(20) Neuromythology – What Happens When You Violate Statistical Premises

Description

A moment ago (this was first published in 2016), probably the biggest publicity bomb I have seen in a long time exploded: A group of Swedish authors, together with an English statistician, have published a huge simulation study. It shows that possibly up to 70% or more of the total of more than 40,000 published neuroscience studies that have used functional magnetic resonance spectroscopy (fMRI) have produced useless results and therefore actually belong to be cleaned out or replicated [1].

This seems to me to be one of the biggest scientific collective scandals of recent times. And one can learn a lot about statistics from it. But in order.

Before we turn to this study: This is not to say that MRI methodology is wrong and that so-called structural imaging methods are useless. It is solely about statements about spatial spread of activity in functional magnetic imaging. But even that is a huge chunk. Follow me.

What happened?

Functional magnetic resonance spectroscopy or imaging (fMRI) is very popular as a research method. The technique is based on the fact that hydrogen atoms – which are found everywhere – can be aligned via strong external magnetic fields. By simultaneously applying and scanning electromagnetic high-frequency waves, the atoms can be localized. Depending on which frequency is chosen, one can also make different types of structures or molecules visible. This can be used, for example, to determine the difference between blood whose red blood cells are saturated with oxygen and that which has given up its oxygen.

This so-called BOLD signal, short for “blood oxygenation level dependent signal”, can be used to deduce how high the metabolic activity is in a certain area of the body, e.g. in an area of the brain. An increase indicates increased oxygen consumption, increased blood supply, increased metabolism and thus increased activity in an area of the brain. A decrease indicates the opposite.

Now, in order to see anything at all in a functional magnetic resonance imaging (i.e., imaging) study, one must of course create differences between experimental and control conditions. This is usually done by having people in the MRI tube do different tasks in a specific sequence, called blocks. For example, they have to read a text on a screen, or think of something specific, or recite a memorized poem in their mind; in another block lie down and relax instead. This happens in fixed sequences. Thus, one can compare the sequences in which something defined happens in the mind with those in which calmness reigns.

The difference in the signals is then used to calculate the difference in the activation levels of the two conditions in specific areas of the brain and to make deductions about which areas of the brain are responsible for which functions. In addition, such conditions are often compared with situations in which control subjects are only measured (“scanned” is the neuro jargon) without anything happening.

To be clear, let me add: One can also use the method to visualize anatomical structures or to record the functionality of connections within the brain. These two applications are not covered by the study discussed here,

but only the activation of brain areas as a result of activity change due to experimental instruction.

Now the signals that arise from the measurement, it is easy to imagine even as a layman, have to run through a series of complex mathematical and statistical procedures before the pretty colourful pictures we admire in the publications and glossy brochures emerge at the end. In which experts then explain that the brain "lights up" when a person does this or that. This "lighting up" refers to the false colour representation of the increase or decrease of the BOLD signal in certain areas, which has been statistically isolated as a significant effect from the background noise. It is this statistical filtering procedure that then leads to the colouring – which is, after all, nothing more than the pictorial implementation of statistically significant signal detection – that was examined in this publication and found to be unreliable in the vast majority of cases. Why?

This statistical filtering procedure is unreliable – why?

Signal detection in an fMRI study is essentially a two-step process. The first step is to pick up the raw signals from the pulsed application of the magnetic fields and their deactivation, and to sample them with a high-frequency electromagnetic field. This provides the raw data about changes in the activity of the blood supply in the brain, i.e. about the oxygen saturation of the blood and the change in the distribution of the blood in the brain. Of course, as you can see immediately, this results in millions of data points that are determined in rapid succession and which, as such, are not usable in raw form.

The second and crucial step is now the statistical discovery and summarization procedure. This is done by analysing the raw data with special programmes. The study discussed here examined the three most popular programmes. In order to understand how complex the whole thing is, one has to imagine that the fMRI signals are initially picked up at different points on the surface of the head and also originate from differently deep areas of the brain. We are therefore dealing with three-dimensional data points, which, analogously to the two-dimensional data points of a screen, where they are now known to all as "pixels", are called voxels. Voxels are therefore three-dimensional pixels that originate from a defined location and vary in intensity. Since voxels cover just 1 cubic millimetre, the image that would emerge would be extremely confusing if one had to analyse them all individually.

For this reason, one usually groups the voxels into larger areas. This is done by making assumptions about how the activity of neighbouring points relate to each other when a larger functional brain area, say the language centre in generating mental monologue, is activated. This happens via so-called autocorrelation functions of a spatial nature. We are all familiar with autocorrelation functions of a temporal nature: If the weather is very nice today, the probability that it will also be very nice tomorrow is higher than if it has already been nice for two weeks. Because then the probability that tomorrow will be worse is gradually higher, and vice versa.

Analogous to such a temporal autocorrelation, one can also imagine a spatial one: Depending on how high the activity is at a point in the voxel universe, the probability that a neighbouring voxel belongs to a functional unit will be higher or lower. In the early days of programme development for the analysis of such data, relatively little information was available. So a reasonable, but as it now turns out wrong, assumption was made: namely, that the spatial autocorrelation function behaves as a spatially propagating Gaussian curve or normal distribution.

