Using meta-analysis to assess reproducibility across analysis teams in fMRI research

Kendra Oudyk
NEUR608 Neuroimaging Data Science
McGill University
Drs. Boris Bernhardt and Bratislav Misic, Instructors
2019-10-28

Over the past few decades, researchers have become more aware of reproducibility issues in science (Ioannidis, 2005). In order for a result to be accepted as scientific knowledge, the result needs to be consistent when the study is performed multiple times with the exact same methods. Further, the result should be stable over small differences in methodology, such as different samples from the same population, different analysis software, and different research groups. Yet, we are becoming more aware that these seemingly minor variables can influence results.

There has been recent concern over these reproducibility problems in functional magnetic resonance imaging (fMRI; e.g., Carp, 2012a, 2012b). In research using fMRI, there are many steps involved in preprocessing and analyzing the images, with different options for how to perform each step, and often different software that implement the same particular option. These researcher degrees of freedom have been shown to influence results (Carp, 2012a; Poldrack et al., 2017). Even some broadly accepted methods, such as cluster-wise inference, have been shown to be more or less problematic in some implementations; this is especially worrisome because the increased false-positive results may be distributed in non-uniform manner over the brain (Eklund, Nichols, & Knutsson, 2016). Further, some variables that do not seem like choices can make a difference; results can vary over different versions of the same software Gronenschild et al., 2012 and different computer operating systems Glatard et al., 2015. It becoming clear that these choices, known as researcher degrees of freedom, have an impact on results.

The present study

However, we do not yet know how severe these effects are. The purpose of this study is to estimate the effect of research group on neuroimaging results. To this end, 55 research teams were given the exact same data and were asked to analyze given hypotheses using their standard pipeline. Here, we use an image-based meta-analysis to assess the consistency of results across these research teams. Further, we qualitatively compare the results of this meta-analysis with a) meta-analytic results from two automated term-based meta-analysis websites (Dockès et al., 2018; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011), b) the

results from the *Science* study that first described this topic (Tom, Fox, Trepel, & Poldrack, 2007), and c) the results shown in the paper describing the validation of the dataset that was shared with analysis teams (Botvinik-Nezer et al., 2019).

The data used here is part of the Neuroimaging Analysis Replication and Prediction Study (NARPS, *https://www.narps.info/analysis.html*), and this is not the official report from that larger study. There were nine hypotheses to be tested in this larger study. Here, we examined one hypothesis: that there would be a parametric effect of gain, specifically, a "positive effect in ventromedial PFC - for the equal indifference group" (`https://www.narps.info/analysis.html`).

Background to the original dataset

The original dataset that was analyzed by the various research teams was fMRI data from participants undergoing the mixed gambles task (Tom et al., 2007). This paradigm is designed to measure decision making in the face of different levels of risk in terms of potential gains and losses. Participants had to decide whether to take a gamble based on the given mixed gamble, that is, the potential gain and potential loss.

There were two particular paradigms, tested on separate groups of participants. In the 'equal indifference' condition, participants the gains could reach double the losses (Tom et al., 2007); in the 'equal range' condition, the possible ranges of losses and gains were equal (De Martino, Camerer, & Adolphs, 2010). Generally, it is thought that people need the gain to be at least twice the potential loss in order to take a gamble (Abdellaoui, Bleichrodt, & Paraschiv, 2007; Kahneman, Knetsch, & Thaler, 1990, 1991; Tversky & Kahneman, 1992, c.f., Gal & Rucker, 2018); hence the name of the equal indifference condition.

Methods

(See Botvinik-Nezer et al. (2019) for a description of the methods relating to data collection and preprocessing; I will re-word these sections for the final draft.)

Acquisition of original data

Participants.
Procedure.
Task.
MRI data collection.

Preprocessing of original data

Anatomical data preprocessing.
Functional data preprocessing.

Analysis of original data

The project was advertised and research teams could apply to analyze the data; in return, a maximum of three researchers from each team could be co-authors on the resulting papers. Participating groups were given access to the raw and the pre-processed MRI data (they could choose which to use), as well as the behavioral data.

