Implementation Errors in the GingerALE Software: Description and Recommendations

Simon B. Eickhoff,[1,2] Angela R. Laird,[3] P. Mickle Fox,[4] Jack L. Lancaster,[4,5] and Peter T. Fox[4,5,6]*

[1]*Institute of Neuroscience and Medicine (INM-1), Research Center Jülich, Germany*
[2]*Institute of Clinical Neuroscience and Medical Psychology, Heinrich-Heine University Düsseldorf, Germany*
[3]*Department of Physics, Florida International University, Miami, Florida*
[4]*Research Imaging Institute, University of Texas Health Science Center at San Antonio, Texas*
[5]*Department of Radiology, University of Texas Health Science Center at San Antonio, Florida*
[6]*South Texas Veterans Health Care System, San Antonio, Texas*

Abstract: Neuroscience imaging is a burgeoning, highly sophisticated field the growth of which has been fostered by grant-funded, freely distributed software libraries that perform voxel-wise analyses in anatomically standardized three-dimensional space on multi-subject, whole-brain, primary datasets. Despite the ongoing advances made using these non-commercial computational tools, the replicability of individual studies is an acknowledged limitation. Coordinate-based meta-analysis offers a practical solution to this limitation and, consequently, plays an important role in filtering and consolidating the enormous corpus of functional and structural neuroimaging results reported in the peer-reviewed literature. In both primary data and meta-analytic neuroimaging analyses, correction for multiple comparisons is a complex but critical step for ensuring statistical rigor. Reports of errors in multiple-comparison corrections in primary-data analyses have recently appeared. Here, we report two such errors in GingerALE, a widely used, US National Institutes of Health (NIH)-funded, freely distributed software package for coordinate-based meta-analysis. These errors have given rise to published reports with more liberal statistical inferences than were specified by the authors. The intent of this technical report is threefold. First, we inform authors who used GingerALE of these errors so that they can take appropriate actions including re-analyses and corrective publications. Second, we seek to exemplify and promote an open approach to error management. Third, we discuss the implications of these and similar errors in a scientific environment dependent on third-party software. *Hum Brain Mapp 00:000–000, 2016*.

Key words: fMRI; statistics; false positives; cluster inference; meta-analysis

Conflicts of Interest: All authors are co-developers of the BrainMap software suite, including the GingerALE application which is discussed herein. Peter Fox and Jack Lancaster are Editors-in-Chief of the journal *Human Brain Mapping*.

Contract grant sponsor: US National Institute of Mental Health; Contract grant number: RO1-MH074457 (to P.T.F.); Contract grant sponsor: Helmholtz Portfolio Theme "Supercomputing and Modelling for the Human Brain" and the European Union Seventh Framework Program; Contract grant number: FP7/2007-2013 under grant agreement no. 604102.

*Correspondence to: Peter T. Fox, 8403 Floyd Curl Drive, San Antonio, Texas 78229. E-mail: fox@uthscsa.edu

Received for publication 11 July 2016; Accepted 29 July 2016.

DOI: 10.1002/hbm.23342

Published online 00 Month 2016 in Wiley Online Library (wileyonlinelibrary.com).

INTRODUCTION

Human neuroscience imaging—as distinguished from clinical, diagnostic imaging—most commonly uses noninvasive, tomographic, whole-brain, image-acquisition modalities (e.g., magnetic resonance imaging, positron emission tomography, and single-photon emission tomography) and grant-funded, non-commercial software to make inferences regarding the structural and functional organization of the human brain in development, in adulthood, in aging, and in a wide variety of neurologic, psychiatric and systemic conditions in an ongoing and programmatic manner [Bandettini, 2012; Rosen and Savoy, 2012]. Despite the impressive power of the neurocomputational techniques shared freely in this field, there are notable limitations. In particular, the generalizability of the information that can be gleaned from a single neuroimaging study is necessarily limited both in reporting differences in activation patterns between task conditions and in reporting differences in grey-matter volume between subject groups [Weinberger and Radulescu, 2015]. Factors contributing to these limitations include sample size (small samples having lower power and higher potential for biased sampling than large samples), an extraordinary degree of experimental-design flexibility and analytic flexibility (both permitting substantive methodological variations between studies apparently reporting on the same effect in the same condition), and the indirect nature of the neuroimaging measures used vis-à-vis the inferred neuronal physiology and pathology [Button et al., 2013, Carp, 2012; Glatard et al., 2015; Rottschy et al., 2013]. Correction for multiple comparisons of datasets representing the brain by hundreds of thousands of voxels (i.e., individual, location-specific data samples) is a complex but critical step for ensuring statistical rigor, but one that has proven particularly problematic. When combined with publication bias (suppression of negative results) and an all-too-common tendency toward overly enthusiastic interpretations of the significance of individual primary-data reports, these factors necessarily foster concerns regarding the reproducibility of neuroimaging results that are similar in import to those voiced in the psychological sciences [Open Science Collaboration, 2015].

