Photography: Paul Stuart

Looking behind the numbers

When the clinical epidemiologist **John Ioannidis** published a paper entitled "Why most published research findings are false" in 2005, he made a lot of scientists very uncomfortable. The study was the result of 15 years' work cataloguing the factors that plague the interpretation of scientific results, such as the misuse of statistics or poor experimental design. Ioannidis tells **Jim Giles** why his conclusion is not as depressing as it appeared, and what he is doing to improve matters

You've been described as the "man who would prove all studies wrong". What was it like to find yourself in this role?
Overall, the reaction I got was very positive. People should not feel threatened by me: science is an evolutionary process, and contradiction and falsification are part of the game. We have tons of literature and a lot of it will eventually be refuted, but that is not bad news. If a small proportion is correct and survives then we will have progress.

How did you end up taking on the whole of science?
My parents were physicians and I trained in internal medicine. I wanted to deal with people and feel that I could improve their health, but I liked mathematics too. It was difficult for me to choose between the two. Then in 1993, I met Tom Chalmers and Joseph Lau. Tom was one of the first people to run a clinical trial and probably the first physician to combine the results of several studies in a meta-analysis – he and Joseph described cumulative meta-analysis in 1992. They introduced me to the idea of evidence-based medicine. That meeting had a great influence on me. It showed me how to inject robust quantitative thinking into clinical work.

Profile

John Ioannidis was born in New York City and raised in Athens, Greece. He studied medicine at the University of Athens, trained as a doctor at Harvard University and returned to Athens to earn a PhD in biopathology. He now holds a joint appointment at the University of Ioannina in Greece and Tufts University in Boston, Massachusetts.

How did you apply this quantitative approach?
Some of the early work I did looked at whether small studies give the same results as larger ones. After looking at hundreds of studies I started to ask: how often do the results of different studies agree with each other? My conclusion was that sometimes small studies disagreed with large ones beyond the level of chance. Early small studies generally tended to claim more dramatic results than subsequent larger studies. It's not just because scientists often oversell their results; it's also because small studies with negative results are often filed away and never published.

A lot of your work relies on complex statistical arguments. Can you explain them in simple terms?
In my 2005 paper "Why most published research findings are false" (*PLoS Medicine*, vol 2, p e124), I tried to model the probability of a research finding being true. By research finding I mean any association that is tested empirically. You can make some inferences based on how big the study is, since bigger studies tend to be more reliable. You also need to know what level of statistical significance a researcher is using when they claim a result. There are other things to try and compensate for, such as the fact that researchers seek out positive results to get further funding. These are the layers of complexity I tried to model.

"What's missing is whether the result is tentative or has high credibility"

What did your modelling reveal?
For some areas, if you get a positive result then it is 85 per cent or even 90 per cent likely to be true. That is the case for very large-scale randomised clinical trials with very clear

protocols, full reporting of results and a long chain of research supporting the finding beforehand. Then there is the other extreme, where an experiment is so poorly designed or so many analyses are performed that a statistically significant finding wouldn't have better than a 1-in-1000 chance of being true.

If most studies are wrong, how can science progress?
Things change as a field matures. In fields that generate data very quickly, you can get one study with an extreme result and then in less than a year you get another with the opposite result. The subsequent research falls somewhere in the middle. I think it works this way because an extreme finding sells in the literature. Once it's published, you get lots of competition in the field. Another group may happen to find something that is extreme in the opposite direction. I don't think it's fraud. We are talking about a sea of analyses. With one click on my computer I can find thousands of results. It's not that difficult for one team to contradict another very quickly.

With so much refutation going on, how can we know what to believe?
By keeping an open mind and trying to be cautious and critical. What's missing from a lot of papers is a sense of whether the result is tentative and unlikely to be true, or whether it has high credibility. This can be really important. For example, we need a lot of certainty about medical care before writing guidelines recommending a certain drug. But in some other fields, research with low credibility is extremely interesting. Most molecular science is highly exploratory and complex, with occasional hints of interesting associations.

How often have you come across high-profile oft-cited papers that later turn out to be wrong?
This is actually a common scenario. Some colleagues and I have looked at high-profile papers, with over 1000 citations each, that were later completely contradicted by large, well-conducted studies. One example is the finding that beta-carotene protects against cancer. It doesn't, but we found a sizeable component of literature where these original beliefs were still supported. It's hard to believe the researchers had never heard they had been refuted.

People aren't willing to abandon their hypothesis. If you spend 20 years on a specific line of thought and suddenly your universe collapses, it is very difficult to change jobs.

How are you trying to improve matters?
I'd like researchers to include credibility estimates in their papers. Many fields rely on a measure of statistical significance called a p-value. The problem is that the same p-value may have very different credibility depending on how the analysis was done and what results preceded it. If you take these into account, you can measure the credibility for a particular finding – the percentage chance that the largest, most perfect study possible would produce the same outcome. There is uncertainty also in the credibility, but I think we can say what ballpark we are in.

How should we promote the studies that produce more credible results, rather than those that are simply statistically significant?
There are several ways to do this. One: do larger, well-designed studies. Two: instead of having 10 teams of researchers, each working behind closed doors, investigators should collaborate and study the same questions. All the data should be made publicly available. If one team comes up with an interesting result then the whole consortium should try to replicate it. Much of the work I've been doing for the past 10 years has been about creating consortia to carry out research. The experience has been very positive. ●