Participating teams were given access to the data in November 2018, and were allowed 3 months to perform a whole-brain corrected analysis of task-related activation for each hypothesis. They were told to use their standard analysis pipeline, and they had the option of using the raw or preprocessed data.

At the end of the analysis period, each team returned the following results:

- A description of how they analyzed the data,
- A yes/no decision for each of the nine hypotheses,
- A group-level thresholded statistical map for each contrast,
- A group-level unthresholded statistical map for each contrast, and
- If available, participant-level contrast maps and variance maps.

Data used for the present meta-analysis

The data for the meta-analysis consisted of the results map from each analysis team, for one of the hypotheses. These are group-level whole-brain unthresholded statistical maps (z-maps). Team identity was anonymized before we received the data. The hypothesis examined here was that there would be a parametric effect of gain, specifically: "Positive effect in ventromedial PFC - for the equal indifference group" (`https://www.narps.info/analysis.html`).

Although this data has not yet been made publicly available, these maps can be visualized at `https://github.com/koudyk/narps_meta/blob/master/analyses/0_data_visualization.ipynb` *(you may have to refresh the page a few times due to a bug in the GitHub rendering of .ipynb files).*

Mixed-effects (MFX) meta-analysis

We meta-analyzed these group-level z-maps using a mixed-effects model, which is considered the gold standard in neuroimaging meta-analysis (Salimi-Khorshidi, Smith, Keltner, Wager, & Nichols, 2009). *(more here)*

We used the implementation in NiMARE (`https://github.com/neurostuff/NiMARE`).

Comparison with online meta-analyses

In order to compare these results against previous literature, we also used two automated term-based meta-analytic tools with searches for the term "decision making". These tools are NeuroSynth (Yarkoni et al., 2011, `https://www.neurosynth.org/`) and NeuroQuery (Dockès et al., 2018, `https://neuroquery.saclay.inria.fr/`).

NeuroSynth was the first automated meta-analysis tool in fMRI research. It has access to a large repository of published research articles that use fMRI, and it meta-analyzes them by automatically extracting coordinates from the papers and performing a coordinate-based meta-analysis. It selects which papers to analyze based on a linguistic analysis of the paper abstracts; the idea is that if a term like 'decision making' is mentioned in the abstract, then the paper is probably about that term, and this probability increases as the term is repeated in the abstract.

NeuroQuery is a more recent tool that has access to a larger dataset of full-text publications, and it uses a different approach to selecting papers. It uses a multivariate predictive model to determine which papers are relevant to a given search, and it allows for searches that are longer and more complex.

Since these tools use publication texts as their data source, they both only have access to the peak coordinates reported in papers, and so they can only perform coodinate-based meta analyses. Image-based meta-analyses are preferable, since they contain more information that is lost in thresholding and peak extraction. However, the strength of these tools lies in the large number of studies that can be assessed, and the speed/convenience of the interfaces. In a search for "decision making", NeuroSynth found 509 relevant articles, and NeuroQuery found 2389.

Results and Discussion

Results of MFX meta-analyses

See Figure 1.

Comparison with online meta-analyses

See Figures 2, and 3.

Comparison with original study results

See Figure 4.

Comparison with validation of participant-level dataset

See Figure 5.

Reproducibilty

The original participant-level data that was used by the analysis teams is publicly available on OpenNeuro.org, at `https://openneuro.org/datasets/ds001734/versions/1.0.4`.

The group-level images from each team will soon be made publicly available, according to Botvinik-Nezer et al. (2019).

The code used in the present study is publicly available at `https://github.com/koudyk/narps_meta`.

