NeuroImage 137 (2016) 70–85

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation

Simon B. Eickhoff [a,b,*], Thomas E. Nichols [c], Angela R. Laird [d], Felix Hoffstaedter [a,b], Katrin Amunts [a,e], Peter T. Fox [f], Danilo Bzdok [g,h,i,1], Claudia R. Eickhoff [a,g,1]

[a] Institute for Neuroscience and Medicine (INM-1), Research Center Jülich, Germany
[b] Institute of Clinical Neuroscience and Medical Psychology, Heinrich-Heine University Düsseldorf, Germany
[c] Department of Statistics and Warwick Manufacturing Group, University of Warwick, Coventry, UK
[d] Department of Physics, Florida International University, USA
[e] C & O Vogt Institute for Brain Research, Heinrich-Heine University Düsseldorf, Germany
[f] Research Imaging Institute, University of Texas Health Science Center at San Antonio, USA
[g] Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University Hospital, Germany
[h] JARA, Translational Brain Medicine, Aachen, Germany
[i] Parietal Team, INRIA, Neurospin, bat 145, CEA Saclay, 91191 Gif-sur-Yvette, France

ARTICLE INFO

Article history:
Received 27 January 2016
Revised 14 March 2016
Accepted 1 April 2016
Available online 11 May 2016

ABSTRACT

Given the increasing number of neuroimaging publications, the automated knowledge extraction on brain-behavior associations by quantitative meta-analyses has become a highly important and rapidly growing field of research. Among several methods to perform coordinate-based neuroimaging meta-analyses, Activation Likelihood Estimation (ALE) has been widely adopted. In this paper, we addressed two pressing questions related to ALE meta-analysis: i) Which thresholding method is most appropriate to perform statistical inference? ii) Which sample size, i.e., number of experiments, is needed to perform robust meta-analyses? We provided quantitative answers to these questions by simulating more than 120,000 meta-analysis datasets using empirical parameters (i.e., number of subjects, number of reported foci, distribution of activation foci) derived from the BrainMap database. This allowed to characterize the behavior of ALE analyses, to derive first power estimates for neuroimaging meta-analyses, and to thus formulate recommendations for future ALE studies. We could show as a first consequence that cluster-level family-wise error (FWE) correction represents the most appropriate method for statistical inference, while voxel-level FWE correction is valid but more conservative. In contrast, uncorrected inference and false-discovery rate correction should be avoided. As a second consequence, researchers should aim to include at least 20 experiments into an ALE meta-analysis to achieve sufficient power for moderate effects. We would like to note, though, that these calculations and recommendations are specific to ALE and may not be extrapolated to other approaches for (neuroimaging) meta-analysis.

Introduction

For more than two decades, functional imaging using Positron-Emission Tomography (PET) and in particular functional Magnetic Resonance Imaging (fMRI) have provided ample information about the location of cognitive, sensory and motor processes in the human brain (Bandettini, 2012; Poldrack, 2012; Rosen and Savoy, 2012). Neuroimaging clinical populations have yielded unprecedented insight into the localization of structural and functional aberrations within the brains of patients with all kinds of neurological and psychiatric disorders (Bullmore, 2012). In spite of its undisputed success, neuroimaging findings carry several important limitations (Weinberger and Radulescu, 2015). The number of participants sampled by neuroimaging experiments have substantially increased since the late 1990s when fewer than a dozen participants were the norm, but are still comparably small relative to other fields of biological and medical sciences (Button et al., 2013). Hence, results from small-data studies are likely to be highly variable. Additionally, the measured signals are indirect hemodynamic proxies rather than direct measures of neuronal activity (Logothetis and Wandell, 2004). Furthermore, the substantial analytic flexibility of neuroimaging pipelines also leads to low reliability and reproducibility of neuroimaging findings (Carp, 2012b; Glatard et al., 2015; Wager et al., 2009). The logistic expenses of neuroimaging studies usually discourage the performance of confirmatory or supplementary experiments. This results in publication of

* Corresponding author at: Institute of Neuroscience and Medicine (INM-1), Research Center Jülich, Leo-Brandt Str. 5, 52425 Jülich, Germany
E-mail address: S.Eickhoff@fz-juelich.de (S.B. Eickhoff).
[1] Equal contributions.

http://dx.doi.org/10.1016/j.neuroimage.2016.04.072
1053-8119/

isolated findings, further aggravating the problem of low reproducibility and false positive findings. Finally, most current neuroimaging studies, more specifically their discussion, may be considered overgeneralizations of context-specific findings. Each investigation into the neuronal correlates of psychological phenomena or pathological states requires operationalization of the experiment, involving choices on the nature of psychological task, the displayed stimuli, the timing and arrangement of individual trials, their modeling in the statistical analysis and the assessed contrast (Carp, 2012a; Rottschy et al., 2012). As an important consequence, the ensuing coordinates of significant effects are conditioned by a large number of subjective selections by the investigator. Yet, they are discussed and cited as neuronal correlates representative of general psychological domains (e.g., working memory) or pathophysiological mechanisms (e.g., dysfunctional working memory in schizophrenia). It may hence come as no surprise that brain regions reported to be associated with a given psychological process may be found almost anywhere in the brain (Cieslik et al., 2015; Nee et al., 2007).

