Citation for published version:
Button, KS 2019, 'Double-dipping revisited', *Nature Neuroscience*, vol. 22, no. 5, pp. 688-690.
https://doi.org/10.1038/s41593-019-0398-z

DOI:
10.1038/s41593-019-0398-z

Publication date:
2019

Document Version
Peer reviewed version

Link to publication

Historical News & Views: Reproducibility

Double dipping revisited

Katherine S. Button

Department of Psychology, University of Bath, Claverton Down, Bath, BA2 7AY, UK.

k.s.button@bath.ac.uk

Robust conclusions require rigorous statistics. In 2009 a seminal paper described the dangers and prevalence of double dipping in neuroscience. Ten years on I consider progress towards statistical rigour in neuroimaging.

The human mind struggles with probabilistic reasoning, tending instead towards mental-shortcuts that leave us prone to cognitive bias and logical fallacies. The scope for these errors increases with the complexity of the analytical pipeline, where decision is layered upon decision, assumption upon assumption. Circularity in analysis is a logical fallacy that occurs where the same data are used twice (or more) in the same analysis: once to select a subset of data of interest and again to test how interesting those same data are. Such double dipping into the data violates the assumption of independence, undermining statistical inferences, inflating effect estimates, and increasing the chance of false positive results. The dangers of double dipping in statistical analyses are well documented. Yet circularity is a seductive trap, beautifying results and feeding our confirmation bias. Methodological precautions can protect us from its allure, but how widely are they employed?

In 2009 Kriegeskorte and colleagues[1] examined 134 functional MRI (fMRI) articles published the year before in *Nature, Nature Neuroscience, Science, Journal of Neuroscience* and *Neuron*. They found that an astonishing 42% contained circular analyses, with the analyses of an additional 14% of papers unclear. Circular analysis is not unique to fMRI, yet Kriegeskorte and colleagues' findings shook the fMRI community to its core. I put this down to several reasons. First, their analysis provided a prevalence estimate that unequivocally demonstrated the ubiquity of this error even in the most prestigious publications. Second, their detailed examples of double dipping in the context of fMRI and electrophysiology experiments provided a tangible way for readers to conceptualise the problem as directly applied to imaging research. That is, they made an abstract problem concrete. Third, and most importantly, they captured the Zeitgeist.

During this time, the high prevalence of double dipping in fMRI studies could be viewed as a symptom of the growing pains of a relatively young 'big-data' discipline, and the wider irreproducibility milieu bubbling away across the biomedical sciences[2]. Since its development as a technique in the early 1990s, fMRI saw two decades of exponential growth, from around 350 articles published in 1998 to over 2600 a decade later in 2008 (Figure 1). At this point the field saw rapid developments in analytic methods and imaging procedures, moving from a diversity of locally developed analysis software to converge on the few open-source analysis packages widely used today[3]. The complexity and high-dimensionality of fMRI data coupled with the myriad analytical packages and pipelines raised a plethora of statistical conundrums. How best to pre-process the data, control for multiple comparisons, or select regions of interest?

In 2005 John Ioannidis published his seminal paper "Why most research findings are false"[2], calling into question the reliability of findings across the biomedical sciences. He demonstrated that widespread use of shoddy research practices such as reliance on underpowered studies, undisclosed flexibility in analyses, and selective reporting of positive results can lead to a worryingly high proportion of false positive results. 2009 was a similar watershed year for the neuroimaging community. Alongside Kriegeskorte et al's[1] 'Double Dipping' paper and Vul et al's[4] related 'Voodoo Correlations' paper, Bennett and colleagues[5] published their Ig Noble prize-winning demonstration of how poor control for multiple comparisons could lead to (false positive) evidence of neural activation during a social perspective-taking task in the brain of a dead salmon. The fMRI community embarked on a period of intense methodological introspection. Ten years on, how far have we come in making commonplace double dipping and related questionable research practices a thing of the past?

As suggested by Kriegeskorte et al., perhaps the simplest way to prevent circular analysis is to split one's data into two independent samples, one for exploration, the other for confirmation. fMRI is expensive and sample sizes have been traditionally very small. While there is some evidence that sample sizes are on the rise, the average sample size in fMRI studies in 2015 was still only 19 participants [3]. Splitting a sample of this size is clearly problematic in terms of loss of statistical power[6]. Individual fMRI studies have therefore instead tended to opt for retaining the full sample in a single confirmatory study (ostensibly at least) and used selection criteria that are demonstrably independent of the hypothesis test, such as using anatomical atlases or functional localiser tasks to define brain regions of interest.

