Bayesian Critique of Statistics in Health: The Great Health Hoax

by Robert Matthews[1]

The plain fact is that 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding. It is time to pull the plug.

If you were going to have a heart attack, it seemed there was never a better time than the early 1990s. Leading medical journals were regularly reporting results from trials of new treatments for heart attacks that weren't just good - they were incredible.

In September 1992. the British Medical Journal published results from trials in Scotland of a clot-busting drug called anistreplase which suggested it could double the survival chances of a heart-attack victim. The following year another "miracle cure" emerged: injections of magnesium, studies suggested, could also halve death rates. Leading cardiologists hailed the injections as an "effective, safe, simple and inexpensive" treatment that could save the lives of thousands.

But then something odd began to happen. In 1995, The Lancet published the results of a huge international study of heart-attack survival rates among 58,000 patients – and the "amazing" life-saving abilities of magnesium injections had simply vanished. Anistreplase fared little better: the current view is that its real effectiveness is barely half that originally claimed.

Other "amazing" heart drugs seem to have suffered the same fate, also mysteriously losing their potency once on the wards. A study published last year in the BMJ compared death rates among cardiac patients in the early 1990s with those back in the early 1980s, in the Dark Ages before the advent of "clinically proven" heart-attack treatments. What Dr Nigel Brown and his colleagues at Queen's Medical Centre in Nottingham found was disconcerting: "Despite an increasing uptake of the 'proved' treat- ments, in-patient mortality ... did not change." The death rate in 1992 was the same as in 1982: 20 per cent - and double what had been found in the trials.

[1]Robert Matthews is Visiting Fellow in the Department of Computer Science, Aston University, Birmingham. and Science Correspondent of the Sunday Telegraph. A full account of the issues raised in this article. "Facts versus Factions: the use and abuse of subjectivity in scientific research" is available from the European Science and Environment Forum. 4 Church Lane. Barton, Cambridge CB3 BE.

Scientists have invented a long list of excuses to account for the disappointments. Some blame the fact that patients in clinical trials tend to-be hand-picked and fussed over by leading experts. Others argue that patients arrived on wards too late for the wonder-drugs to work. Or, perhaps the original trials simply hadn't been big enough. But there is another explanation, and its implications are so serious that scientists who understand them often refuse to go on the record to talk about them. They have been discussed behind closed doors by leading scientific journals and academic institutions - only to be swept under the carpet. Meanwhile, millions of pounds of taxpayers' money are being wasted following up illusory breakthroughs.

At the center of this scandal is a simple statistical technique used by scientists as the basis for supposed research breakthroughs. It is called 'significance testing'. And it is fatally flawed.

When used to analyze clinical trials, significance testing can easily double the apparent effectiveness of a new drug, and turn a borderline result into a "significant" breakthrough. It can throw up convincing - yet utterly spurious - evidence for links between diseases and any number of supposed causes.

But even more astonishing than these dangerous flaws is the fact that experts have been warning about the methodology for more than 30 years - but the scientific community has refused to act. In the meantime, a host of implausible "breakthroughs", which raise the hopes of patients and families, are announced with a fanfare ... and then quietly disappear.

Take another case. In 1994, Dr Michael Mendall and colleagues at St George's Hospital Medical School, London, made the claim that heart disease was linked to a bacterium in the human stomach. Quite how a bug in the stomach could possibly damage the heart was far from obvious. No matter: the standard textbook statistical tests pointed to a "significant" link. The existence of the link was confirmed by other studies - again using the standard statistical tests. But last November, a team at Bart's Hospital published the results of the biggest-ever investigation into the link. And it had simply vanished.

A similar story surrounds claims that emerged in the mid-1980s that aspirin could prevent pre-eclampsia, a potentially fatal condition that affects up to one in seven pregnant women. By the early 1990s, many studies seemed to confirm the theory, reporting a far lower rate of pre-eclampsia among women given low doses of aspirin. Once again, no one knew why it should work – the cause of pre-eclampsia is unknown – but the findings were still

"statistically significant". In 1994, however, the aspirin and pre-eclampsia link went the same way as claims for stomach bugs and heart disease: a major international study failed to find what the small studies had seen.

Health scares offer further evidence of the exaggerating effect of significance tests. Claims of a link between living near pylons and leukaemias in children have been made since the 1970s. The most impressive evidence emerged in 1992, when a team from the well-respected Karolinska institute of Stockholm found a "highly significant" increase of three- to four-fold in the risk of leukaemias. Again, quite why the risk should be so huge was unclear - and yet again, when the biggest-ever study into the claim was published last year, the link had evaporated.

Vitamin K injections and leukaemias; silicone breast implants and connective-tissue disease; salt and high blood pressure; passive smoking and lung cancer - the list of statistically based public "scares" goes on.

Just why has the scientific community failed to act? The answer lies in its squeamishness about subjectivity. It is hard to convey the strength of emotion aroused within the scientific community by the "S-word". Subjectivity is seen as the barbarian at the gates of science, the enemy of objective truth. So when, in the 1920s, the brilliant Cambridge mathematician and geneticist Ronald Aylmer Fisher came up with an apparently objective way of drawing conclusions from experiments, it was seized upon by the scientific community.