These circumstances have severely limited the knowledge on brain organization that can be gained from individual neuroimaging studies. This can however be compensated by the sheer number of neuroimaging experiments that have been published over the last decades. Derfuss and Mar. estimated that until 2007 almost 8000 neuroimaging studies had been published (Derrfuss and Mar, 2009). In autumn 2015, a similar PubMed search found more than 21,000 fMRI and PET studies in addition to ~3000 morphometric MRI papers. Searches for studies investigating major brain disorders such as schizophrenia, depression, dementia, stroke or Parkinson's disease likewise revealed many hundred clinical neuroimaging papers. The systematic integration of dispersed findings from previous studies - across different populations and experimental variations — may thus overcome the limitations raised above. They provide robust insights into the location of psychological and pathological effects (Kober and Wager, 2010; Nickl-Jockschat et al., 2015; Radua and Mataix-Cols, 2012; Schilbach et al., 2012; Yarkoni et al., 2011). Such structured integration is enabled by community-wide standards of spatial normalization (Amunts et al., 2014; Evans et al., 2012; Mazziotta et al., 2001) and the convention of reporting peak locations in stereotaxic reference spaces.

Activation likelihood estimation (ALE) has originally been introduced more than a decade ago (Turkeltaub et al., 2002). It is among the most commonly used algorithms for coordinate-based meta-analyses and it is part of the BrainMap software suite (Laird et al., 2009, 2011a) (http://brainmap.org/ale). The key idea behind ALE is to not treat activation foci reported in neuroimaging studies as dimensionless points but as spatial probability distributions centered at given coordinates. This approach thus accommodates the spatial uncertainty associated with neuroimaging findings by using the reported coordinates as the best point estimator but at the same time employing a (Gaussian) spatial variance model (Eickhoff et al., 2009). ALE maps are then obtained by computing the union of activation probabilities across experiments for each voxel (Turkeltaub et al., 2012). Finally, true convergence of foci is distinguished from random clustering of foci (i.e., noise) by testing against the null-hypothesis of random spatial association between experiments (Eickhoff et al., 2012; Turkeltaub et al., 2002).

In addition to previously frequent but now generally discouraged reports of uncorrected p-values, it is possible to account for multiple comparisons in the whole-brain setting by three approaches:

a) Family-wise error correction on the voxel level controls the chance of observing a given Z-value if foci were randomly distributed (Eickhoff et al., 2012).
b) Cluster-level family-wise error correction involves the use of an uncorrected cluster-forming threshold and employing a cluster-extent threshold that controls the chance of observing a cluster of that size if foci were randomly distributed (Eickhoff et al., 2012; Woo et al., 2014).
c) Correction for multiple comparisons using the false-discovery rate (Laird et al., 2005; Wager et al., 2007) has also been frequently employed, although this method may be less appropriate for neuroimaging data (Chumbley and Friston, 2009).

Another question that frequently comes up when planning, discussing or reviewing meta-analysis projects pertains to the necessary sample size, i.e., number of experiments that is needed to perform robust meta-analyses. In our view, this question may be investigated along two different aspects: First, is there a lower bound such that meta-analyses involving fewer experiments may not be considered valid? Second, can we provide power-calculations for ALE meta-analyses?

The present investigation addresses these questions by means of massive simulations of ALE analyses. These systematically vary the overall number of experiments and that of experiments activating the simulated "true" location. Four types of significance testing are then performed on the more than 120,000 simulated ALE analyses: voxel-level family-wise error correction *[FWE]*; cluster-level family-wise error correction *[cFWE]*; voxel-level false discovery rate correction *[FDR];* and uncorrected thresholding at $p < 0.001$ with an additional extent threshold of 200 mm^3. This allowed us to extend the scope of this work to a comparison of the influence of these methods on sensitivity, ensuing cluster sizes, the number of incidental clusters and statistical power.

In summary, the presented empirical simulations should provide a comprehensive overview on the behavior of ALE meta-analyses in addition to answering the key question, i.e., "how many experiments are needed to perform ALE?"

Methods & results

Experimental and conceptual setting

To avoid the ambiguous word "study", there is a preference for the terms "paper" for an entire published scientific work and "experiment" for the individual comparisons reported therein. Even though the present simulation does not rely on the paper unit but on sets of distinct experiments to be assessed for convergence, we retain the general terminology for the sake of clarity.

For all performed simulations, we selected the ground-truth location (i.e., the hypothetical "true" location of the effect of interest) to be at MNI coordinates −42/32/26, situated in the posterior part of the left middle frontal gyrus, without any particular motivation. Each simulation then entailed a variable number of experiments among the total numbers of experiments in that particular simulation, which activated at a location around this coordinate. That is, each simulation contained a certain number of experiments that featured activation foci in the target region, but importantly, these activation foci were not all located precisely at the target-location but rather randomly spread through the vicinity of this coordinate. This spatial spread captures the random variations of reported neuroimaging activation findings on a particular topic around a supposed true location of the underlying effect. For each combination of "total number of experiments" and "experiments activating the target-location", we generated 500 simulations using empirically derived spread parameters as outlined next.

One of the challenges when constructing a realistic simulation of ALE meta-analyses is to use simulation parameters that realistically reflect the distance of reported coordinates what may be considered the true location. Importantly, this spread is distinct from the spatial uncertainty associated with each individual focus in an ALE analysis. The latter describes how much the results for a single neuroimaging experiment may vary depending on between-subject and between-template variability, while the former describes the dispersion of reported foci from different experiments assessing the same effect. In order to provide empirical estimates on this spread that will then be used to construct the simulations, we assessed two pools of meta-analyses.