Coordinate-based meta-analysis offers a powerful remedy for the lack of generalizability potentially impacting any individual neuroimaging study. The vast majority of the neuroscience imaging literature—several tens of thousands of peer-reviewed publications—uses anatomically standardized stereotaxic space (*x-y-z* coordinates referenced to a published anatomical template) as a framework within which results are computed and reported, typically as local maxima of significant statistical contracts. This standard has been employed since the inception of the field [Fox et al., 1988; Fox and Mintun, 1989; Friston et al., 1991], and its impact has been repeatedly reviewed [Fox, 1995; Fox, Parsons and Lancaster, 1998; Fox et al., 2014]. When sufficiently large subsets of this literature are combined using rigorous selection criteria and appropriate statistical methods, robust insights into the functional and structural organization of the human brain and its disease processes can be obtained [Yarkoni et al., 2010; Crossley, Fox and Bullmore, 2016; Eickhoff and Etkin, 2016]. As with primary-data analyses, coordinate-based meta-analyses are performed voxel-wise over the entire brain and also apply corrections for multiple comparisons.

Activation Likelihood Estimation (ALE) was one of the first algorithms developed for coordinate-based meta-analysis [Turkeltaub et al., 2002] and remains one of the most widely used (http://brainmap.org/pubs). A core concept of the ALE algorithm is to model reported *x-y-z* addresses as centroids of 3-D Gaussian probability distributions, thereby accommodating the spatial uncertainty of neuroimaging findings caused jointly by inter-individual neuroanatomical variability and the intrinsic signal-to-noise and spatial-resolution limitations of non-invasive neuroimaging modalities. Since its introduction, ALE has benefitted from a series of functional enhancements, most notably for present purposes, in its corrections for multiple comparisons. The original implementation of ALE applied no correction for multiple comparisons [Turkeltaub et al., 2002]. Corrections based on false-discovery rate [FDR, Laird et al., 2005] and on cluster-level and voxel-level family-wise error (FWE) estimation [Eickhoff et al., 2012] were subsequently added. Other developments include the replacement of the initial fixed-effects modeling with random-effects analyses of convergence over experiments rather than individual foci [Eickhoff et al., 2009] and a correction to avoid summation of within-group effects [Turkeltaub et al., 2012]. In addition, algorithms have been provided for meta-analytic contrast analyses using fixed-effects [Laird et al., 2005] and random-effects [Eickhoff et al., 2011] models.

The most widely used implementation of the ALE algorithm is GingerALE, a software application distributed as part of the BrainMap meta-analysis environment and software suite [Fox and Lancaster, 2002; Laird et al., 2009, 2011; Fox et al., 2014; http://brainmap.org/ale]. GingerALE has included FDR multiple-comparison correction since V1.0, and has included voxel- and cluster-level FWE correction since V2.2. Implementation errors in FDR were first suspected in May, 2015, when inconsistencies were noted in the output of large-scale, replication simulations performed by a member of the BrainMap user community and reported to the BrainMap development team. The source of the inconsistencies was identified rapidly, and a new build (V.2.3.3) was released within weeks. The error in the FWE correction was first suspected in January, 2016, also via a report from a BrainMap user-community member. This error was confirmed, identified and corrected with a new build (V2.3.6) released in April, 2016. Both errors and their corrections were described on the BrainMap online forum (http://www.brainmap.org/forum).

Abbreviations

ALE Activation likelihood estimation;
FEW Family-wise error;
NIH National Institutes of Health

Posting errors in this manner is common practice among software developers in the field and this transparency is to be commended. However, this valuable information is poorly discoverable and cannot easily be cited by the users when writing up their findings. In the following, we describe these implementation errors and their potential impact; we make recommendations for corrective actions; and, we discuss this meta-analysis-specific situation in the larger context of current neuroimaging research, suggesting potential future management strategies.