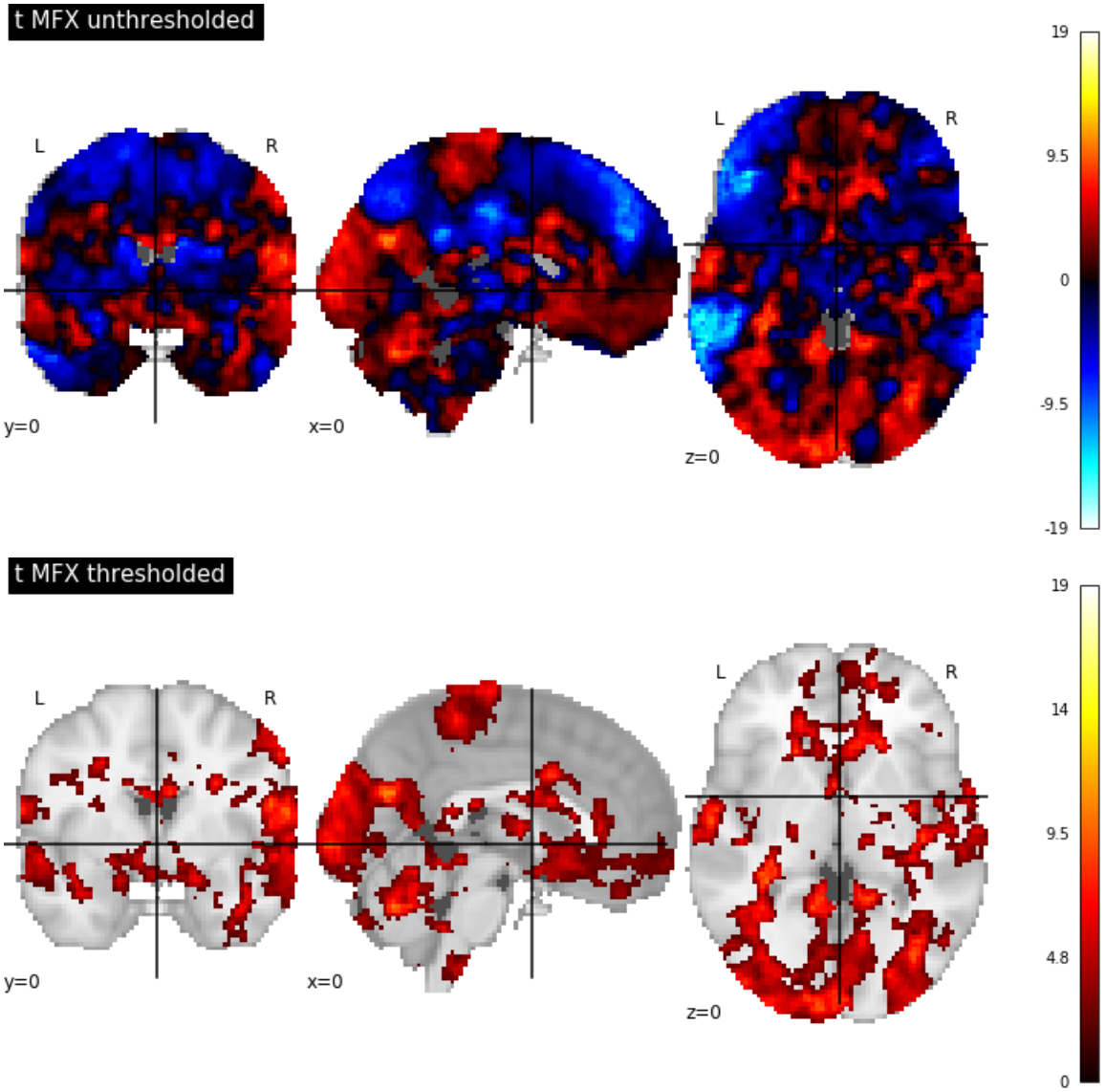

Figure 1. Results of the mixed-effects model meta-analysis, unthresholded and thresholded with a FDR-corrected threshold at $p < .05$