The first pool contained 15 hand-coded datasets (Bzdok et al., 2012; Cieslik et al., 2015; Hardwick et al., 2013; Kohn et al., 2014; Rottschy et al., 2012). The second pool consisted of 105 datasets that were automatically extracted from the BrainMap database (www.brainmap.org; (Fox et al., 2014; Laird et al., 2009, 2011a)). This was achieved by combinations of the paradigm class and behavioral domain classifications (Laird et al., 2011a) that yielded between 30 and 200 experiments. For each of the ensuing meta-analysis datasets, we performed an ALE meta-analyses and subsequently identified the peaks of the ensuing Z-map (thresholded equivalent to $p < 0.001$ uncorrected). In order to avoid assessment of multiple sub-peaks within a larger cluster, we ranked the Z-encoded activation peaks by significance and started selection from the top, eliminating all lower peaks within 2 cm distance of the selected peaks. Next, we identified all experiments from the respective dataset that contributed to the respective peak location. Here contribution was defined by truncating the Gaussian probability distribution of each experiment at 90% mass (i.e., its 3D confidence ellipsoid) and considering those experiments to contribute for which the peak was located within this Gaussian distribution with truncated tails. The distance between the coordinates for the contributing foci and the peak were then recorded and the sigma of the spatial spread (under isotropic Gaussian assumptions) computed from these. That is, we treated the peaks of convergence obtained from ALE analyses of the different datasets as the best estimators of the true effect and quantified the spatial spread around these.

As illustrated in Fig. 1, the average of the standard deviation sigma reflecting the spread of contributing foci to a peak location was just over 3.5 mm, corresponding to a FWHM of ~9.5 mm. In spite of the differences between both pools of datasets, our assessment furthermore showed that for the vast majority of peaks from both hand-coded and automatically generated meta-analyses, the contributing experiments feature a spread between 3 and 4 mm. Small numbers of peaks showed spreads up to 2 and 5 mm, respectively. Importantly, these findings seem independent of the number of experiments that constitute the respective meta-analyses. Consequently, we configured the 500 simulations for each combination of "total number of experiments" and "experiments activating the target-location" to used spreads (i.e., sigma parameter of the generated Gaussian distributions) of 2 mm (50 times), 3 mm (200 times), 4 mm (200 times) and 5 mm (50 times) around the true location of the synthetic activation blobs. In this context, we would like to acknowledge, that both the computation of the spatial spread as well as the generation of the simulations are based on the assumption of stationary Gaussian distributions, i.e., the same assumptions that also underlie ALE. This is not only a strong assumption but necessary in the absence of voxel- wise empirical data on spatial uncertainty, but may also yield to somewhat optimistic estimates of sensitivity and power.

Experimental details of the simulations

In this study, we performed ALE analyses on simulated datasets containing between 5 and 30 total experiments. Among these, between 1 and the minimum of 10 and the total number of experiments were randomly chosen to feature activation at the target-location. This yielded 245 distinct combinations of "total number of experiments" and "experiments activating the target-location", i.e., distinct cases of meta-analyses. For each of these cases, we then generated 500 random datasets, yielding in total 122,500 datasets that provided the basis for the systematic assessment of the behavior of the ALE algorithm and the different methods for statistical thresholding.

To create artificial datasets that are as realistic as possible, we used the BrainMap database to provide a distribution of the properties, in particular, the number of subjects as well as the number and location of foci, based on the current neuroimaging literature (www.brainmap.org). Importantly, this distribution was solely based on those experiments in BrainMap that report foci from healthy adults (i.e., studies with pathological populations or children were excluded) and were coded as "normal mapping studies" (i.e., intervention studies and group comparisons were excluded). Moreover, we restricted the

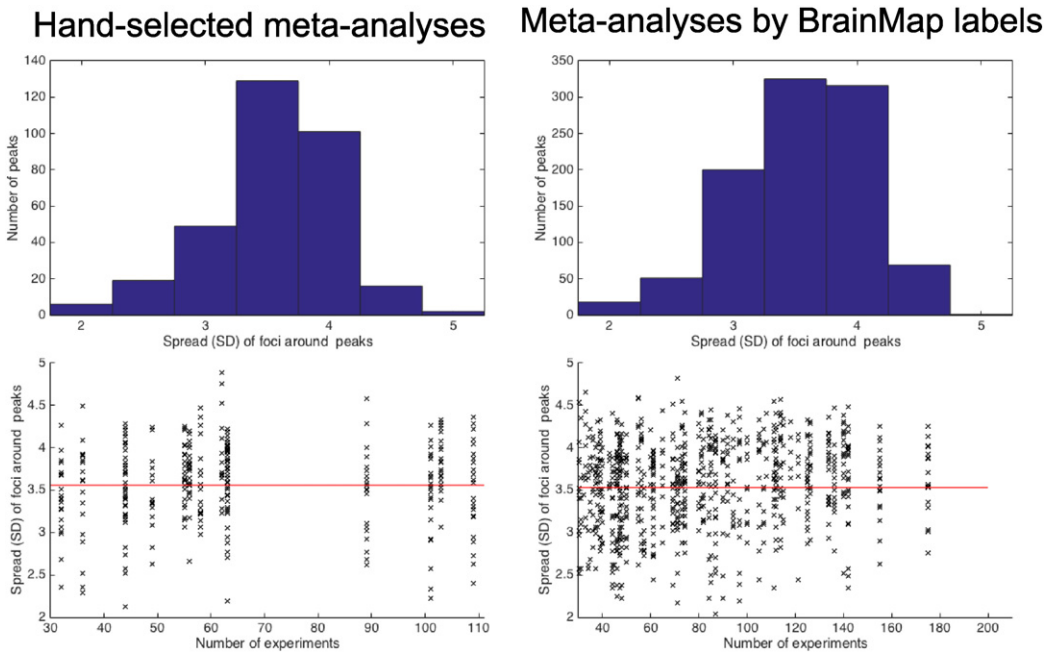

Fig. 1. Evaluation of the activation spread to be used in the simulations. We assessed 15 hand-coded datasets for topic-based ALE meta-analyses (left side) as well as 105 datasets defined by combinations of the Behavioral Domain and Paradigm class meta-data from BrainMap (right side). After identifying the principal significant peaks in the ALE meta-analysis map, we computed the standard deviation of foci that contributed to that peak. In spite of the differences between both sets of meta-analyses with respect, they show homogeneity with the contributing experiments with a similar concentration (3–4 mm standard deviation) in each case (upper panels). Importantly, this spread seems independent of the number of experiments that constitute the respective meta-analyses (lower panels). Based on these data, the 500 simulations used spreads of 2 mm (50 times), 3 mm (200 times), 4 mm (200 times) and 5 mm (50 times) around the true location of the simulated effect.