ERROR IN THE FDR CORRECTION CODE

FDR thresholding is designed to control the expected proportion of errors among rejected hypotheses, i.e., false discoveries. GingerALE's implementation of FDR uses the Benjamini–Hochberg procedure, which starts by converting the 3-D *P*-value image into a sorted 1-D array of ascending *P* values. The sorted *P* values are then compared in a step-up fashion against a boundary criterion depending on the overall number of parallel tests and assumptions regarding independence. Critically, a small mistake in the customized code for sorting floating-point numbers (*P* values) has persisted until GingerALE 2.3.3. As a result of this error, the *P* values were not completely sorted, leaving some high *P* values distributed through the lower *P* values at the beginning of the sorted vector. This error right-shifts the "observed" *P* values relative to the line setting the boundary criterion and allows *P* values that should have been above the cut-off to remain underneath it. That is, the effective threshold became too lenient and did not fully control the FDR at the desired level.

The impact of this error on FDR-corrected inference in GingerALE is heterogeneous and dataset specific because, in FDR, the corrected significance of a particular location depends on the overall shape of the curve of sorted *P* values. Also contributing to the variability of the effect, the magnitude of the sorting error depends on the dataset, in particular the distribution of *P* values therein and their spatial location, i.e., initial indexing. Ultimately, the potential impact of this coding error is highly dependent on the properties of the individual study, though it will almost inevitably lead to thresholds that are too liberal. Generally this will mean that with re-analysis the observed cluster sizes will be smaller than previously reported and that smaller clusters may not reach significance. Actual ALE scores and peak locations locations should be unaffected.

ERROR IN THE CLUSTER-LEVEL CORRECTION CODE

Cluster-level FWE thresholding is designed to apply a "cluster-forming threshold" (typically and standard in GingerALE: $P < 0.001$), and then compare the size of the individual clusters in this excursion set to a distribution of cluster sizes arising from the same initial threshold under a null-hypothesis of random spatial location. In the non-parametric, Monte-Carlo approach for establishing this null-distribution in the context of ALE, foci are randomly distributed throughout the brain followed by application of the cluster-forming threshold. The size of the largest cluster is recorded, and the procedure repeated many thousands of times. By removing clusters in the actual excursion set that are smaller than the top 5% of the recorded values, the cluster-level FWE is controlled given that only 5% of all random realizations of the null-hypothesis will entail one or more clusters larger than the ones that were deemed significant.

Cluster-level FWE thresholding was introduced into GingerALE in V2.2 and, unfortunately, the procedure for establishing the null-distribution of cluster-sizes through V2.3.5 contained a small but important error. Rather than recording the size of the largest cluster in the excursion set, versions of GingerALE before V2.3.6 recorded all cluster sizes following application of the cluster-forming threshold on the data generated under the null-hypothesis. This approach yielded thresholds that did not control the FWE of the clusters, but rather resulted in inference based on uncorrected cluster-level *P* values. While these are substantially more conservative than uncorrected *P* values at the voxel-level given the two-step inference and initial cluster-forming threshold, the use of uncorrected cluster-level *P* values still resulted in inadequately liberal inference. The overall effect will be that some smaller clusters will not reach significance.

SPECIFIC RECOMMENDATION FOR STUDIES USING THE AFFECTED VERSIONS OF GINGERALE

We recommend that published meta-analyses using the GingerALE versions with implementation errors in the multiple-comparisons corrections be repeated using the latest version of GingerALE (V2.3.6), and the results compared to those of the original report. Depending upon the magnitude and potential impact of the differences, authors should consider corrective communications in consultation with the journal in which their original report appeared, as discussed below.

When weighing their course-of-action options, we suggest authors consider the argument that unintended errors in reporting statistical thresholds do not necessarily invalidate the results and conclusions of their published studies. Choice of a statistical threshold and the ensuing trade-off between type-I and type-II errors is, at base, an arbitrary and ultimately subjective decision [Lieberman and Cunningham, 2009]. On the other hand, readers should expect to receive correct information about the statistical thresholds applied.

We also note that a case can be made that even correctly performed voxel-wise FDR correction may be inappropriate for inferences on topological features such as regions of significant convergence of a smooth dataset [Chumbley and Friston, 2009]. This shortcoming of FDR was recently

confirmed in a large-scale simulation study [Eickhoff and Etkin, 2016], which demonstrated that voxel-level FDR correction entails both relatively low sensitivity and a high susceptibility to false-positive findings. Moreover, that work also highlighted another negative property of FDR thresholding, namely that the chance of a voxel being declared significant depends on the strength of convergence in other parts of the brain [Genovese et al., 2002]. For maximal statistical rigor, FWE thresholding should be used for ALE analyses in preference to FDR. Further, to have sufficient power to detect moderately sized effects, ALE analyses should be based on workspaces containing 17-20 experiments or more [Eickhoff and Etkin, 2016].