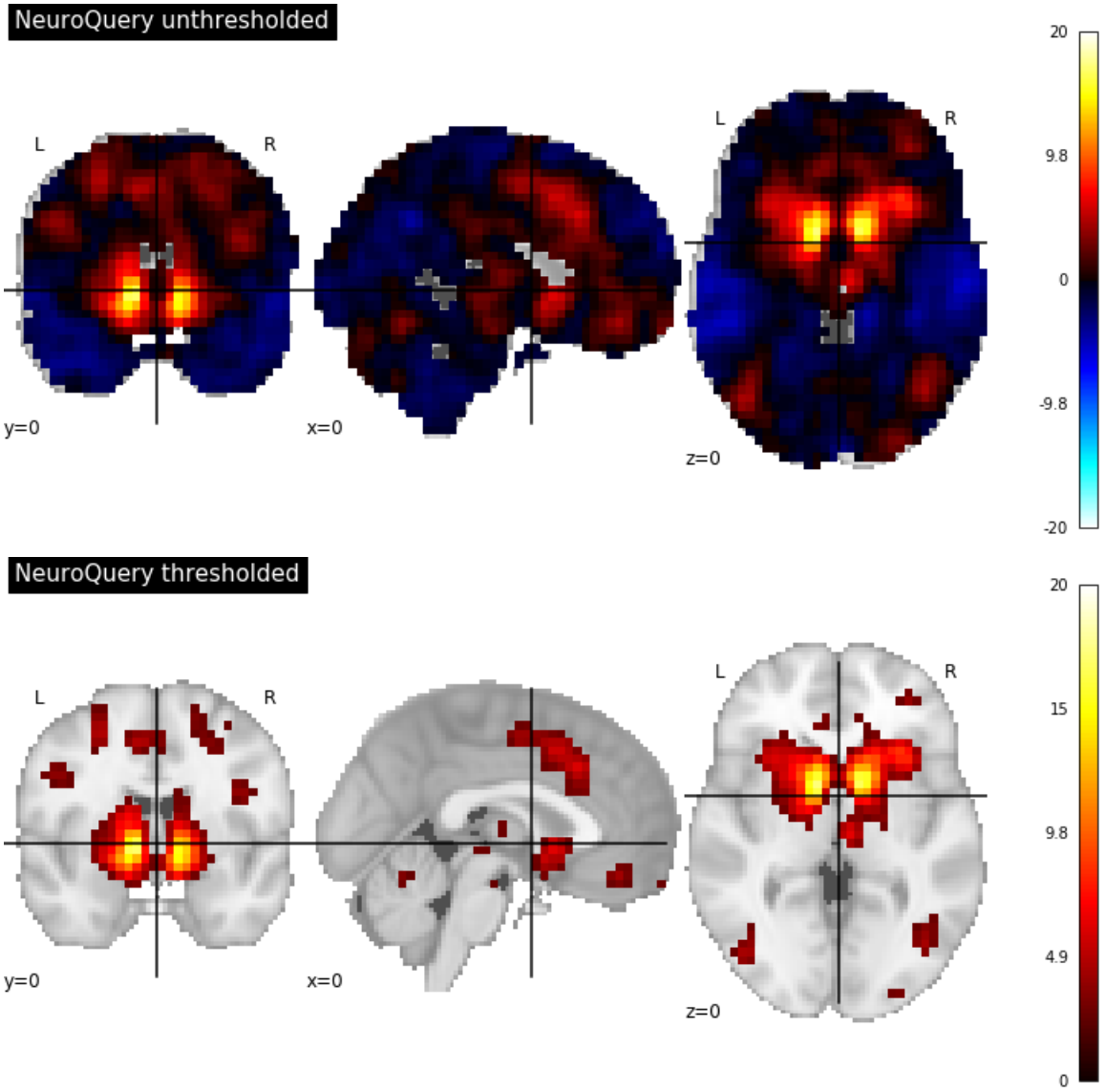

Figure 2. Results of the term-based automated meta-analysis on NeuroQuery (Dockès et al., 2018) for the term "decision making", unthresholded and thresholded (by me) with a FDR-corrected threshold at $p < .05$