analysis to peaks reflecting task-based activations, and discarded deactivations from the analysis. This resulted in approximately 7200 eligible experiments at the time of analysis. For each experiment in each of the simulation datasets, we first determined the "number of subjects" by randomly drawing this parameter from the aforementioned 7200 experiments. Independent of sampling the empirical subject numbers, we drew the "number of reported foci" for each experiment, again based on the distribution of this parameter among the normal mapping experiments in BrainMap (Laird et al., 2009, 2011a). Finally, for each experiment and independently from the other steps, the respective foci from the empirical distribution thereof in the BrainMap database were drawn. That is, for each experiment the number of subjects, the number of reported foci and the spatial distribution of foci in the simulation all corresponded to the distribution of neuroimaging literature stored in BrainMap, but were randomly reassembled by independent draws.

Critically, for those experiments that were supposed to feature a "true" activation of the target-location, we replaced one of the foci by a coordinate that was located at −42/32/26 (the coordinates of the simulated effect of interest) with a random Gaussian displacement corresponding to the spread as described above.

We than conducted state-of-the-art ALE analyses on the ensuing 122,500 datasets and performed statistical inference on significant convergence using voxel- and cluster-level FWE corrected at $p < 0.05$ (Eickhoff et al., 2012), voxel-level FDR corrected at $p < 0.05$ (Laird et al., 2005) and uncorrected $p < 0.001$ with an additional extent-threshold of 200 mm^3.

All simulations were run using in-house MATLAB scripts implementing the ALE algorithm as described in the respective publications (Eickhoff et al., 2009, 2012; Turkeltaub et al., 2012).

Behavior of ALE and p-values

Fig. 2 depicts the behavior of the ALE-scores and p-values at the local peak around the target-location (4 mm search radius around the true location) in the 245 cases assessed in the current simulation, i.e., the distinct combinations of "total number of experiments" and "experiments activating the target-location". ALE scores quantify the convergence of foci given the associated spatial uncertainty, while the p-values quantify, how likely such convergence is under the hull-hypothesis of random spatial association. Several important, reassuring observations on the obtained ALE scores and p-values may be noted.

As the number of experiments activating the target-location increases, the average ALE score (across simulations) increases almost linearly. In this context it should be reiterated that the ALE values are the union (rather than sum) of the respective probabilities for the individual experiments (Turkeltaub et al., 2002). Yet, in the considered range of values these are virtually identical. The ALE values also increased slightly but consistently as a function of the total number of experiments, independently of the number of experiments activating the target-location. This effect is most likely attributable to interference (superposition) of unrelated foci, which probabilistically increases as the number of experiments is increased when averaging across the 500 distinct simulations of each case.