GENERAL CONSIDERATIONS ON THE EFFECTS OF ERRORS IN NEUROIMAGING SOFTWARE

Fully automated, voxel-wise, whole-brain, image-analysis methods concurrently analyzing data from multiple subjects in anatomically standardized 3-D arrays were first introduced more than twenty-five years ago [Fox et al., 1988; Friston et al., 1991]. As these statistical parametric imaging methods have advanced in sophistication, standardization, ease-of-use, and community acceptance, they have largely supplanted user-scripted tools. Following a "survival of the fittest" evolutionary process, the vast majority of neuroimaging researchers now rely on a limited number of grant-supported, freely distributed, non-commercial software libraries with SPM [Ashburner, 2012], FSL [Jenkinson et al., 2012] and AFNI [Cox, 2012] being among the most popular. While the wide scope of use of these packages, inviting scrutiny and cross validation by many researchers, will eventually detect and eliminate errors [Nosek et al., 2015], the impact of as-yet-undetected errors on the published literature can be substantial. This predicament is best illustrated by a recent study that identified a problem in the multiple-comparison correction strategies implemented in several widely used packages, which collectively effect several thousand peer-reviewed neuroimaging publications [Eklund et al., 2016].

Implementation errors (reported here) and algorithmic errors [Eklund et al., 2016] in widely used image-analysis software creates the unfortunate situation wherein well-intentioned researchers who have followed developers' recommendations and established best practices may still have published flawed results—typically erroneous statistical confidence levels or cluster sizes. To best serve the neuroscientific community, corrections to the literature should be two-fold. First, the software developer should highlight the errors and need for re-analysis, as we are doing here. Second, the authors should be encouraged and enabled to self-correct such errors in a concise, rapidly implemented, non-pejorative manner. Given that the magnitude and impact of the errors will vary, the most appropriate self-correction measure will also vary. For minimal corrections, a comment on PubMed Central confirming the previous results should suffice. For minor corrections, publication of an *erratum* or *corrigendum* linked to the original publication may be needed. For more substantive corrections, a Comment-type article citing the original publication likely is the appropriate course of action. For older or underpowered meta-analyses, particularly in domains for which additional publications have appeared in the interim, a more comprehensive, original publication will likely be the most valuable contribution to the literature.

In light not only of the present error report but also given the wider implications of the topic as noted above, we believe there is a need for dialogue among journal editors, scientific organizations (e.g., the Organization for Human Brain Mapping), and the neuroscience community at large to develop a generally acceptable best-practices policy. We hope that this article encourages the more open reporting of errors in public software or data and also serves as a starting point for this important dialogue.

REFERENCES

Ashburner J (2012): SPM: A history. Neuroimage 201262:791–800.

Bandettini PA (2012): Twenty years of functional MRI: The science and the stories. Neuroimage 62:575–588.

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafo MR (2013): Power failure: Why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376.

Carp J (2012): On the plurality of (methodological) worlds: Estimating the analytic flexibility of FMRI experiments. Front Neurosci 6:149.

Chumbley JR, Friston KJ (2009): False discovery rate revisited: FDR and topological inference using Gaussian random fields. Neuroimage 44:62–70.

Cox RW (2012): AFNI: What a long strange trip it's been. Neuroimage 15;62:743–747.

Crossley NA, Fox PT, Bullmore E (2016): Meta-connectomics: Human brain network and connectivity meta-analyses. Psychol Med 46:897–907.

Eickhoff SB, Etkin A (2016): Going beyond finding the "Lesion": A path for maturation of neuroimaging. Am J Psychiatry 173: 302–303.

Eickhoff SB, Laird AR, Grefkes C, Wang LE, Zilles K, Fox PT (2009): Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. Hum Brain Mapp 30:2907–2926.

Eickhoff SB, Bzdok D, Laird AR, Roski C, Caspers S, Zilles K, Fox PT (2011): Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation. Neuroimage 57:938–949.

Eickhoff SB, Bzdok D, Laird AR, Kurth F, Fox PT (2012): Activation likelihood estimation meta-analysis revisited. Neuroimage 59:2349–2361.

Eickhoff SB, Nichols TE, Laird AR, Hoffstaedter F, Amunts K, Fox PT, Bzdok D, Eickhoff CR (2016a): Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. Neuroimage 137:70–85.

Eickhoff S, Nichols TE, Van Horn JD, Turner JA (2016b): Sharing the wealth: Neuroimaging data repositories. Neuroimage 124: 1065–1068. 569.