Control data

Now there are thousands of data sets of people measured by MRI scanners for control purposes, so to speak, without any tasks, and thanks to the possibility of open platforms, these data are made openly available to scientists. Anyone can download it and make analyses with it. Taking advantage of this opportunity, the scientists have recalculated data from nearly 500 healthy people from different regions of the world, measured in a scanner without any task, using simulated analysis methods by applying the three most popular analysis software

packages to them.

In total, they tested 192 combinations of possible settings in more than 3 million simulation calculations. So, somewhat simplistically, the scientists have pretended that the data from these 500 people came from real fMRI experiments with on and off blocks of specific tasks or questions. But it is clear that this was not the case because the data was control data.

One would expect in such a procedure that a certain number of false positives would always be found, i.e. results where the statistics say: “Hurrah, we have found a significant effect”, but where in fact there is no effect. This so-called error of the first kind or alpha error is controlled by the nominal significance level, which can be set by convention and which is often 5% (p = 0.05), but in the case of fMRI studies is often set lower from the outset, namely at 1% (p = 0.01) or 0.1% (p = 0.001). This is because this alpha error indicates how often we make a mistake when we claim an effect, although there is none. At 5% level of alpha error, we make such an error 5 times out of 100. At a 1% level of alpha error in one out of 100 cases. And at a 1 per thousand level in one out of 1,000 cases.

Now, of course, if we apply many statistical tests in parallel to the same set of data, this error multiplies because we get the same probability of making a mistake again in each test if we make a factual claim that is not in fact true. The nominal probability of error of p = 0.05, i.e. 5%, then becomes the probability of error of p = 0.1 or 10% for two simultaneous tests We therefore make twice as many errors. So, in order to comply with the nominal probability of 5%, the individual probabilities must be set to p = 0.025 for two simultaneous tests on the same data set, so that the joint error probability p = 0.05 remains. This is called “correction for multiple testing”.

Because a large number of tests are made at once in the fMRI evaluation packages, one sets the detection threshold there for what one is prepared to accept as a significant signal right from the start at p = 0.01 (i.e. an error probability adjusted for 5 simultaneous tests) or even at p = 0.001. This is an error probability adjusted for 50 simultaneous tests and thus adheres to the nominal error level of 5% for 50 tests. This correction is already built into the software packages studied; thus, the problem found is not related to it.

All these parameter settings were used in the study conducted here. At the same time, scenarios were run that are common practice in the real world of fMRI research, i.e. that one takes, for example, 8 mm clusters and merges the neighbouring voxels with a detection threshold of p = 0.001, which seems to be quite reasonable.

Then, in complex simulation calculations, all sorts of putative experimental comparisons were superimposed on this control data, and it was documented how often the various software packages make significant “discoveries”, even though it is known that there are no signals hidden in the data at all.

When clusters are formed, i.e. voxels are combined into larger areas, false positives, i.e. signals where there are none, are found in up to 50% of analyses. Or to put it another way: some software packages detect signals with a 50% probability of error where there are no signals at all. Put another way, in 50 out of 100 studies, the analysis says “there is a significant effect here” where there is no effect at all.

When the cluster size is smaller and the threshold for combining voxels into clusters is higher, the probability of error approaches the 5% nominal significance threshold. For voxel-based analysis, i.e. when one makes no assumptions about the correlation of voxel activities and accepts that one has to interpret a chaotic image of many voxels, the analysis remains close to the error probability of 5% for almost all software packages.

And for the so-called non-parametric method, i.e. a statistical analysis based on a simulation calculation in which the probability is not derived from an underlying and assumed distribution, but from an actual simulation calculation based on the available data, the nominal significance values are always preserved.

The problem is, however: The software packages are used because one does not want to do a laborious interpretation of a voxel-based evaluation oneself, but delegate it to the computer, and because one does not want to carry out weeks of simulation calculations to determine the true probability. In addition, signal noise or artefacts, such as those caused by movements, would be too much of a factor in a voxel-based evaluation. So one tries to find supposedly more robust quantities, precisely those clusters, which one then tests.

For a very common scenario, the 8 mm clusters described above with an apparently conservative detection threshold of p = 0.001 from voxel to voxel before one is inclined to consider a cluster "significantly activated" or "significantly inactivated", the values look grim: the error frequency rises up to 90% depending on the program, and **a 70% error probability across the literature is a robust estimate**.

Only a non-parametric simulation statistic would not make excessive errors here, either. However, this one is almost non-existent. Incidentally, the same problem was also found for active data from real studies. Here, too, a so-called inflation of the alpha error or a far too frequent detection of effects where there are none at all has been demonstrated.

Where does the problem come from?

