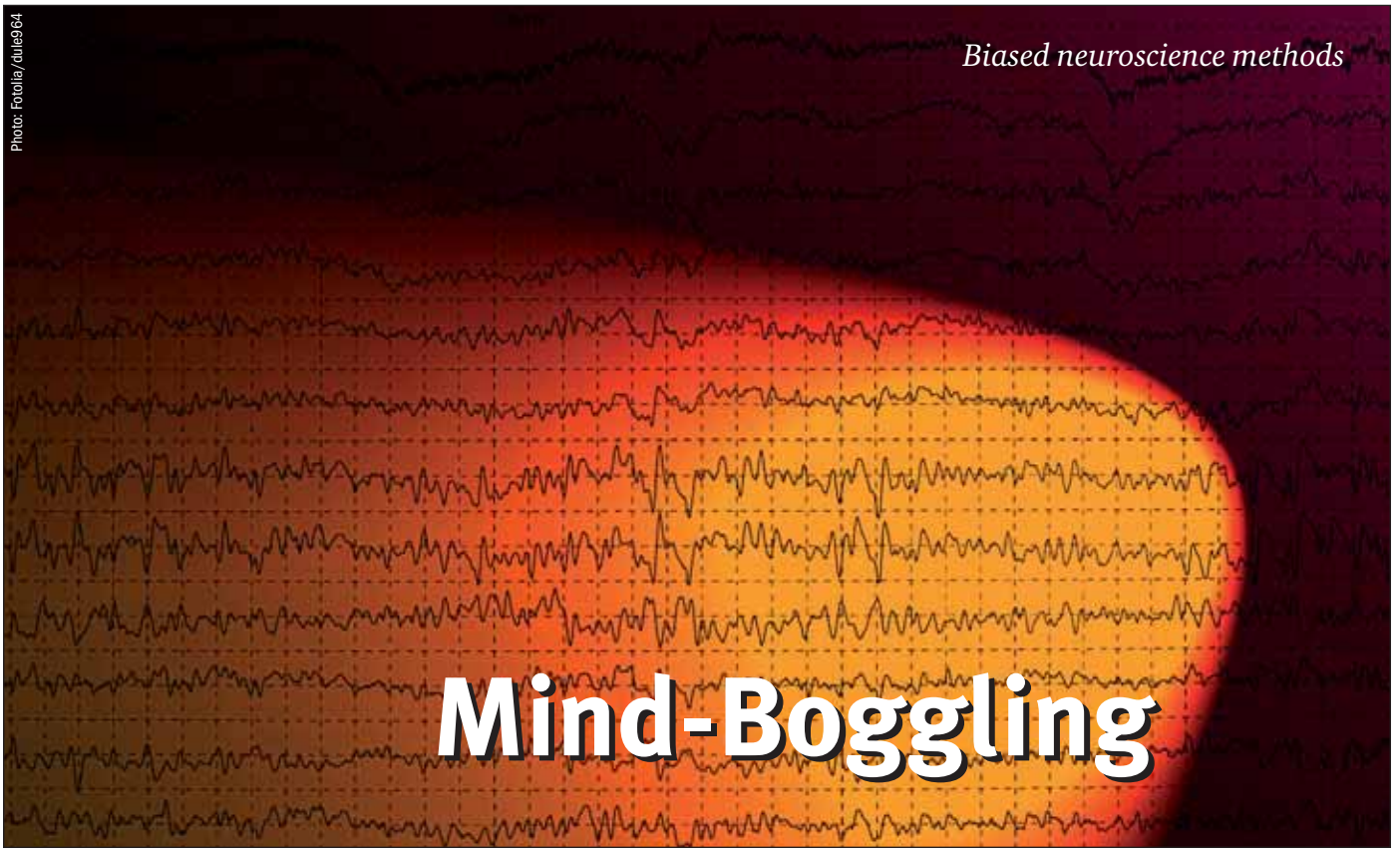

Mind-Boggling

Scientists impartially seek and disseminate truth – ideally. The reality looks different, as we all know. New looks at methods in neuroscience remind us of how bias can creep into our findings.

"Why Most Published Research Findings Are False" is the title of an article in *PLoS Medicine* in 2005 (2(8):e124). It caused quite a stir. While there is some critique about wording, scope or the misleading title for laypeople, the article's main argument is widely accepted by most doctors and scientists.

