Commentary

Interpretation of published meta-analytical studies affected by implementation errors in the GingerALE software

Jane R Garrison[1], John Done[2], Jon S. Simons[1]

[1] Department of Psychology and Behavioural & Clinical Neuroscience Institute, University of Cambridge, UK

[2] School of Psychology, University of Hertfordshire, Hatfield, Hertfordshire, UK.

Correspondence should be addressed to Dr. Jane Garrison, Department of Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK. E-mail: jrg60@cam.ac.uk. Phone: +44 1223 333566. Fax: +44 1223 764760.

Acknowledgements: JRG was supported by Wellcome Trust Collaborative award made to Hearing the Voice (2). JSS was supported by a James S. McDonnell Foundation Scholar award.

GingerALE (http://brainmap.org/ale/) is a widely used, freely distributed software package used to undertake co-ordinate based activation likelihood estimation (ALE) meta-analysis of neuroimaging data. The developers of the software (Eickhoff, Laird, Fox, Lancaster, & Fox, 2017) have recently reported their discovery of two implementation errors which affected versions of the software prior to version 2.3.6 (released in April 2016). These errors, which have been discussed previously in *Neuroscience and Biobehavioral Reviews* (Tanasescu, Tench, Cottam, Constantinescu, & Auer, 2015; Tench, Tanasescu, Cottam, Constantinescu, & Auer, 2016) affected the multiple comparisons correction procedure resulting in the application of more liberal statistical thresholds than should have been the case. The first error, involving calculation of the threshold for the False Discovery Rate (FDR) correction, was amended in GingerALE V2.3.3 (May, 2015) but affected all earlier versions of the software. The second error, in the cluster-level Familywise Error (FWE) correction process dating from V2.2 (May 2012), was corrected in April 2016 in V2.3.6.

Several hundred published meta-analysis studies (http://www.brainmap.org/pubs/) have used versions of the GingerALE software affected by these errors. This number includes two neuroimaging meta-analyses by the present authors, published before the errors came to light (Garrison, Erdeniz, & Done, 2013; Zmigrod, Garrison, Carr, & Simons, 2016). The GingerALE developers have recommended that the authors of affected studies repeat their analyses with the latest version of the software, and compare their results with the original findings (Eickhoff et al., 2017). Consistent with a few other authors of studies that used versions of GingerALE now known to have been affected by these implementation errors (e.g. Smith & Delgado, 2017), we

have done this, and we summarise our findings below. We also use our experience to make suggestions for the interpretation of other published meta-analytical studies affected by the GingerALE software errors, and discuss the implications for interpreting statistical analyses more generally that may be affected by similar problems relating to the use of non-open-source, third party software products.

The implementation error in the GingerALE FDR code affected calculation of the statistical threshold for determining activation significance, meaning that clusters that would otherwise have been excluded were falsely shown to have achieved significance (Eickhoff et al., 2017). Importantly, this error did not affect the calculation of individual activation likelihood effect sizes, nor the application of the statistical threshold once it had been calculated. As such, reported uncorrected ALE p values calculated from the modelled activation maps are unaffected, as are the peak locations identified in the analysis, with the implementation error impacting only on which peaks were designated as being significantly above threshold (Eickhoff et al., 2017). However, the scale of the error is variable and dependent on the particular properties of the data, being affected by both the number of neuroimaging experiments in the dataset and the number of foci in each experiment: smaller datasets being typically more affected than larger ones (Eickhoff et al., 2016; M Fox. personal communication).

The effect of correcting this error on data from our two published ALE analyses was a large reduction in the number of clusters that exceeded the statistical threshold. Our first study, a meta-analysis of neuroimaging data relating to prediction error in reinforcement learning (Garrison et al., 2013), was based on a full dataset of 35

experiments and 445 foci. The significance threshold used, FDR correction with $p < .05$, implemented in GingerALE V2.1.1, pN (a conservative setting making no assumption about data correlation), and a minimum cluster size of $50mm^3$, had been chosen to mirror similar meta-analyses published a few years previously (e.g. Liu, Hairston, Schrier, & Fan, 2010). Re-analysis of the prediction error data revealed that for the top level 'All Studies' prediction error analysis, only four of the originally reported 33 activation peaks survived correction using these FDR settings when implemented in the corrected version of the software (GingerALE V2.3.6). The impact of the error on smaller datasets was similar, so for example only three activation peaks survived for the instrumental and reward analyses using these FDR settings (previously 21 peaks each). In light of current arguments that FDR may not, in any event, be an optimal correction method for ALE analyses (Eickhoff et al., 2016; Eickhoff, Bzdok, Laird, Kurth, & Fox, 2012), we further analysed the All-Studies prediction error data with GingerALE V2.3.6 using FWE voxel correction ($p < .05$), and cluster-level FWE correction (cluster-forming threshold of $p < .001$, cluster-level correction of $p < .05$) as recommended in the GingerALE manual (http://www.brainmap.org/ale/manual.pdf). Four activation peaks survived correction using FWE and five for cluster level correction.