You can use this example to study the importance of preconditions for the validity of statistics. First, the software packages and the users make assumptions about the interrelation of the voxels via spatial autocorrelation functions, as I described above. Users also choose the size of the areas to be studied, and the smoothing used in the process. These original assumptions were reasonable to begin with, but they were made at a time when there was relatively little data. No one checked them. Until now. And lo and behold, precisely this central assumption describing the mathematical relationship of neighbouring voxels was wrong. So: back to the books; modify software programs, implement new autocorrelation functions closer to empirical reality. And recalculate.

Other assumptions have to do with assuming statistical distributions for the data. This is something that is done often. So the inference procedures involved are called "parametric statistics" because you assume a known distribution for the data. You can normalize the known distribution. One then interprets the area under the curve as "1". If you then plot a value somewhere on the axis and calculate the area behind it, you can interpret this area fraction of 1 as a probability.

So, for example, more than 95% or less than 5% of the area lies behind the axis value "2" (or "-2") of the standard normal distribution. Because the area is normalized to "1", this can then be interpreted as a probability. So you can calculate error probabilities from a known distribution. A common distribution assumption is that based on normal distribution, but there are plenty of other statistical distribution curves where you can then calculate the area fraction of a standardized curve in the same way and thus determine the probability.

On the other hand, we rarely know whether these assumptions are correct. Therefore, as this analysis shows, a non-parametric procedure, i.e. one that makes no distributional assumption about the data, is actually wiser. The discussion about this is already very old and well-known, as are the procedures [2]. We have used them on various occasions, especially in critical situations [3,4]. If you use such simulation or non-parametric statistics properly, you actually have to use the empirically found data. You let the computer generate new data sets, say 10,000, that have similar characteristics, e.g. the same number of points with certain features, and then count how often the features that appear in the empirically found data also appear in the simulated data. If you then divide

the number of features that occurred empirically by the number of features found by chance, you have the true probability that the empirical finding could have occurred by chance.

Of course, such simulations, often called Monte-Carlo analyses (because Monte-Carlo is the big casino) – or non-parametric analyses – are very costly. Even modern, fast computers often need weeks to perform complex analyses.

Anyway, you can see from this example what happens when you violate statistical assumptions: You can no longer interpret probability values based on assumptions and feed the results to the rabbits. In this specific case, a huge literature of neuromythology has emerged. More than half, perhaps as many as 70%, of the approximately 40,000 studies on fMRI methodology would actually have to be repeated or at least re-evaluated, the authors complain. If the data were publicly available, this would be feasible. Unfortunately, in most cases they are not. This is where the complaint of the neuroscientific community meets the call just made by psychologists for everything, but really everything, to be made publicly accessible, protocols, results, data [5]. The authors are calling for a moratorium: first do your homework, first work through the old problems, then do new studies. This will not work everywhere. Because in many cases study data was deleted after 5 years due to applicable laws.

Fabled

Now that's beautifully silly, I think. One has to consider: Most major clinical units in hospitals and most major universities in Germany and the world maintain MRI scanners; the English Wikipedia estimates 25,000 scanners are in use worldwide. The problem with these devices is that once they have been put into operation, they are always connected to the power grid and thus generate high operating costs. You can't simply switch them off like a computer, because that could damage the device, or switching them off and starting them up is itself a very complex and time-consuming process. That's why these devices have to be kept in continuous use, so that their purchase, now worth several million euros, is worthwhile. That is why so many studies are done with them. Because whoever does studies pays for scanner time. No sooner does someone come up with an idea that seems reasonably clever – "let's see which areas of the brain are active when you play music to people or show them pictures they don't like" – than he finds the money to get such a study funded, even in today's climate.

> That brain research has a number of other problems has been noticed by others, as brain researcher Hasler points out in an easy-to-read article.

And so it comes to pass that we have a huge stock, by now we must say, of storybooks about what can happen in the brain when Aunt Emma knits and little Jimmy memorizes nursery rhymes. Beautiful pictures, pretty narratives, all suggesting to us that the most important thing in the world of science at present is knowledge about what makes the brain tick. Except that, in the majority of cases, all these stories have little more value than the sagas of classical antiquity. The sagas of classical antiquity sometimes contain a kernel of truth and are at least exciting. Whether the kernel of truth of the published fMRI studies is greater than that of the sagas? Indeed: the colourful images of the fMRI studies are the baroque churches of postmodernity: beautiful, pictorial narratives of a questionable theology.

Sources and literature

1. Eklund, A., Nichols, T. e., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Science, early edition. Doi: https://doi.org/10.1073/pnas.1602413113.

2. Edgington, E. S. (1995, orig. 1987). Randomization Tests. 3rd Edition. New York: Dekker.
3. Wackermann, J., Seiter, C., Keibel, H., & Walach, H. (2003). Correlations between brain electrical activities of two spatially separated human subjects. Neuroscience Letters, 336, 60-64.
4. Schulte, D., & Walach, H. (2006). F.M. Alexander technique in the treatment of stuttering – A randomized single-case intervention study with ambulatory monitoring. Psychotherapy and Psychosomatics, 75, 190-191.
5. Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.

Date Created
10 May 2022