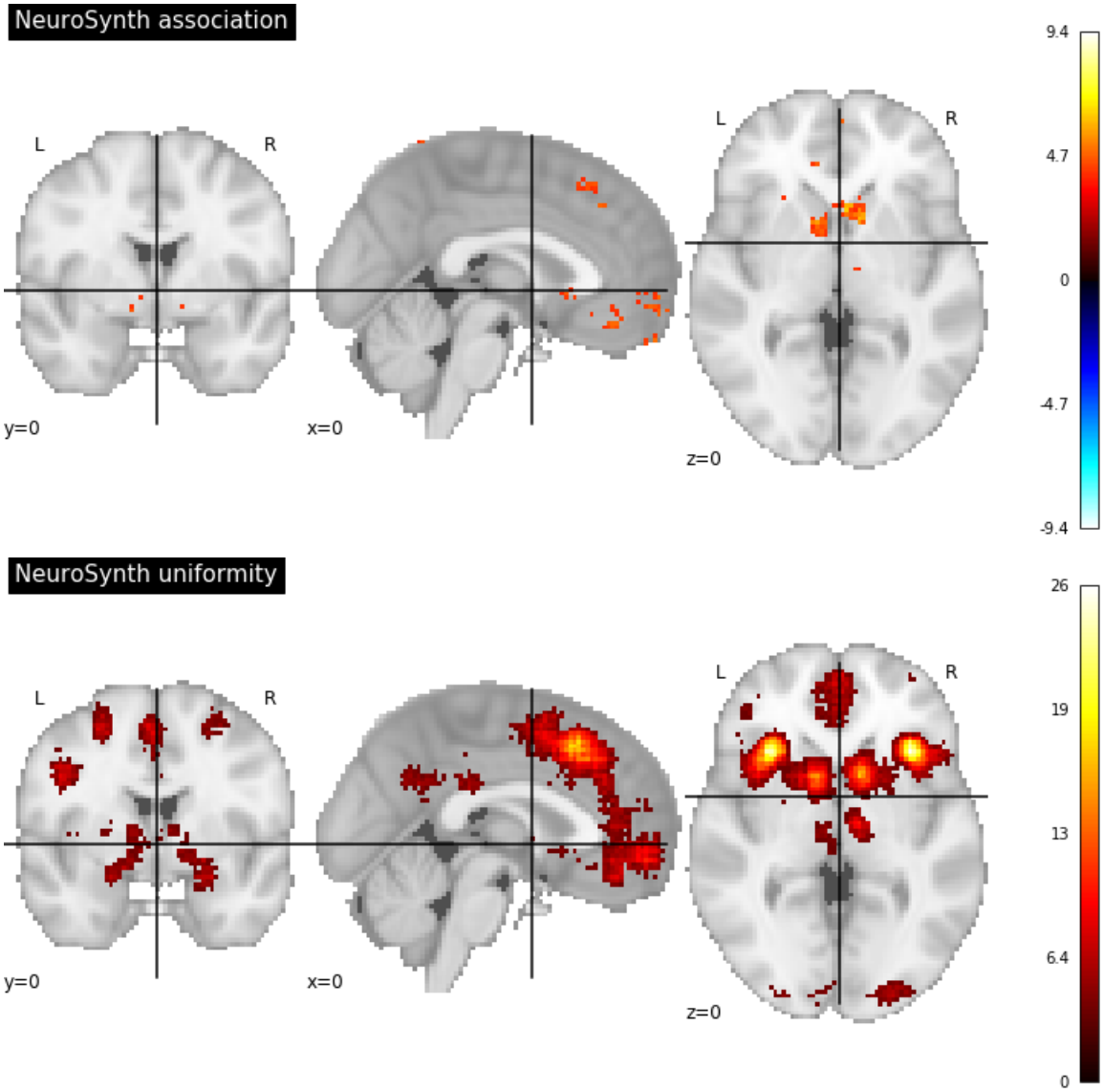

Figure 3. Results of the term-based automated meta-analysis (association test and uniformity test) on NeuroSynth (Yarkoni et al., 2011).