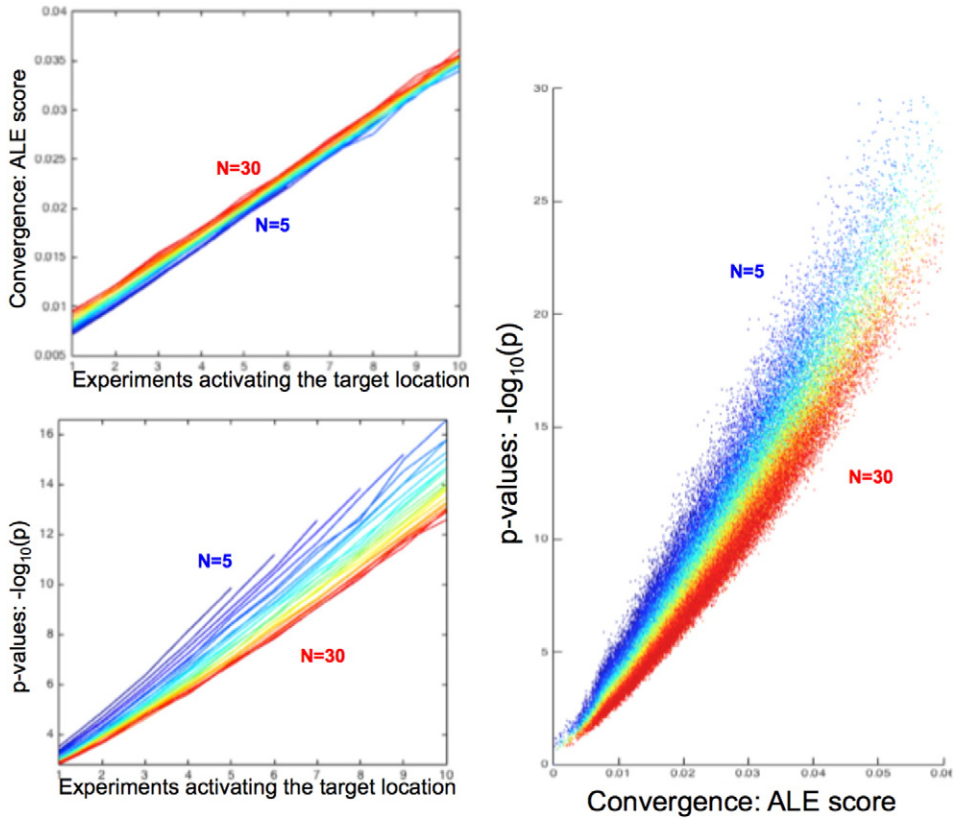

Fig. 2. Characteristic behavior of the ALE scores and the corresponding p-values under the different simulation conditions as observed across 122,500 simulated ALE analyses. The total number of experiments in the respective simulated ALE is coded in a spectral sequence from 5 experiments (dark blue) to 30 experiments (dark red). The top left panel shows the average ALE-score (across simulations) at the simulated location (again using the highest local maximum within 4 voxels of the "true" location). ALE scores increase with number of experiments due to additional contribution of noise from other unrelated foci. The bottom left panel shows the average p-value (across simulations) at the simulated location (using the highest local maximum within 4 voxels of the "true" location). It shows that given the same number of experiments activating at a particular location, lower total numbers of experiments (i.e., greater prevalence of activated experiments) yield lower p-values. The right panel demonstrates the inverse relationships evident from the left panels by plotting ALE scores vs. p-values for the simulated location (again using the highest local maximum within 4 voxels of the "true" location) for all 122,500 simulations. Higher total numbers of experiments lead to higher p-values for the same ALE scores or, conversely, require higher ALE scores for the same p-value.

As expected, the p-value of the local peak at the target-location is also strongly dependent on the number of experiments featuring "true" activation at the target location. However, it decreases when the total number of experiments is increased. That is, given the same number of true activations, a higher amount of experiments that do not feature activation in the target location increases the p-value of the observed convergence, even though the ALE score is on average slightly increased due to random superposition. This behavior reflects the shift in the null-distribution, i.e., the distribution of ALE scores under the null-hypothesis of random spatial association, to higher values if the total number of experiments is increased.

Plotting the peak ALE-score at the target-location against their p-values across all 122,500 simulations summarizes the above observations (Fig. 2, right). While higher ALE scores yield lower p-values, this relationship is systematically shifted according to the total number of experiments included in the simulated meta-analysis. A higher total number of experiments lead to higher p-values for the same ALE scores or, conversely, requires higher ALE scores for the same p-value.

Excessive contribution of individual studies

Empirically, it has repeatedly been noted that significant effects may be largely driven by even a single experiment if the total number of experiments is relatively low (Bzdok et al., 2011; Raemaekers et al., 2007; Turkeltaub et al., 2012; Wager et al., 2007). This comes as no surprise. In a very small ALE analysis of, e.g., 6 experiments, ALE scores of a single experiment may already be close to significance relative to the overall null-distribution. This is especially the case with a higher number of subjects that result in a tighter Gaussian distributions and hence higher ALE scores (Eickhoff et al., 2009). Consequently, even minimal overlap of a second MA map may yield ALE scores that become significant. This observation has prompted anecdotal recommendations that "at least 10–15 experiments" should be included in order to perform a meaningful ALE analysis (Eickhoff and Bzdok, 2013). In order to remedy this unsatisfactory situation and to provide quantitative guidelines on the minimal number of experiments needed for thorough ALE analyses, we quantified the contribution of the most dominant individual experiments to the significant clusters under different thresholding conditions. Importantly, this analysis was based on the "true" location of the effect according to the structure of the BrainMap database (Langner et al., 2014). For each cluster that survived statistical thresholding, we computed the fraction of the ALE value that was accounted for by the one and two, respectively, most dominant experiments and, as well as the fraction accounted by the two most dominant experiments (Fig. 3). This was computed as the ratio of the ALE values with and without the respective experiment. In this context, we need to acknowledge that the ALE computation (union of the MA-values) is actually a non-linear operation. However, as demonstrated in Fig. 2, in the actual (low) range of probability values encountered in ALE, the union in essentially equivalent to a linear summation.

It may be noted, that for voxel-level FWE thresholding, 8 experiments are enough to ensure that on average the contribution of the most dominant experiment is lower than 50%, but the two most dominant experiments explain more than 90% of the total ALE score. In turn, using cluster-level FWE thresholding, 17 experiments ensure that on average the contribution of the most dominant experiments is less than 50% whereas the contribution of the two most dominant experiments is less than 80%. When using uncorrected thresholding at $p < 0.001$ at the voxel level with an additional extent-threshold of 200 mm^3, as commonly seen in ALE papers, 28 experiments are needed to restrict the average contribution of the most dominant experiment to <50%. Finally, as we will explain later in detail (cf. Fig. 6), for FDR thresholding the number of the additional clusters is strongly dependent on the number of experiments activating the "true" location. Consequently, FDR thresholding, which will figure least strongly in the following, is not included in the comparison.

In summary, these data suggest that cluster-level thresholding does a very good job of controlling excessive contribution of one experiment