The pattern of findings with our second hallucination meta-analysis (Zmigrod et al., 2016) was also marked. In this study, we compared neuroimaging data reporting brain activity during auditory verbal hallucinations (16 experiments, 236 foci) with that during visual hallucinations (7 experiments, 77 foci). FDR correction with a $p < .05$ threshold was used, implemented in GingerALE V2.3.2, pN, and $200mm^3$ minimum cluster size, chosen to mirror four earlier GingerALE meta-analyses of hallucination

data by other researchers (Jardri, Pouchet, Pins, & Thomas, 2011; Kompus, Westerhausen, & Hugdahl, 2011; Kühn & Gallinat, 2012; Van Lutterveld, Diederen, Koops, Begemann, & Sommer, 2013). No activation peaks exceeded this FDR threshold when implemented in the current version (V2.3.6) of the GingerALE software, for either auditory verbal hallucinations (originally 31) or visual hallucinations (10). The FWE $p < .05$ voxel-level threshold correction also resulted in no significant peaks for either analysis. Use of cluster-level FWE correction (cluster-forming threshold of $p < .001$, cluster-level correction of $p < .05$), as now recommended in GingerALE V2.3.6, resulted in three activation peaks designated as significantly above threshold for both the auditory and visual hallucination analyses. The underlying GingerALE text files for both the prediction error and hallucination meta-analyses will be made publicly available (https://doi.org/10.17863/CAM.15181) to enable interested parties to further explore the data, and we encourage authors of other affected GingerALE studies to do the same.

GingerALE Interpretation

These re-analyses suggest that while the locations of consistent activation peaks across neuroimaging studies were accurately identified in the original analyses, the designation of which peaks were significant was incorrect. This highlights an important issue in terms of sample sizes for coordinate based meta-analyses. It is likely that the re-analyses did not reproduce the earlier delineation of significant peaks using either the original FDR settings, or using voxel-wise FWE or cluster based thresholds, due to insufficient power based on the number of neuroimaging experiments available. This was the case even for the 'All Studies' prediction error analysis which utilised substantially more than the 17-20 minimum number of

experiments recommended by the GingerALE developers to ensure that meta-analysis results are not driven by a single experiment (Eickhoff et al., 2016). Notably, our hallucination analyses (Zmigrod et al., 2016) built upon the four earlier GingerALE meta-analyses by other researchers that utilised smaller datasets than our own (Jardri et al., 2011; Kompus et al., 2011; Kühn & Gallinat, 2012; Van Lutterveld et al., 2013), and which also used the FDR thresholding settings now known to be unreliable. This issue of sample size and its effect on power is important and should be borne in mind when interpreting published meta-analyses that used versions of the GingerALE software affected by the implementation errors. We echo Eickhoff et al. (2016)'s recommendation that data from smaller samples be re-analysed using a corrected version of the software to understand the extent to which the original results can be reproduced.

Despite the recommendations for reanalysis and communication above, it is likely that for a large number of meta-analyses, there will be no published assessment of the impact of the GingerALE thresholding errors. Eickhoff et al., (2017) point out in their discussion of the GingerALE software issue that unintended errors in reporting statistical thresholds do not necessarily invalidate the results and conclusions of published studies, as the choice of statistical threshold is an arbitrary and ultimately subjective decision. Many statisticians argue that p-values are a poor basis for making scientific inferences, and that effect sizes are more informative measures (Wasserstein & Lazar, 2016). As relative effect sizes (ALE values) and uncorrected p values are unaffected by the errors, these previous meta-analyses retain considerable value in identifying the degree to which brain regions were activated consistently across underlying experiments. This information can be the main interest for many readers of

meta-analyses who use them to identify the handful of regions that are most frequently associated with a cognitive function, rather than being solely concerned with the statistical significance of that frequency.