That statistical significance alone does not determine whether a finding is true or not was, in fact, more of a reminder than a discovery. The author, John Ioannidis, a professor of epidemiology at the University of Ioannina in Greece – highly skilled in mathematics – shows that the probabilities for research findings to be true ranged from an acceptable 85 percent in best case of well-conceived, randomised, controlled trials down to just 0.1 percent (!) for discovery-oriented, massive testing. Most types of performing studies would not reach the 50 percent threshold.

Thoughtless trials

The reason for these low probabilities are trials designed to confirm bias, unpublished failures to confirm a hypothesis and the low odds that the hypothesis is true, even before starting the study. The hotter the research field, the more likely the trials are done in a thoughtless fashion.

One such a hot field is brain research. It just fascinates most of us, how this roughly one-and-a-half kilogram mass in our skull enables our muscles to handle a pipette, remember what we learned at school or enjoy the conscious experience of being right here, right now.

Timing a potential

Indeed, recently a few studies about the function of the brain have been published that show how false positive findings creep into the scientific literature when electric currents are measured on the skull in electroencephalography (EEG), when the blood oxygen levels are measured in functional magnetic resonance imaging (fMRI) and even when technically simple psychology experiments are performed.

The invention of measuring small currents from the brain dates back as far as 1875. While it gives a rather crude measure with a poor spatial resolution, it is still the fastest method that allows to precisely time events in the brain. Event-related potentials (ERPs) are characteristic peaks that are used for timing.

The famous experiment by Benjamin Libet in 1983, for example, found out that an ERP could be measured in the motor cortex 300 milliseconds before participants reported feeling like taking the decision. While we will probably never reach agreement over the implications for the idea of free will that this experiment has, it remains a fascinating area of research.

David Acunzo, PhD student in neuroinformatics at the University of Edinburgh and colleagues wanted to measure the influence of emotion and attention on the earliest ERP detectable when something catches our eye – 40 to 100 milliseconds after the visual stimulus. There is a debate going on among electroencephalographers on whether emotions would influence this so called component 1 of the visual stimulus. "Some people state that emotion is processed particularly early in the brain. Component 1 modulation by emotional faces has been reported before but not clearly replicated," says Acunzo. He displayed some photographs of happy, neu-

tral and fearful faces to a few participants and measured changes in the electric potential at component 1. One reviewer of the article from Acunzo and colleagues, however, found that there may be a problem with the filters used to enhance the signal.

Tricky filters

As the electric waves are usually quite noisy and the measurements also drift away as a technical artefact. That is why filters are commonly used to remove the high and low frequencies from the signal. But, as Acunzo had to realise, filtering influences the data. And not just data alone. "By curiosity, we looked at other papers and their methods to compare, and found that a very large portion of them actually used wrong filter parameters. So we wrote a paper to point out this problem and to prevent people from repeatedly committing this very basic mistake. More than one-third of the published ERP results of the recent years may be wrong because of bad filters!" says Acunzo.

In fact, filters distort the signal: they shift peaks slightly. Some types of filters delay the signal while others make it appear early. Both make the correct timing of the ERP difficult. And particularly annoying, as Acunzo and colleagues explain in the article, "The early ERP component 1 is very sensitive to such effects." Acunzo recommends inspecting the data before filtering, only using filters when absolutely necessary and, if they have to be used, putting them at very low frequencies by default (*J Neurosc Meth* 2012, 209:212).

Checking on haemoglobin

These findings about filters also apply to other techniques like magnetoencephalography – the EEG magnetic corollary– and even to a technique as different as fMRI. In fact, fMRI relies heavily on filtering. The signal usually accounts for only ten percent of the signal measured – the rest is noise.

The fMRI technique measures how much oxygen is used by a particular brain region – more precisely, the portion of deoxygenated haemoglobin in the blood of that region. When a region is used more intensively to perform a task, the maximum signal is obtained about five seconds later. Compared to EEG the spatial resolution is very high, as each voxel (three dimensional pixel) can be as small as one millimetre cubed. The downside is that the time resolution is very low as recording the whole brain takes about two seconds to complete.

Catherine Davey, postdoc at the Defence Science and Technology Organisation in Melbourne, and Australian colleagues spotted a common flaw in the analysis of fMRI due to filtering. They put participants into a scanner and let them press either the left or the right button, according to constantly changing instructions on a screen. The participants had to do this for half-a-minute and then let their brain rest for another half-a-minute by looking at a cross on the screen.