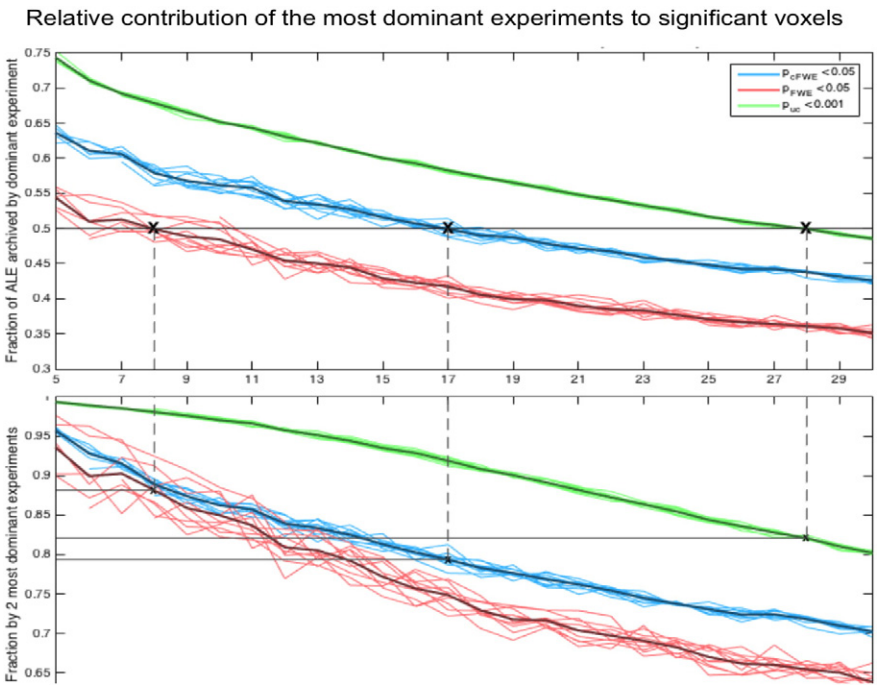

Fig. 3. To quantify the empirical observation that significant effects may be largely driven by a single experiment if the total number of experiments is relatively low and hence to provide quantitative guidelines on the minimal number of experiments needed for valid ALE analyses, we quantified the number of experiments contributing to the significant clusters under different thresholding methods. This analysis is not based on the "true" location of the effect but rather on those at which random convergence happened through the structure of the BrainMap database. For each of these additional clusters, surviving statistical thresholding, we computed the fraction of the ALE value that was accounted for by the most dominant (top panel) and two most dominant (lower panel) experiments. The light lines denote the average across the different number of experiments activating the "true" location and illustrate the robustness of these findings.

to the ensuing significant findings if 17 or more experiments are included in the ALE analysis. This number is lower for voxel-level thresholding, where 8 experiments ensure a contribution of <50% for the most dominant and 14 experiments a contribution of <80% of the two most dominant experiments. This, however, comes at a substantially reduced sensitivity as will be outlined in the next section.

It may be noted, that for voxel-level FWE thresholding, 8 experiments are enough to ensure that on average the contribution of the most dominant experiment is lower 50%, but the two most dominant experiments explain more than 90% of the total ALE score. Using cluster-level FWE thresholding, 17 experiments ensure a top-contribution of less than 50% and a contribution of the two most dominant experiments of less than 80%. Given that for FDR thresholding the number of the additional clusters was strongly dependent on the number of experiments activating the "true" location as later seen in Fig. 6, we did not consider FDR in this analysis. In summary, this data suggests that cluster-level thresholding does a very good job of controlling excessive contribution of one experiment if 17 or more experiments are included in an ALE analysis.

Sensitivity

As a next step, we quantified the sensitivity of ALE meta-analyses under different conditions, i.e., systematically varying the number of experiments activating the target location and the total number of experiments, using the four different thresholding approaches. In this context, sensitivity was provided by the fraction of the 500 random realizations that yielded a statistically significant finding at the location of the simulated true activation site. As illustrated in Fig. 4, the sensitivity is strongly related to the number of experiments featuring activation in the target location in a roughly sigmoid fashion.

Three key aspects that characterize the sensitivity of ALE analyses may be noted. First, a higher total number of experiments leads to a right-shift in the sensitivity curves. That is, given the same number of experiments activating the target location, the sensitivity to correctly detect this convergence decreases when the total number of experiments increases. This evidently mirrors the behavior of (local peak) p-values described above. Second, independently of the chosen statistical thresholding method and total number of experiments, sensitivity converges to almost perfect recovery once a sufficiently high number of experiments activate the "true" location. How quickly perfect detection rate is observed, however, depends on the total number of experiments, reflecting the aforementioned right-shift, and importantly also the thresholding method chosen. As a third observation, the quite marked differences in sensitivity between the four assessed thresholding methods. As expected, the uncorrected method is most sensitive. In particular, uncorrected voxel-level thresholding in combination with an (arbitrary) extent threshold of >200 mm^3 has at least 80% power when 4 experiments activate the modeled true location (out of up to 30 experiments considered). Corrected inference methods have less power but differ considerably in the slope of the sensitivity curve. Cluster-level FWE correction represents the most sensitive approach, followed by voxel-level FDR correction. In turn, voxel-level FWE is the least sensitive approach, reaching 100% sensitivity only in cases where 10 experiments feature activation around the target location. This is consistent behavior in traditional brain imaging where cluster-level inference typically exhibits superior power to voxel-level inference (Friston et al., 1996), though this difference must be considered in light of cluster-inference's reduced spatial specificity relative to voxel-level inference.

As shown in Fig. 5, across all thresholding approaches the size of the excursion set strongly increases with the number of experiments activating around the target location. In turn, clusters become smaller when the total number of experiments increases given the same number of experiments activating at the target location. Strikingly,