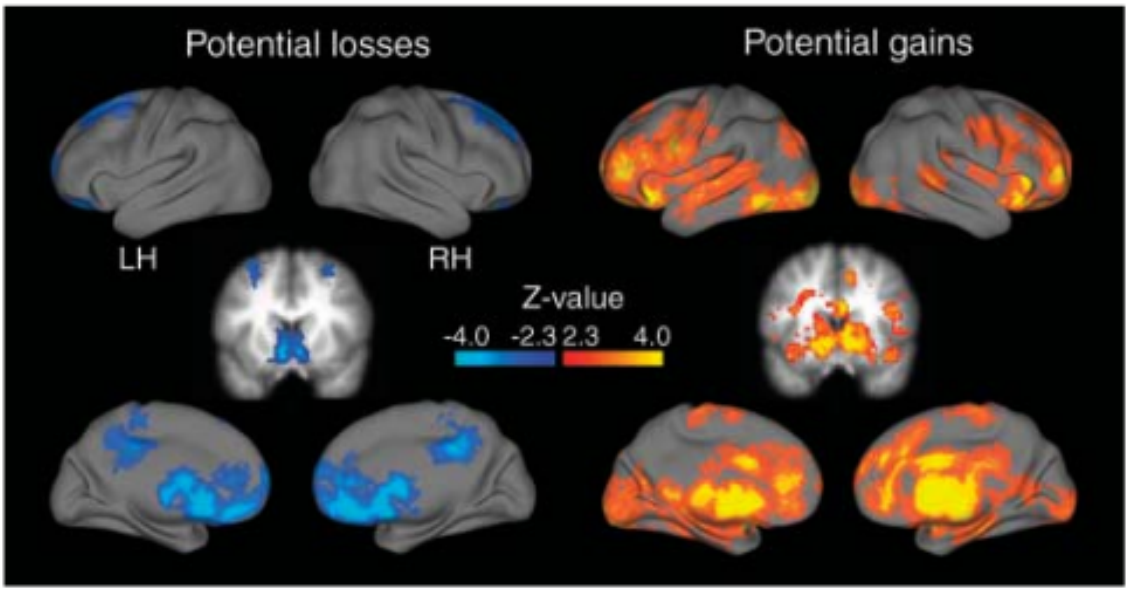

Figure 4. Results from the original paper on this topic by Tom et al. (2007). Their figure caption is: "Whole-brain analysis of parametric responses to size of potential loss (left) or gain (right). Statistical maps were projected onto an average cortical surface with the use of multifiducial mapping in CARET software (Van Essen, 2005); coronal slices (y = 10) are included to show ventral striatal activation. All maps are corrected for multiple comparisons at the whole-brain level by means of cluster-based Gaussian random field correction (Worsley, Evans, Marrett, & Neelin, 1992) at $P < 0.05$. LH, left hemisphere; RH, right hemisphere." (pp. 516)

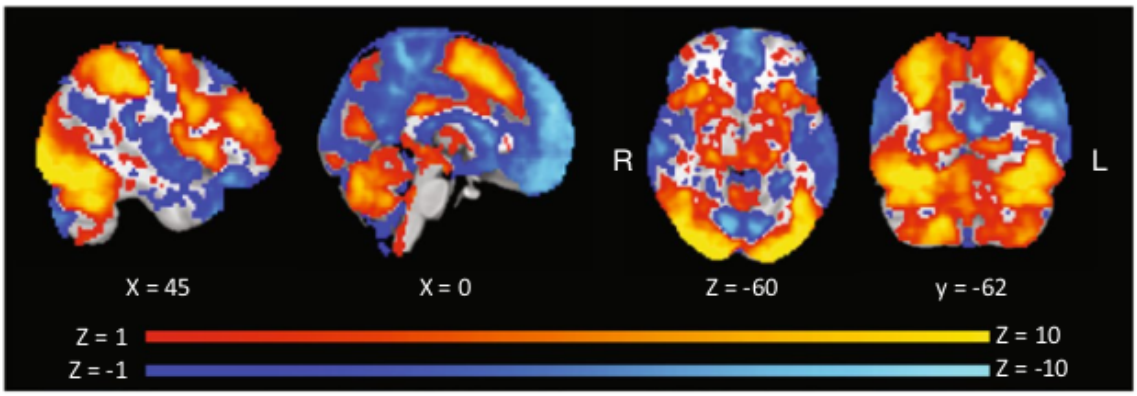

Figure 5. Results from the paper by Botvinik-Nezer et al. (2019), describing their validation of the original dataset that was used by the analysis teams. Their figure caption is: "Uncorrected results of task versus baseline. Uncorrected Z values are presented, thresholded at $Z > 1$ for positive activations (hot colors) and $Z < -1$ for negative activations (cold colors). This analysis was only used for validation." (pp. 7)

References

Abdellaoui, M., Bleichrodt, H., & Paraschiv, C. (2007). Loss aversion under prospect theory: A parameter-free measurement. *Management Science*, *53*(10), 1659–1674.