Modifications – not filters!

Davey then analysed, which voxels were active due to the task performed. The images of the brain showing the significantly correlated voxel were lit up like a bonfire – roughly a third of the voxels were significantly correlated. A very unrealistic finding. Filtering the data made it even worse. In their mathematically very heavy article the Australian neuroscientists point out that the filtering introduced artificial correlation (autocorrelation) between the different recordings, violating the assumed independence for statistical analysis. They propose a correction method for unintentionally introduced variation (*Neuroimage* Epub, 24 August 2012).

This is not just a minor aesthetic problem but points out that, until now, fMRI analysis "inflated the false positive rate and artificially induced connectivity", as explained in the article. Davey says, "The effect of filtering is enormous. Particularly if the number of frequencies you remove is high. It's impossible for me to say if the effects people have included so far would pass the more stringent but more correct test required. Perhaps their results would just be more focussed on a specific region?" Davey and colleagues also deplore the absence of standards in the filtering and sampling rate, which "impedes comparison of results across different experimental configurations".

A blogger with the pseudonym Neuroskeptic even considers filtering a misleading term. "It implies that all you are doing is removing the unwanted noise, leaving pristine, crystal clear data. In fact mathematical filters can put stuff into the data as well as take it out. So should we stop using that word and just call them what they are: modifications?"

35,000 ways to Paradise

But the errors are not just due to filters. The bare choice of which filters, corrections and statistics causes false positives to increase tremendously. If one way of processing the data does not show the expected power, it is perfectly acceptable to switch to a better technique. But exactly the practice of switching analysis method when there is no signal but keeping the method when there is a nice signal, overall distorts published results.

Joshua Carp, a PhD student in neuroscience with the department of psychology at the University of Michigan, Ann Arbor (USA) made a very thorough study of the most common analysis pipelines used by fMRI experts. This allowed him to analyse the data in almost 35,000 different ways. And still, Carp wrote that "the approach used likely underestimated the true

Photo: Fotolia/guy

Those are definitely the wrong filters for electroencephalography.

flexibility of fMRI analysis methods". The neuroscientist took a freely-available dataset of thirteen participants that were put into the scanner and were asked to push and hold a button when a start signal appeared and to let go of the button when a stop signal appeared. A comparison of the two signals showed, which parts of the brain were more active during the stop task than during the start task.

The data is usually pre-processed to correct for the equipment artefacts, like unusually high spikes caused by magnet artefacts or slice-timing correction that accounts for the different timing as the brain slices are recorded one after the other. Equally, model artefacts are corrected, like the different shape of each individual's brain, the movement of the head or, as in the Australian Davey's case, the suspected autocorrelation of their recordings. In addition, there are different statistical ways of determining whether the threshold has been exceeded or not. Each method can be run with different parameters.

Results à la carte

In the 35,000 analysed conditions, nearly each voxel was activated at least once. Carp writes, "A sufficiently persistent researcher determined to find significant activation in virtually any brain region is quite likely to succeed." Because there are 32 different software packages and no default parameters, any fMRI researcher can reach practically whatever conclusion he or she likes. Indeed, sometimes false positive findings are estimated to reach over 50 percent.

Carp suggests amending this extremely unsatisfactory situation by optimising analysis pipelines. In his study, the PhD student did not look at, which of the methods would be the best. He instead points to a tool for finding the optimal and, therefore, correct analysis strategy developed by the biophysicist Stephen Strother from the University of Toronto. "But for reasons I don't understand, these tools don't seem to be used very often," deplores Carp. On the other hand, Carp is convinced that "tools for sharing analysis code have improved very quickly over the past few years, so this practice may grow more popular".

False positives psychology

How many analysis pipelines one might try out to reach a conclusion is not only important in high-tech science, such as fMRI, but in reality applies to every field of science, such as psychology. A fantastic demonstration of this was unintentionally provided by Daryl Bem, a psychology professor at Cornell University. Bem was able to show the existence of the psi phenomenon of feeling the future, like knowing in advance, on which side of the screen an erotic picture would appear or remembering numbers before learning them – no joke! (*J Pers Soc Psychol* 2011, 100:407). The psychology postdoc Tal Yarkoni from the University of Colorado, Boulder commented on this in his blog *[citation needed]*: "When you combine data peeking, liberal thresholds, study recombination, flexible hypotheses and selective measures, you have a perfect recipe for spurious results."