Eklund A, Nichols TE, Knutsson H (2016): Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci U S A 113:7900–7905.

Fox PT (1995): Spatial normalization origins: Objectives, applications, and alternatives. Hum Brain Mapp 3:161–164.

Fox PT, Mintun MA (1989): Noninvasive functional brain mapping by change-distribution analysis of average PET images of $H_2^{15}O$ tissue activity. J Nucl Med 30:141–149.

Fox PT, Lancaster JL (2002): Opinion: Mapping context and content: The BrainMap model. Nat Rev Neurosci 3:574 319–321.

Fox PT, Mintun MA, Reiman EM, Raichle ME (1988): Enhanced detection of focal brain responses using intersubject averaging and change-distribution analysis of subtracted PET images. J Cereb Blood Flow Metab 8:642–653.

Fox PT, Parsons LM, Lancaster JL (1998): Beyond the single study: Function/location meta-analysis in cognitive neuroimaging. Curr. Opin. Neurobiol 8:178–187.

Fox PT, Lancaster JL, Laird AR, Eickhoff SB (2014): Meta-analysis in human neuroimaging: Computational modeling of large-scale databases. Annu Rev Neurosci 37:409–434.

Friston KJ, Frith CD, Liddle PF, Frackowiak RSJ (1991): Comparing functional (PET) images: The assessment of significant change. J Cereb Blood Flow Metab 11:690–699.

Genovese CR, Lazar NA, Nichols T (2002): Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15:870–878.

Glatard T, Lewis LB, Ferreira da Silva R, Adalat R, Beck N, Lepage C, Rioux P, Rousseau ME, Sherif T, Deelman E, Khalili-Mahani N, Evans AC (2015): Reproducibility of neuroimaging analyses across operating systems. Front Neuroinform 9:12.

Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM (2012): FSL. Neuroimage 62:782–790.

Laird AR, Fox PM, Price CJ, Glahn DC, Uecker AM, Lancaster JL, Turkeltaub PE, Kochunov P, Fox P (2005): ALE meta-analysis: Controlling the false discovery rate and performing statistical contrasts. Hum Brain Mapp 25:155–164.

Laird AR, Eickhoff SB, Kurth F, Fox PM, Uecker AM, Turner JA, Robinson JL, Lancaster JL, Fox PT (2009): ALE meta-analysis workflows via the brainmap database: Progress towards a probabilistic functional brain atlas. Front Neuroinform 3:23.

Laird AR, Eickhoff SB, Fox PM, Uecker AM, Ray KL, Saenz JJ, Jr, McKay DR, Bzdok D, Laird RW, Robinson JL, Turner JA, Turkeltaub PE, Lancaster JL, Fox PT (2011): The BrainMap strategy for standardization, sharing, and meta-analysis of neuroimaging data. BMC Res Notes 4:349.

Lieberman MD, Cunningham WA (2009): Type I and Type II error concerns in fMRI research: Re-balancing the scale. J Soc Cogn Affect Neurosci 4:423–428.

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishiyama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon TA, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Paluck EL, Simonsohn U, Soderberg C, Spellman BA, Turitto J, VandenBos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T (2015): Promoting an open research culture. Science 348:1422–1425.

Open Science Collaboration (2015): Estimating the reproducibility of psychological science. Science 349:aac4716.

Rosen BR, Savoy RL (2012): fMRI at 20: Has it changed the world? Neuroimage 62:1316–1324.

Rottschy C, Caspers S, Roski C, Reetz K, Dogan I, Schulz JB, Zilles K, Laird AR, Fox PT, Eickhoff SB (2013): Differentiated parietal connectivity of frontal regions for "what" and "where" memory. Brain Struct Funct 218:1551–1567.

Turkeltaub PE, Eden GF, Jones KM, Zeffiro TA (2002): Meta-analysis of the functional neuroanatomy of single-word reading: Method and validation. Neuroimage 16:765–780.

Turkeltaub PE, Eickhoff SB, Laird AR, Fox PM, Wiener M, Fox PT (2012): Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. Hum Brain Mapp 33:1–13.

Weinberger DR, Radulescu E (2015): Finding the elusive psychiatric "Lesion" with 21st-century neuroanatomy: A note of caution. Am J Psychiatry173:27–33.

Yarkoni T, Poldrack RA, Van Essen DC, Wager TD (2010): Cognitive neuroscience 2.0: Building a cumulative science of human brain function. Trends Cogn Sci 14:489–496.