For example, one recent paper (Chen, Lambon Ralph, & Rogers, 2017) took exactly this descriptive approach in reporting the results of their GingerALE meta-analysis, with the emphasis on effect sizes, and associated p-values not reported. To explore this further, we reanalysed our own meta-analytical findings to calculate the frequency with which each of the reported clusters was observed in the underlying experimental papers. We defined a contribution as a focus of peak activity lying within 5mm of the reported GingerALE cluster. The cluster ALE value was very strongly correlated with the number of contributing studies ($r = .907$, $N = 48$, $p < .001$), suggesting that focusing on ALE effect size values can provide useful insight into the descriptive accuracy of the results, and can be used to aid interpretation of the results of previously published meta-analyses that are now known to be subject to software errors.

There is a broader issue here regarding the interpretation of statistical analyses that have employed other software packages, which could potentially be subject to similar errors. There is an overriding need to read critically and with an awareness of the possibility of error not only in the data, but in the analysis software. It is notable that there was not one, but two errors discovered in the GingerALE code, which appear to have been present across many versions of the software, affecting the results reported in a large number of published meta-analyses. There may be similar errors in other statistical analysis software packages that are, as yet, undiscovered. Understanding whether statistical results have been replicated using alternative software packages or

taking the opportunity to undertake such re-analysis oneself, as well as knowing whether the software code has been made open-source and been subject to some form of independent verification, may help to address the uncertainty attached to the results. In their discussion of the effect of the FWE cluster-level error on their own published meta-analysis data, Smith & Delgado (2016) called for the effective communication of implementation errors once discovered by software developers, and users can inform themselves further by reading on-line support forums to be aware at the earliest stage of issues that may arise. In summation, there is a clear need to promote openness, in making available source code, the underlying data, and provision of early and informed communication of issues whenever these arise.

References

Chen, L., Lambon Ralph, M. A., & Rogers, T. T. (2017). A unified model of human semantic knowledge and its disorders. *Nature Human Behaviour*, *1*, 39.

Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., & Fox, P. T. (2012). Activation likelihood estimation meta-analysis revisited. *NeuroImage*, *59*(3), 2349–2361.

Eickhoff, S. B., Laird, A. R., Fox, P. M., Lancaster, J. L., & Fox, P. T. (2017). Implementation errors in the GingerALE Software: Description and recommendations. *Human Brain Mapping*, *38*(1), 7–11.

Eickhoff, S. B., Nichols, T. E., Laird, A. R., Hoffstaedter, F., Amunts, K., Fox, P. T., … Eickhoff, C. R. (2016). Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. *NeuroImage*, *137*, 70–85.

Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *37*, 1297–1310.

Jardri, R., Pouchet, A., Pins, D., & Thomas, P. (2011). Cortical activations during auditory verbal hallucinations in schizophrenia: a coordinate-based meta-analysis. *The American Journal of Psychiatry*, *168*(1), 73–81.

Kompus, K., Westerhausen, R., & Hugdahl, K. (2011). The "paradoxical" engagement of the primary auditory cortex in patients with auditory verbal hallucinations: a meta-analysis of functional neuroimaging studies. *Neuropsychologia*, *49*(12), 3361–3369.

Kühn, S., & Gallinat, J. (2012). Quantitative meta-analysis on state and trait aspects of auditory verbal hallucinations in schizophrenia. *Schizophrenia Bulletin*, *38*(4), 779–786.

Liu, X., Hairston, J., Schrier, M., & Fan, J. (2010). Common and distinct networks underlying reward valence and processing stages: A meta-analysis of functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *35*(5), 1219–1236.

Smith, D. V., & Delgado, M. R. (2017). Meta-analysis of psychophysiological interactions: Revisiting cluster-level thresholding and sample sizes. *Human Brain Mapping*, *38*(1), 588–591.

Tanasescu, R., Tench, C. R., Cottam, W. J., Constantinescu, C. S., & Auer, D. P. (2015). Coordinate based meta-analysis does not show grey matter atrophy in narcolepsy: Commentary. *Neuroscience and Biobehavioral Reviews*, *57*, 297–298.

Tench, C. R., Tanasescu, R., Cottam, W. J., Constantinescu, C. S., & Auer, D. P. (2016). Coordinate based meta-analysis does not show grey matter atrophy in narcolepsy: Discussion. *Neuroscience and Biobehavioral Reviews*, (In Press).

Van Lutterveld, R., Diederen, K. M. J., Koops, S., Begemann, M. J. H., & Sommer, I. E. C. (2013). The influence of stimulus detection on activation patterns during auditory hallucinations. *Schizophrenia Research*, *145*(1–3), 27–32.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70(2)*, 129–133.

Zmigrod, L., Garrison, J. R., Carr, J., & Simons, J. S. (2016). The Neural Correlates of Hallucinations: A Quantitative Meta-Analysis of Neuroimaging Studies. *Neuroscience & Biobehavioural Reviews*, *69*, 113–123.