In 1925, Fisher published his techniques in a book. Statistical Methods for Research Workers. It has become one of the most influential texts in the history of science, and forms the foundation of virtually all the statistics now used by scientists. Its methods are taught today. On the face of it, Fisher had found techniques anyone – sceptic or advocate – could use to prove the significance of a new finding.

Critical to Fisher's method is the so-called P-value – defined as the chances of getting at least as impressive evidence as that actually seen if mere fluke were at work. These P-values are worked out mathematically from the raw experimentla data. According to Fisher, if the resulting P-value is below 0.05, then it is safe to label a finding "significant".

Combining simplicity and apparent objectivity, Fisher's P-value method was an immediate hit and its popularity endures to this day. Open any leading scientific journal and you will see the phrase, P less than 0.05 in papers on every conceivable area of research, form astronomy to zoology.

Indeed, national governments, including ours, still use Fisher's standard to decide whether a new life-saving drug should be approved for use.

But no sooner were P-values taken up throughout the scientific community than other statisticians began to ask some awkward questions. Most importantly, just how did Fisher know that his figure of 0.05 was a safe point at which to declare a result 'significant'? Incredibly, as Fisher himself admitted, he didn't know at all. He simply chose the figure of 0.05 because, he said, it was 'convenient'.

The implications are stark. It means that vital scientific questions – whether a new heart drug is seen as effective or whether breast implants trigger disease, for example – are being decided by an entirely arbitrary standard.

The first hints of this appalling flaw in Fisher's methods emerged as long ago as the early 1960s, following a resurgence of interest in a 200-year-old mathematical formula known as Bayes's Theorem. Put simply, Bayes discovered a mathematical recipe for working out how to update one's belief in a theory as new evidence emerges. It was a fundamental discovery, for Bayes's Theorem gave scientists a way of working out just how much more plausible their theories become as the data rolls in.

There was a problem, however: Bayes's Theorem revealed that before any new finding can be deemed 'significant', a crucial factor must be included: its plausibility. Non-scientists may hardly regard that as a problem at all: surely it makes sense to gauge just how plausible a result is before declaring it a breakthrough?

But within the scientific community, Bayes's Theorem acquired a reputation for being dangerously subversive. Its insistence on plausibility means that different people can reach different conclusions about the 'significance' of findings. And that has tainted Bayes's Theorem with that most repellent concept known to scientists: subjectivity.

Regardless of how awful they found Bayes's Theorem, however, scientists could not evade it. And its implications for Fisher's P-values turn out to be grave indeed. In the 1960s, at the University of Michigan, a team of statisticians including Prof Leonard Savage – one of the most distinguished experts on probability – showed that P-values were easily capable of boosting the apparent significance of implausible results by a factor of 10 or more. They went on to issue a warning that P-values were 'startlingly prone' to attribute significance to fluke results.

Despite being published in the prestigious Psychological Review, the warning went unheeded. Over the next 30 years, other statisticians also sounded the alarm bell, again without effect. During the 1980s, Prof James Berger of Purdue University – a world authority on Bayes's Theorem – published an entire series of papers alerting researchers to the "astonishing" tendency of the standard statistical tests to mislead. 'Significant evidence', Berger warned, 'can actually arise when the data provide very little or no evidence in favor of an effect.' The warning could not have been clearer. But again, it was ignored.

In 1986, one scientist decided to take direct action against the failings of Fisher's methods. Prof Kenneth Rothman, of the University of Massachusetts, editor of the respected American Journal of Public Health, made a bold stand and told all researchers wanting to publish in the journal that he would no longer accept results based on P-values.

It was a simple move that had a dramatic effect: the teaching in America's leading public health schools was transformed with statistics courses revised to train students in alternatives to Fisher's formula. But two years later, when Rothman stepped down from the editorship, his ban on P-values was dropped, and researchers went straight back to their bad old ways.

The story in Britain has been similar. In 1995, The British Psychological Society and its counterpart in America quietly set up a working party to consider introducing a ban on P-values in its journals. The following year, the working party was disbanded – having made no decision. 'The view was that it would cause too much upheaval for the journals.' said one senior figure. All research submitted would have to be vetted by a panel of statisticians.

Leading British medical journals have also considered taking decisive action – but they too have shied away. Instead, they are quietly trying to nudge researchers towards alternative ways of stating findings. The most popular are known as "confidence intervals" yet these too are known to suffer from similar flaws to P-values, exaggerating both the size of implausible effects and their significance.

Now, more than 30 years after the first warnings were sounded, it is clear that the scientific community has no intention of taking decisive action to tackle the critical flaws in Significance Testing. It has grown used to seeing its supposed breakthroughs come and go: the flaky claims of health risks from a host of implausible causes, the "wonder- drugs" that lose their amazing abilities outside clinical trials.

Taking action would mean "radical re-training", some scientists lamely argue. Curiously for a profession supposedly dedicated to discovering truths, the reliability of research findings is never mentioned. It is hard to avoid the conclusion that the real explanation for all the endless evasion is not scientific at all. It is simply that if scientists abandon significance tests like P-values, many of their claims would be seen for what they really are: meaningless aberrations on which taxpayers' money should never have been spent.

The plain fact is that 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding. It is time to pull the plug.