Reproducibility

Pipeline choices alter neuroimaging findings

Martin Lindquist

Seventy laboratories that analysed the same neuroimaging data each produced different results. This finding highlights the potential consequences of a lack of standardized pipelines for processing complex data. **See p.84**

For most types of big data, from genome sequences to medical images, there is no single 'best' way to process the data. This issue is exemplified by the substantial differences in how individual laboratories preprocess and analyse data from functional magnetic resonance imaging (fMRI) experiments, which generate information about brain activity. Indeed, a survey of fMRI studies found that nearly every study used a different analysis pipeline[1]. Botvinik-Nezer *et al.*[2] provide further evidence of this variability on page 84, highlighting how analytical choices made by individual researchers can greatly influence the findings gleaned from an fMRI data set. The work is bound to spark lively discussion.

Functional MRI experiments produce a series of images of the brain at work. These images go through several stages of processing and analysis to determine which brain regions show significant activity. However, the choice of pipeline can alter the outcome of an fMRI study. In 2012, for instance, 6,912 unique processing and analysis pipelines were applied to the same fMRI data set[3]. Many of the pipelines indicated that the same general brain regions were active, but the locations at which neural activity was deemed to be highest varied widely depending on the pipeline used.

Botvinik-Nezer *et al.* explored this phenomenon further. The authors gave 70 independent research teams the same fMRI data set, generated from 108 people performing decision-related tasks. The teams were asked to use the data set to test nine hypotheses, each of which posited that activity in a specific brain region related to a particular feature of the tasks. This allowed Botvinik-Nezer *et al.* to evaluate the impact of analytical flexibility on fMRI results 'in the wild' (rather than performing the analyses themselves, as was done in the 2012 study). Notably, no two teams chose identical workflows to analyse the data, resulting in substantial variation in the results (Fig. 1).

"The fact that each team's results were so pipeline-dependent is highly problematic."

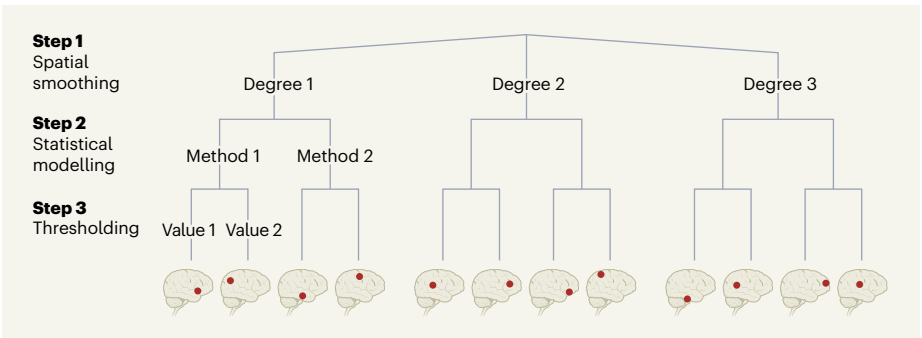

Figure 1 | Implications of choosing a neuroimaging pipeline. Botvinik-Nezer *et al.*[2] report that researchers process neuroimaging data using a wide variety of pipelines, which can produce varying results. In this simplified example, the pipeline has three steps: spatial smoothing of the images to reduce noise, which in this example is done to three different degrees; statistical modelling, which in this example can be performed in one of two ways; and 'thresholding' of statistical tests associated with these models to determine the level at which neuronal activity in each brain region is deemed to be significant, which in this example is set to two different values. Making different choices for each step leads to a different end point – the red dots represent how activation moves throughout the brain depending on which pipeline is used.

It is standard in neuroimaging to test types of hypothesis such as the nine put forward by Botvinik-Nezer *et al.* by generating a statistical map. The map comprises a compendium of statistical tests performed on different parts (voxels) of the image. The results of these tests are subjected to a process called thresholding to set the level, for each voxel, at which activity picked up by the experiment is deemed to indicate real neuronal activity rather than noise. The authors found considerable differences between each research team's results even when the underlying statistical maps they had used were highly correlated.

The strongest factor in explaining the differences between each team's results was the spatial smoothness of the data being analysed. Spatial smoothing is a preprocessing step in which the activity of each voxel is averaged with that of its neighbours – a process designed to reduce noise. Higher estimated smoothness was associated with a greater likelihood of reaching a significant outcome for each hypothesis.

The fact that each team's results were so pipeline-dependent is highly problematic, particularly because the exact configuration of analytical pipelines is often poorly described in research articles. Moreover, sensitivity analyses – which assess how different pipeline choices might affect an experiment's outcome – are rarely performed in neuroimaging. However, Botvinik-Nezer and colleagues offer several reasonable suggestions for addressing the concerns that their work will raise.

The first is to share unthresholded activity maps, because this will allow image-based meta-analysis. The authors found that such an analysis, which aggregated information across teams, yielded consensus results, no doubt aided by the fact that the spatial patterns in the activity maps were highly correlated across groups.