How easy it is to produce statistically significant but completely nonsensical results is beautifully shown by Joseph Simmons, associate professor in psychology at the Wharton University of Pennsylvania, who studies consumer behaviour. Together with two other US-American psychologists he shows in an entertaining article that whilst listening to the Beatles song "When I'm Sixty-Four", the study participants felt, on average, a year-and-a-half younger than when they listened to a neutral song.

They were able to influence the age of the participants by allowing a high "researchers degree of freedom". For example, by not defining when the data collection would be terminated, allowing a peek at the data and stopping whenever a significant result was reached. They also used different variables (songs) but only published the result of one. Apart from the father's age, they also used many other con-

Too much freedom in the experimental set-up can lead to wrong results.

Photo: Fotolia / XtravaganT

trol variables but only published the one that gave a significant result (*Psychol Sci* 2011, 22:1359).

Raising awareness

So again: making use of your degrees of freedom in conducting the experiment and in choosing what to publish, you can bring any claim you want to statistical significance. You may prove whatever you want! Simmons and colleagues showed by simulation that for the traditional significance level of five percent and combining four different researchers' freedoms, you could reach likelihoods for obtaining false positives of over 60 percent.

In the article the authors state that "in conversations with colleagues, we have learned that many believe this practice exerts no more than trivial influence on false-positives rates". And Simmons adds, "We conducted simulations and realised that the effects of using those freedoms are bigger than we suspected. Most scientists in our field knew that these practices were not strictly kosher from a statistical point of view but I do not think almost any of them realised, just how powerful all of this flexibility could be. I personally did not realise this and I had engaged in some of these practices in the past."

Simmons is indeed in good company. Also, the EEG scientist Acunzo only realised how filtering can affect the results after a reviewer pointed out the problems to him. And the problems are not restricted to brain and behaviour.

In his new book "Bad Pharma", the British psychiatrist and science writer Ben Goldacre convincingly tells the story of broken pharma, despite the control of professional regulators (see book review on p. 61).

What can be done?

The scientific process does not live up to expectations and usually publication pressure is blamed for the deplorable state of science. While part of science is busy correcting itself, not everyone welcomes the critique by fellow scientists. Some see it as a possibility for pseudoscience (psi, homoeopathy and the like) to claim that science is wrong. Ioannidis dismissed this critique as he told the US-magazine *The Atlantic*, "If we don't tell the public about these problems, then we're no better than non-scientists who falsely claim they can heal."

Most of the comments Simmons received on his article were positive. Some less so: "The negative comments mistook our article as an attempt to discredit psychology. We are worried that psychology is already being discredited, and our article was an attempt to change all that." Who needs to fix it? The fMRI neuroscientist Carp thinks, "It is not productive to blame any particular group for flexibility in research practices. But I do think that granting agencies have the most power to improve these practices."

All the scientists mentioned, so far, recommend one thing in common. The details of experimental design, the used filters and the whole analysis pipeline have to be described thoroughly in the publication, including critical decisions. Simmons asks authors to declare that they have *only* used so many participants, *only* done a particular analysis and *only* used a particular statistics test. If the authors deviate from this rule, the decision has to be justified. They also agree with Carp that standardised reporting procedures for their respective fields would be a good idea. Both Simmons and Ioannidis also urge fellow scientists to value replicating experiments and publish whatever the results are. Simmons and colleagues write, "Our goal as scientists is to publish as many articles as we can, but to discover and disseminate truth."

The psychologists warn of non-solutions: like adjusting the alpha levels for statistical significance, use of Bayesian statistics. Both have their virtues but make the problem worse as they introduce more degrees of freedom for the researchers. Furthermore, open data is a good thing but does not solve the problem of secretly dropping data.

Self-serving interpretations

Most of this has been known for ages. Why has it not been changed? The author, Ben Goldacre, accuses journal editors, for example, for having broken their promise of only publishing pre-registered clinical trials. Simmons thinks that earlier warnings on the abuse of the researchers' freedom "had mostly been ignored, probably because they were too statistical and abstract, or offered solutions that were too impractical".

Simmons and co-authors give scientists the benefit of the doubt, "This is not driven by a willingness to deceive but by the self-serving interpretation of ambiguity, which enables us to convince ourselves that whichever decision produced the most publishable outcome, must have also been the most appropriate."

Florian Fisch