Both have limitations; anatomical selection works well for small, clearly defined anatomical regions such as the amygdala, but less well for large structures such as the medial prefrontal cortex. Functional localisers, where regions of interest are identified using a separate task thought to activate the same neural processes as those under investigation, are often preferred for larger anatomical areas. However, functional localisers are subject to several assumptions and suffer the same issues of signal-to-noise in small datasets as do the tests of hypothesis[6]. So how do you achieve separation of exploration and confirmation in a field that has been traditionally dominated by small, expense-constrained datasets? One answer is through better data-sharing, collaboration, and data reporting.

Historically there has been little tradition of data-sharing in fMRI. Even the data in fMRI papers, presented in the form of peak voxel coordinates, were of limited use to other researchers wishing to replicate or build from an initial study's finding, as they provide a poor summary of the vast amounts of data and analysis performed in the typical fMRI experiment. However recent years have seen a growing number of tools for supporting open fMRI data-sharing, and their use is gaining in popularity. For example, the COINS service currently hosts data on over 50,000 participants in 702 studies (http://coins.mrn.org) and the NeuroVault repository (http://neurovault.org) hosts over 1000 public collections. Neurosynth (http://neurosynth.org) provides a data-synthesis service that summarizes available evidence from published peak voxel data, which is ideal for independently selecting regions of interest.

Recent years have also seen the development of successful neuroimaging consortia such the ENIGMA (Enhancing Neuro Imaging Genetics by Meta-Analysis) consortium[7] and the 1000 Functional Connectomes Project and its International Neuro-Imaging Data-Sharing Initiative (INDI)[8]. These initiatives pave the way for the creation of large datasets, such as The Human Connectome Project (http://www.humanconnectomeproject.org/), the UK Biobank (http://imaging.ukbiobank.ac.uk/), and prospective cohort studies such as Imagen (https://imagen-europe.com/about/project/) which make their datasets freely available to academic researchers. These datasets can be used for exploratory hypothesis generation and independent selection criteria, confirmatory replication, or both.

However, while a selection method that ensures independence, such as split and/or shared datasets, is necessary to prevent circular analysis, it is not sufficient. It must also be demonstrated that the selection method was chosen before data collection to ensure that a presumably independent selection criterion is not applied retrospectively after having seen a potentially interesting result. This is akin to Hypothesising After the Results are Known (HARKing), a similar form of circular thinking where hypotheses are retrofitted to exploratory findings. Preregistration is widely recognised as the most powerful way of preventing HARKing and demonstrating that selection criteria are independent of subsequent analysis [3,6,9]. It involves registering the study with a detailed pre-specification of the study design, primary outcome, and analysis plan in advance of data-collection. In this way, confirmatory research testing a priori hypotheses (i.e., those made before data collection) are clearly differentiated from exploratory post-hoc analyses which are used to generate hypotheses after the data are observed.

Preregistration has been standard practice in clinical trials many years[10] . However, despite its wide advocation[3,6,9-11], preregistration has yet to gain traction within the neuroimaging community. In 2013 the Open Science Framework (http://osf.io/) provided a service to preregister studies across various fields of science, including neuroscience. Since then more than 28,500 studies have been registered on OSF. Of these, only 102 relate to “fMRI” (search date 21 March 2019). To put this into context, searching Web of Science found 26,068 “fMRI” articles were published over the same period. By contrast, the field of “eye-tracking” registered 328 studies on OSF, and published 5,029 articles.

Transparent reporting of results and methods is the bedrock for reproducible science, yet historically, reporting standards in fMRI studies have been inconsistent[12,13]. To address this the Organisation for Human Brain Mapping (OHBM) convened a Committee on Best Practices in Data Analysis and Sharing (COBIDAS) in 2015-16, which issued a detailed set of reporting guidelines (hhtp://www.humanbrainmapping.org/COBIDAS)[9]. Relative to other reporting checklists such as those for clinical trials (http://www.equator-network.org/reporting-guidelines/consort-abstracts), the COBIDAS MRI checklist is formidable. This reflects the length and complexity of analytic pipelines and the extent of information required for another researcher to be able to replicate an fMRI finding [14].

