Norman Swan: And just to translate that for the audience we're talking here about a10% or 40% increase risk which sounds a lot but in fact isn't.

John Ioannidis: It's a relative risk so it's not an absolute risk increase so let me give you another perspective of that. If someone has let's say 1% chance of getting a heart attack without having the bad gene, with the bad gene that 1% becomes 1.1%. So it's not that it becomes 1% plus 10%, 11%. It goes from 1% to 1.1% so we are reaching a point where sooner or later most of the associations that may be left to discover may be very,

very subtle effects.

Norman Swan: Because there are no big cholesterol or blood pressure type risk factors left?

John Ioannidis: If we have relative risks of 1.05 meaning the absolute risk of 1% going to 1.05% that's very, very tough to find even with extremely large studies. However, if we have very small studies one can prove that mathematically that what we get is either false signals or extremely inflated signals so it would get the wrong answer no matter what. This means that we need to push for larger good quality studies.

Norman Swan: Another rule is the one we did earlier which is that the more you look for the more you're likely to find something that's a false positive?

John Ioannidis: Yes I think that this is very important and it has many ramifications. One is how many tests one is doing and there are some fields that run hundreds of thousands of tests just by routine so you expect that most of the significant signals will just be chance and no more than chance.

Norman Swan: Can you give me an example of that?

John Ioannidis: Well much of epidemiology is like that. For example both genetic and nutritional epidemiology they can test very, very large numbers of items but probably nutritional epidemiology or lifestyle epidemiology where you can think of hundreds of nutrients and their interactions and composites in foods and so forth, it can be a real mess.

Norman Swan: So just by accident you find that mung beans might prevent colon cancer?

John Ioannidis: Exactly and that's every day in the news that some type of beans causes or prevents this type of cancer and the credibility of observations of this kind is very low.

Norman Swan: Now the fourth one is hard for the general public to analyse because they don't necessarily have the paper in front of them but it is about the design, that if they've been a bit slack in how they've designed it then again they are giving themselves latitude for a false finding.

John Ioannidis: Yes, in some fields researchers have decided that we have very strict rules on how to define what is a bad outcome for a patient, but there is some vibration also in the effect definition sometimes in some studies. Let me give you an example - death.

Norman Swan: Death?

John Ioannidis: I don't think that any one person can question that someone is dead.

Norman Swan: I don't know why I'm laughing here.

John Ioannidis: Even so, even for death there can be some variability in the definition. For example some studies may say let's take a look at whether that risk factor has any association with death - we find nothing. OK - how about disease related death, let's say that these are breast cancer patients, how about deaths that are related to breast cancer? This means that you have to decide whether a death was really related to breast cancer or it happened because of some other reason. This is a very difficult decision sometimes, I mean how did this woman die and then why? Or death from a particular toxicity, again it's a judgement call. So I gave you the example of death because this is very clear cut that there shouldn't be any ambiguity about whether somebody is dead or not. But once you go to more subtle issues like for example did the patient respond to treatment, if the treatment response is objective like for example headache, how do you measure the response of headache, or nausea and vomiting - how do you measure response if someone has nausea and vomiting. Then you can have lots of analyses using these different definitions and on top of that you can have lots of different statistical models that you can apply to each one of these definitions.

So even though it seems that one is doing one analysis, we may end up doing fifty analyses and again it's by chance.

Norman Swan: So you could have wobble on wobble?

John Ioannidis: Exactly and just by chance one out of these fifty may show something.

Norman Swan: Now the other one, the next one is financial and other interests. This sounds like being ideologically sound here, you know it's just wrong for the pharmaceutical industry to sponsor studies but you are actually saying that there is evidence that the person paying the piper does influence the tune.

John Ioannidis: Yes, I think that there are lots of empirical studies that have shown that repeatedly, studies that are supported by a specific commercial sponsor are about three times more likely to show statistically significant results in favour of the product of the sponsor compared to studies that don't have this kind of sponsorship.

Norman Swan: Is that a crude bias that they are just manipulating the data or -

John Ioannidis: Well I think it's a very subtle situation. I think we have no evidence that studies that are conducted by the industry are of worse quality compared to studies that have no sponsoring by the industry. Actually if anything the trend is in the opposite direction.

Norman Swan: So couldn't it just be what you mentioned earlier that, I'm not here to defend the pharmaceutical industry, but could it just be that they are doing large randomised trials of things that are well proven and therefore they are getting that 90% hit rate which is what you predicted at the beginning of our discussion?

John Ioannidis: This is one possibility however looking at a very large sample of such studies in different fields it seems that this is not really the main explanation. It may be that this is not involved at all. The most likely explanation is that when a sponsor is willing to sponsor a trial he wants to get a result that would be favourable out of it. Now you can design a study in a way that you can get a favourable result that would be marketable even though it would be clinically useless. There are many tricks that one could play with that. For example one could run a trial that has the same sample size as one that is not sponsored by the industry. However one may pick a different outcome instead of picking death, or clinical progression that would be important for patients you may pick up some biological marker.

Norman Swan: So one example of this for example is there was a recent controversy over an anti-Alzheimer's drug which had fancy testing there but possibly made no difference to either the person with dementia or their carers.

John Ioannidis: Exactly so if you want to prove that this makes a difference to the patient or to their family then you would get a negative result. But if you select to choose that by making that intervention or giving that drug you can change some pattern of the MRI picture or CT scan or PET or some weird biological thing, then you can get a significant result and it's not that difficult. So this is just one way that you can get a statistically significant result that is likely not to be very clinically relevant.