Second is a call to publicly share both data and code, making it easier for others to attempt to reproduce a paper's findings. In this regard, the authors model good behaviour by making all their data and processing pipelines publicly available.

Third is the use of pre-registration – in which a hypothesis and analysis plan is made public before the experiment is performed. It is unfortunately common for researchers to explore various pipelines to find the version that yields the 'best' results, ultimately reporting only that pipeline and ignoring the others. This practice can lead to errors and make it difficult to replicate findings[4]. Pre-registration would make it easy to detect cases in which researchers had explored various pipelines.

Fourth is to analyse all data through

multiple pipelines and use the results to obtain consensus findings. This could be achieved by implementing the type of meta-analysis used by Botvinik-Nezer and co-workers.

Another goal of the study was to evaluate how accurately researchers could predict the number of teams that would report significant results for each hypothesis. To study this, the authors ran separate 'prediction markets', one for the analysis teams and one for researchers who did not participate in the analysis. In them, researchers attempted to predict the outcomes of the scientific analyses and received monetary payouts on the basis of how well they predicted performance. Participants — even researchers who had direct knowledge of the data set — consistently overestimated the likelihood of significant findings. Botvinik-Nezer *et al.* do not explicitly explore how the analysts' prior beliefs affected their findings and pipeline choices. For example, if a research finding does not initially coincide with expectations, will groups seek to alter pipelines until expectations and results align? The prevalence of pipeline exploration implies that this is likely.

What other improvements could be made for the future? One approach is to use pipeline-optimization tools[5,6] to reduce analysis flexibility. These tools automatically identify pipelines that maximize reproducibility, which can reduce the risk of excessive pipeline exploration and selective reporting. In addition, increased use of sensitivity analyses to evaluate the effects of pipeline decisions would provide a better understanding of the link between analysis choices and research findings. The fact that activity maps were highly correlated across groups implies that multivariate statistical approaches, which identify spatial patterns in the data, might provide more-consistent results across pipelines than would a series of tests performed at individual voxels.

Ultimately, neuroimaging results should be carefully verified using independent data sets to demonstrate generalizability across samples, research contexts and populations. A positive example of this already being done comes from an approach for developing predictive models on the basis of brain activation, which can be shared, tested in multiple contexts and used in applied settings[7].

It seems unlikely that the fMRI field will ever coalesce on a standard workflow that is applicable to all types of study, because studies tend to be too varied for one pipeline to always be appropriate. But Botvinik-Nezer *et al.* conclude their paper by calling for an increased awareness of the situation, and a drive to improve the quality of method reporting. This is wise and prudent advice that researchers in any field analysing high-dimensional data would be well advised to heed.

Martin Lindquist is in the Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA.
e-mail: mlindqui@jhsph.edu

1. Carp, J. *NeuroImage* **63**, 289–300 (2012).
2. Botvinik-Nezer, R. *et al. Nature* **582**, 84–88 (2020).
3. Carp, J. *Front. Neurosci.* **6**, 149 (2012).
4. Simmons, J. P., Nelson, L. D. & Simonsohn, U. *Psychol. Sci.* **22**, 1359–1366 (2011).
5. Strother, S. C. *et al. NeuroImage* **15**, 747–771 (2002).
6. Churchill, N. W. *et al. Hum. Brain Mapp.* **33**, 609–627 (2012).
7. Woo, C. W., Chang, L. J., Lindquist, M. A. & Wager, T. D. *Nature Neurosci.* **20**, 365 (2017).

This article was published online on 20 May 2020.

Optical physics

Light turned into exotic Laughlin matter

Laura Corman

A Laughlin state is a phase of matter that has remarkable features, such as excitations that behave as a fraction of a particle. The long-sought creation of a photonic Laughlin state is a milestone for the field of quantum simulation. **See p.41**

Solving a difficult physics problem can be surprisingly similar to assembling an interlocking mechanical puzzle. In both cases, the particles or pieces look alike, but can be arranged into a beautiful structure that relies on the precise position of each component (Fig. 1). In 1983, the physicist Robert Laughlin made a puzzle-solving breakthrough by explaining the structure formed by interacting electrons in a device known as a Hall bar[1]. Although the strange behaviour of these electrons still fascinates physicists, it is not possible to simulate such a system or accurately measure the particles' ultrashort time and length scales. On page 41, Clark *et al.*[2] report the creation of a non-electronic Laughlin state made of composite matter–light particles called polaritons, which are easier to track and manipulate than are electrons.

To picture a Laughlin state, consider a Hall bar, in which such states are usually observed (Fig. 2a). In these devices, electrons that are free to move in a two-dimensional plane are subjected to a strong magnetic field perpendicular to the plane. In classical physics, an electron at any position will start moving along a circular trajectory known as a cyclotron orbit, the radius of which depends on the particle's kinetic

GETTY

Figure 1 | An interlocking mechanical puzzle. Certain many-body phases of matter, such as Laughlin states, depend on the precise position of all the particles in the system, just like the pieces in an interlocking puzzle.