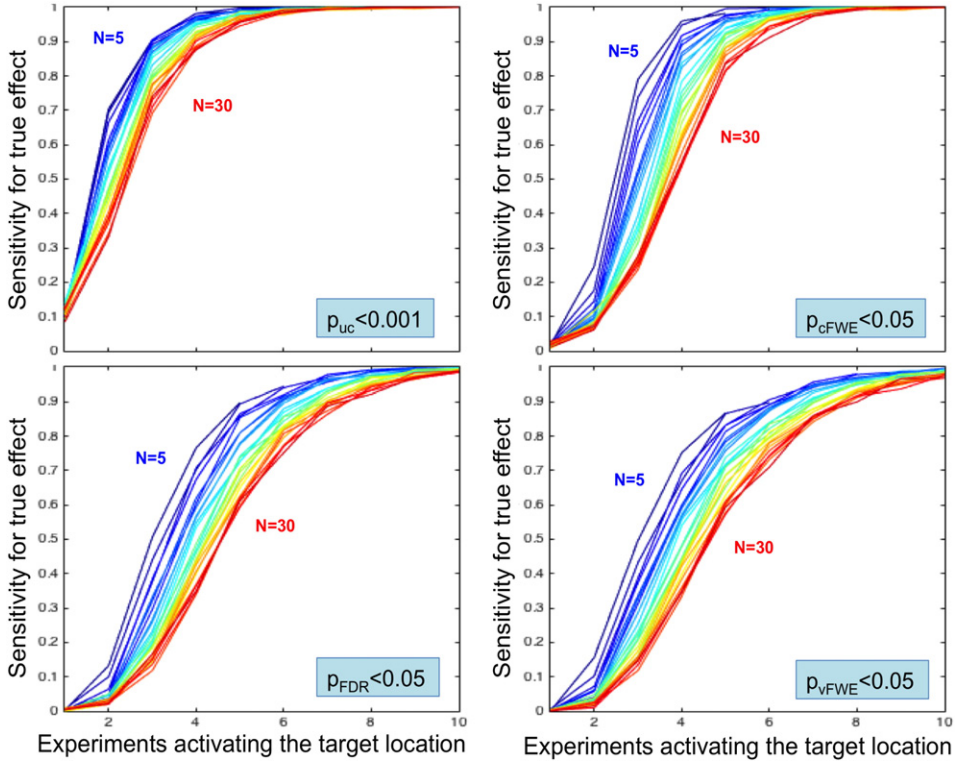

Fig. 4. Sensitivity of ALE to detect the simulated true convergence given the number of experiments activating the target location. The total number of experiments in the respective analyses is coded in a spectral sequence from 5 experiments (dark blue) to 30 experiments (dark red). Three key aspects may be noted. First, a higher total number of experiments leads to a right-shift in the sensitivity curves. Second, independent of the chosen statistical thresholding method and sample size, sensitivity curves converge to 100% when a sufficiently high number of experiments activates the "true" location. Third, cluster-level correction shows a higher sensitivity than voxel-level FDR and particularly voxel-level FWE thresholding.

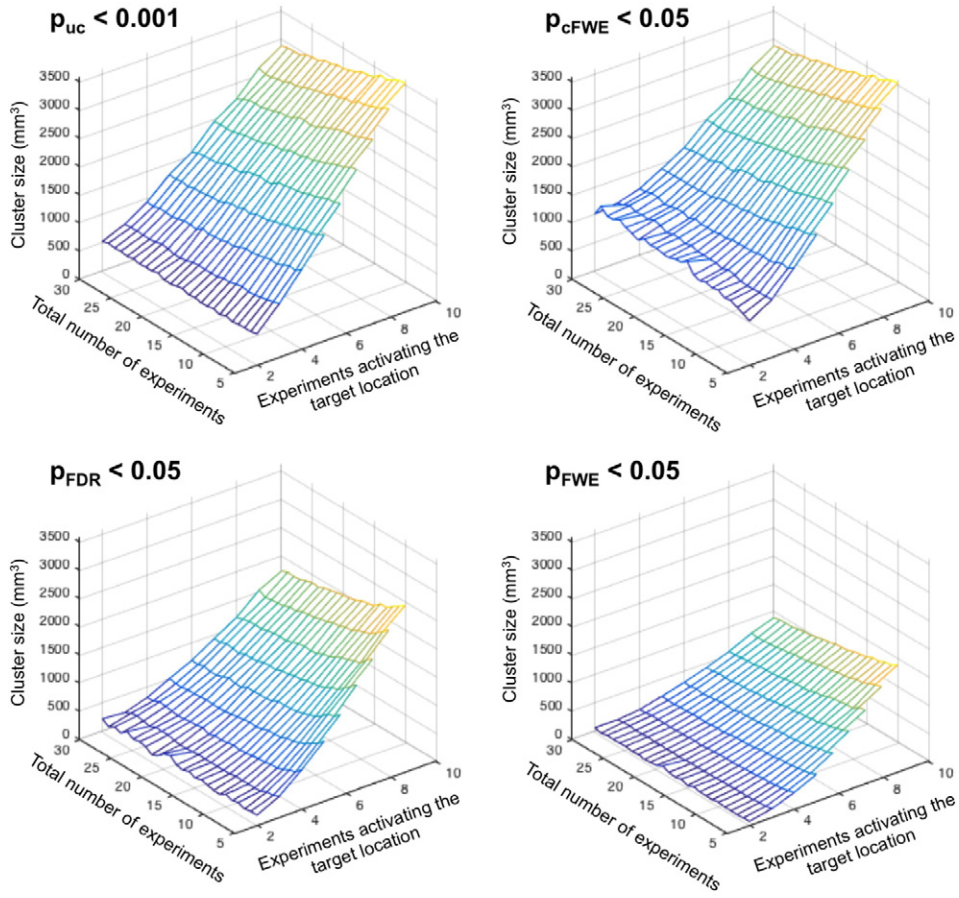

Fig. 5. Cluster-size of the super-threshold cluster at the "true" location, i.e., the target of the simulation, in relationship to the total number of experiments and the number of experiments activating the target location. It becomes apparent that the cluster size increases strongly with the number of experiments activating the "true" location. In turn, clusters become smaller when the total number of experiments increases given the same number of experiments activating the target location. Finally, we note that FDR and in particular voxel-level FWE thresholding yields much smaller clusters than cluster-level FWE thresholding.

however, the ensuing clusters are much smaller when using voxel-level FWE or FDR correction as compared to cluster-level FWE correction, reflecting some combination of cluster-level's greater power and voxel-level's greater spatial specificity. In addition to underlining the differences in sensitivity between the different approaches, this observation also leads to another important consideration. When using voxel-level FWE thresholding in an ALE study, it is quite likely that the ensuing significant clusters are of very small size. In this context, it is worth reiterating that even a single voxel may allow for a significant finding, given the voxel-wise correction for multiple comparisons across the whole brain. Nevertheless, such findings are often very difficult to display and interpret, and may require an arbitrary re-thresholding to more clearly demonstrate the area of activation.