Botvinik-Nezer, R., Iwanir, R., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., ... Schonberg, T. (2019, July). fMRI data of mixed gambles from the Neuroimaging Analysis Replication and Prediction Study. *Scientific Data*, *6*(1), 1–9. Retrieved 2019-10-26, from `https://www.nature.com/articles/s41597-019-0113-7` doi: 10.1038/s41597-019-0113-7

Carp, J. (2012a). On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. *Frontiers in Neuroscience*, *6*. Retrieved 2019-09-25, from `https://www.frontiersin.org/articles/10.3389/fnins.2012.00149/full` doi: 10.3389/fnins.2012.00149

Carp, J. (2012b, October). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300. Retrieved 2019-09-30, from `http://www.sciencedirect.com/science/article/pii/S1053811912007057` doi: 10.1016/j.neuroimage.2012.07.004

De Martino, B., Camerer, C. F., & Adolphs, R. (2010, February). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences*, *107*(8), 3788–3792. Retrieved 2019-10-26, from `http://www.pnas.org/lookup/doi/10.1073/pnas.0910230107` doi: 10.1073/pnas.0910230107

Dockès, J., Wassermann, D., Poldrack, R., Suchanek, F., Thirion, B., & Varoquaux, G. (2018). Text to brain: predicting the spatial distribution of neuroimaging observations from text reports. In *International conference on medical image computing and computer-assisted intervention* (pp. 584–592).

Eklund, A., Nichols, T. E., & Knutsson, H. (2016, July). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*(28), 7900–7905. Retrieved 2019-09-15, from `http://www.pnas.org/lookup/doi/10.1073/pnas.1602413113` doi: 10.1073/pnas.1602413113

Gal, D., & Rucker, D. D. (2018). The loss of loss aversion: Will it loom larger than its gain? *Journal of Consumer Psychology*, *28*(3), 497–516.

Glatard, T., Lewis, L. B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., ... Evans, A. C. (2015). Reproducibility of neuroimaging analyses across operating systems. *Frontiers in Neuroinformatics*, *9*. Retrieved 2019-10-28, from `https://www.frontiersin.org/articles/10.3389/fninf.2015.00012/full` doi: 10.3389/fninf.2015.00012

Gronenschild, E. H. B. M., Habets, P., Jacobs, H. I. L., Mengelers, R., Rozendaal, N., van Os, J., & Marcelis, M. (2012, June). The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. *PLoS ONE*, *7*(6), e38234. Retrieved 2019-10-28, from `http://dx.plos.org/10.1371/journal.pone.0038234` doi: 10.1371/journal.pone.0038234

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the coase theorem. *Journal of political Economy*, *98*(6), 1325–1348.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives*, *5*(1), 193–206.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., ... Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*(2), 115.

Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D., & Nichols, T. E. (2009, April). Meta-analysis of neuroimaging data: A comparison of image-based and coordinate-based pooling of studies. *NeuroImage*, *45*(3), 810–823. Retrieved 2019-10-27, from `http://www.sciencedirect.com/science/article/pii/S1053811908012901` doi: 10.1016/j.neuroimage.2008.12.039

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007, January). The Neural Basis of Loss Aversion in Decision-Making Under Risk. *Science*, *315*(5811), 515–518. Retrieved 2019-10-26, from `http://www.sciencemag.org/cgi/doi/10.1126/science.1134239` doi: 10.1126/science.1134239

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*(4), 297–323.

Van Essen, D. C. (2005). A population-average, landmark-and surface-based (pals) atlas of human cerebral cortex. *Neuroimage*, *28*(3), 635–662.

Worsley, K. J., Evans, A. C., Marrett, S., & Neelin, P. (1992). A three-dimensional statistical analysis for cbf activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, *12*(6), 900–918.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, *8*(8), 665.