By adopting stringent statistical criteria, independent replication, large collaborative consortia, complete reporting of statistical results, and routine sharing of fine-grained statistical results, fields such as genetics have seen a step-change in their rate of scientific discovery[15]. Many hundreds more reproducible findings have been found in recent years since whole-genome methods were developed than were produced in 15 years of small-scale candidate-gene studies. Similarly, clinical trials, which have widespread adoption of preregistration and adherence to transparent reporting guidelines (at least in the top journals), have resulted in a flourishing field of evidence-synthesis, with high-quality systematic reviews and meta-analysis forming the basis of national and global healthcare policies.

So is double dipping in fMRI research is a thing of the past? The pessimistic answer is, no. A more optimistic answer is, not yet but it soon could be. Recent years have seen the technological ingredients for rigorous and reproducible functional brain imaging fall into place. Widespread adoption of practices such as preregistration for confirmatory analyses, adherence to recommended best-practices in analysis and data-sharing, transparent reporting of results, large-scale collaboration, and a cultural-shift towards independent replication have the potential to bring about a step-change in the reproducibility of fMRI findings. With a shift in the reward structures to promote routine use of such rigorous methods over the next ten years, commonplace errors such as double dipping may indeed become a thing of the past.

Acknowledgments

I am grateful to Tom Freeman, George Stothart, Esther Walton and the wider Neurostim group for helpful discussions.

1 Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci* **12**, 535-540, doi:10.1038/nn.2303 (2009).

2 Ioannidis, J. P. Why most published research findings are false. *PLoS medicine* **2**, e124, doi:10.1371/journal.pmed.0020124 (2005).

3 Poldrack, R. A. *et al.* Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature reviews. Neuroscience* **18**, 115-126, doi:10.1038/nrn.2016.167 (2017).

4 Vul, E., Harris, C., Winkielman, P. & Pashler, H. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspect Psychol Sci* **4**, 274-290, doi:10.1111/j.1745-6924.2009.01125.x (2009).

5 Bennett, C. M., A., B. A., Miller, M. B. & Wolford, G. L. Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction. *Journal of Serendipitous and Unexpected Results* **1**, 1-5 (2009).

6 Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience* **14**, 365-376, doi:10.1038/nrn3475 (2013).

7 Thompson, P. M. *et al.* The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav* **8**, 153-182, doi:10.1007/s11682-013-9269-5 (2014).

8 Biswal, B. B. *et al.* Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 4734-4739, doi:10.1073/pnas.0911855107 (2010).

9 Nichols, T. E. *et al.* Best practices in data analysis and sharing in neuroimaging using MRI. *Nat Neurosci* **20**, 299-303, doi:10.1038/nn.4500 (2017).

10 Dickersin, K. & Rennie, D. Registering clinical trials. *JAMA : the journal of the American Medical Association* **290**, 516-523, doi:10.1001/jama.290.4.516 (2003).

11 Munafò, M. R. *et al.* A manifesto for reproducible science. *Nature Human Behaviour* **1**, 0021, doi:10.1038/s41562-016-0021 (2017).

12 Carp, J. The secret lives of experiments: Methods reporting in the fMRI literature. *Neuroimage* **63**, 289-300, doi:10.1016/j.neuroimage.2012.07.004 (2012).

13 Guo, Q. *et al.* The Reporting of Observational Clinical Functional Magnetic Resonance Imaging Studies: A Systematic Review. *PloS one* **9**, doi:ARTN e94412

10.1371/journal.pone.0094412 (2014).

14 Carp, J. On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. *Frontiers in neuroscience* **6**, doi:10.3389/fnins.2012.00149 (2012).

15 Ioannidis, J. P., Tarone, R. & McLaughlin, J. K. The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology* **22**, 450-456, doi:10.1097/EDE.0b013e31821b506e (2011).

Figure 1.

Number of fMRI articles published per year from 1990 to 2018. The graph depicts exponential growth until 2014, after which growth flattens. The curve is overlaid with examples of key papers that signalled field-wide concerns about reliability of research findings, as well as key initiatives to address these concerns and promote reproducible science. This is not an exhaustive list, but it serves to illustrate the emergence of the key ingredients for reproducible science, such as platforms to support open data-sharing, automatic evidence-synthesis of published results, and preregistration of study protocols, as well as the publication of standardised guidance for data-analysis and transparent reporting of methods and results. Widespread adoption of these practices could bring about a step-change in the reliability of fMRI findings, protecting against errors such as circular analysis and other related dubious practices that were common in 2008. Searches performed on 11 March 2019 on Web of Science ((TS = (fMRI OR functional Magnetic Resonance Imaging)) AND DOCUMENT TYPES: (Article))