Norman Swan: And the final one you deal with is that the hotter a scientific field the less likely the research findings are to be true.

John Ioannidis: Yeah, this is a little bit of a paradox but it stems from the same reasoning that if you have a highly competitive field where people don't really tolerate negative results and results that are negative kind of disappear, never come to the surface, then you have a very strong selection force for people to select the very extreme findings that they get.

Norman Swan: Can you give me an example here?

John Ioannidis: Well I think that one can take lots of molecular medicine fields where there is high technology involved and lots of pressure to generate significant results. And even though we have for example like probably 50,000 research articles on gene expression profiling and their relationship to prognosis or prediction, so improving the forecasting of outcomes in patients with cancer and using these tests efficiently to select who to treat and who not to treat, out of 50,000 publications we have only two tests that have been approved for clinical use. And both of them apparently when we ran larger studies have very subtle benefits if any at all and randomised trials are still ongoing.

Norman Swan: So what can you do about this, how can you remedy the situation?

John Ioannidis: I think that is something that needs to be approached at many different levels; it's an issue of how we as researchers and scientists are doing our work and also what the society really expects of science and scientists. There's no doubt in my mind that science is extremely important and it's really the one objective way that we have to approach the world surrounding us and really make a difference, help people, especially by medical science improve health. It's something that is really useful. However I think that we are forced to make too many promises, we are forced to compete against domains of social life that make very strong claims for example politics, or religion, or sports, or you know you just name it, where people just make promises that they can show off and accomplish lots of things and probably they don't. And science is trying to catch up for various reasons, maybe not for just visibility but even just for sustaining itself. Unfortunately in most countries, even in countries that are wealthy, science gets a very small piece of the budget. So I think that if there was less pressure that scientist should really come out and say that we've made long chains of discoveries and we have a huge discovery every day, people would feel much better that they can really focus on exactly what we have found, what is its credibility, how do we move that forward, how do we refute that and some of that would survive to move to the next steps.

I think this is a very important aspect and something that probably a wide audience should be aware of.

Norman Swan: Each of these people that you've described, the 50,000 people that have done gene profiling studies in cancer for example to detect susceptibilities, or drug prescription appropriateness and so on and so forth, every single one of them would have got a grant from somebody either the National Institutes of Health, the Medical Research Council of the United Kingdom, the equivalent in Greece, the National Health and Medical Research Council here in Australia, so is it a problem of the grant giving bodies that they are just helping out their mates and allowing inadequate studies to go through to the keeper? Because they all complain that they are only allowing the best to go through, but if this huge percentage of studies are falsely positive, they have obviously not been rigorous enough?

John Ioannidis: Well I'm not sure that that would be the conclusion that I would make because going through these corollaries it's not that these studies are false because people have necessarily manipulated the data or done something that is fraud, I think that this is -

Norman Swan: No, no, no, that's not what I'm alleging, but their design is not rigorous enough and it's not being -

John Ioannidis: I would say that the design of a study has inherent limitations so I think it is important to recognise these limitations and to be just very straightforward about it and to say that with this kind of design, what we expect to get 80% of the time is not likely to be true. And people should be satisfied with that, if that's the type of design that they can afford and that's what they can only do. Obviously if they can run larger studies, better controlled studies with more strict quality control parameters, that's even better. But I think just because it's not possible to run trials of a million patients for each medical question of interest I think at a minimum we should just acknowledge what are the limitations of each study design that we adopt. And also describe what we find in the light of the limitations.

Norman Swan: We've been talking about biomedical science here and epidemiology but presumably what you've said applies to social research, psychological research that we hear about every day. So child care is damaging, child care is good for you, where presumably a lot of these definitional problems, design problems are actually inevitable because we are talking about something that's messy which is our existence in the social world.

John Ioannidis: I think that the basic concepts that we have been talking about apply to any type of empirical research where people make measurements and try to find associations, something is related to something else. It would not apply to scientific fields like pure mathematics where you know a equals b, b equals c, therefore a equals c, that would not be an issue but depending on where they operate different fields may have different problems. I think the take-home message is that scientific research is extremely exciting and we should just be as straightforward about what we have found and how credible it is.

Norman Swan: And just a final question then, the extent to which it has changed your research, in other words if you're actually walking the talk then presumably you're doing very few scientific studies because they are massive.

John Ioannidis: Well I wouldn't necessarily say so. I think that recognising the limitations is part of the research work. This doesn't mean that we should not do research, it means that we just have to be careful -

Norman Swan: To be up front?

John Ioannidis: To what we claim about our results. Results are interesting no matter if they are positive, negative, ambivalent, conclusive they are results, they are data, people have paid for that information to be collected and people have wasted lots of hours to get that information. It's a private and public treasure that we need to cherish.

Norman Swan: So maybe I'll just hold onto that resignation and just not put it in yet and maybe there'll be another Health Report next week you may or may not be pleased to hear. Professor John Ioannidis is Chairman of the Department of Hygiene and Epidemiology at the University of Ioannina in Greece and he's also at Tufts University, School of Medicine in Boston.

Reference:

John P.A. Ioannidis Why Most Published Research Findings Are False. *PLoS Medicine* August, 2005;2;8:696-700

Guests

Professor John Ioannidis
Chairman of the Department of Hygiene and Epidemiology University of Ioannina Greece and Tufts University Boston, Massachusetts USA

Further Information

PLoS Medicine

Presenter

Norman Swan

Producer

Brigitte Seega

Mon 8.30am
repeated **Mon 8pm**
Presented by **Norman Swan**

In This Program