Susceptibility to spurious convergence

In a next step, we assessed the average (across random realizations of the same simulated ALE analysis) number of clusters of significant convergence outside of the target location. We emphasize that these are not equivalent to false positives in a strict sense, which would be controlled through correction for multiple comparisons. Rather, they reflect incidental convergence through the structure of the BrainMap database (Langner et al., 2014). Such convergence arises from the fact that the frequency of observed activations in neuroimaging experiments and hence also BrainMap is not homogeneous across the brain. Some brain regions like the anterior insula (Kurth et al., 2010; Yarkoni et al., 2011) and the posterior-medial frontal cortex (Muller et al., 2015) are more often activated than, for instance, the lateral temporal cortex. By randomly sampling the noise foci from BrainMap, the current simulations thus reflect this bias, leading to incidental convergence outside the target location. Given that a similar structure should also be present in most real-world ALE analyses, as it most likely reflects task-set effects or epiphenomena of the neuroimaging setting, they should represent a very good approximation of how likely spurious findings are found. They would usually be considered false positives in usual neuroimaging settings. Yet, we would like to stress the distinction between false positive findings in the statistical sense and such incidental effects, which are true positives but conceptually spurious convergence. The advantage of the present set-up thus lies in the fact that the ground truth is known and by drawing resamples from the known BrainMap sample (which is itself a sample of the unknown population of neuroimaging studies) we can quantify the uncertainty (as interval estimates), which is conceptually similar to bootstrapping approaches (Efron and Tibshirani, 1994).

Reassuringly, both voxel- and cluster-level FWE correction yielded very low numbers of additional clusters independently of the number of experiments activating the target location and the total number of experiments (Fig. 6). That is, controlling the FWE seems to be a very potent approach to limiting the likelihood of finding incidental convergence. In turn, when performing uncorrected thresholding ($p < 0.001$ at the voxel level with an extent threshold of 200mm^3) the number of additional clusters is not only substantially higher but also depends heavily on the total number of experiments entering the ALE. While this effect can be expected given that the chance for additional overlap increases if more experiments are present, it also indicates that uncorrected thresholding even with an additional (arbitrary) extent threshold does a very bad job in controlling for incidental findings. Consequently, most results based on uncorrected thresholds

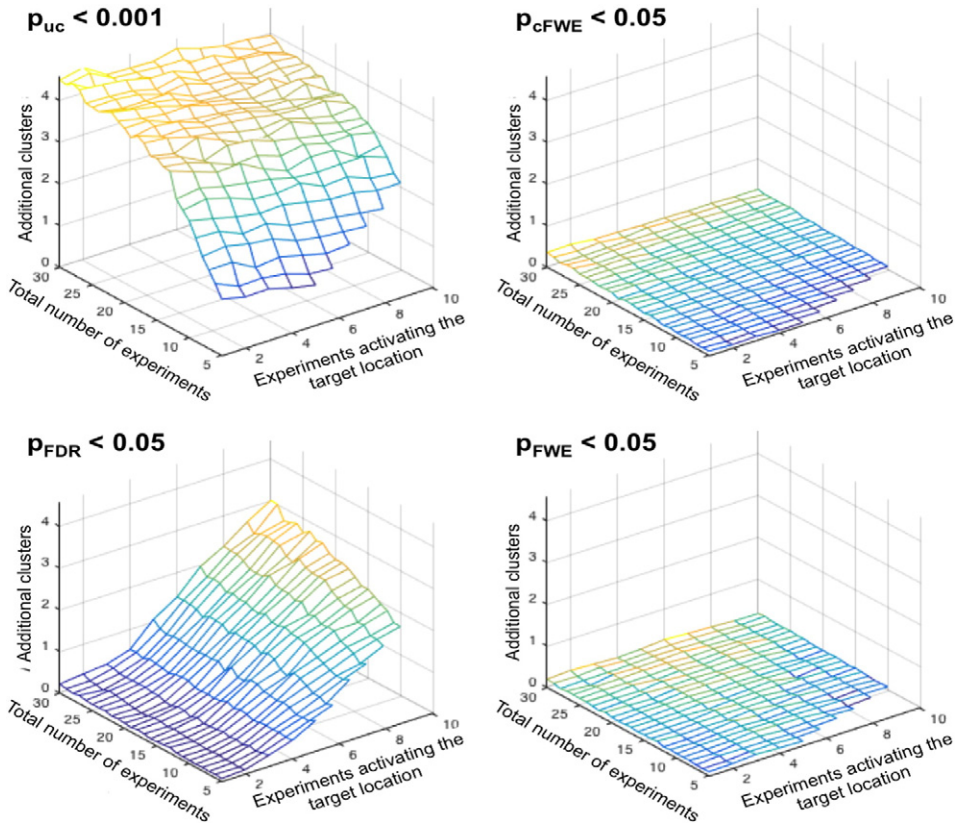

Fig. 6. Average number of additional clusters of significant convergence outside of the "true", i.e., target location in relationship to the total number of experiments, and the number of experiments activating the target location and the significance thresholding. Given that the distribution of the entire BrainMap database is known, these analyses allow quantifying deviations from the ground truth, similar to false positive findings. As can be seen, both voxel- and cluster-level FWE correction yield very low numbers of additional clusters. In turn, two interesting and orthogonal patterns may be noted for uncorrected thresholds and voxel-level FDR correction. When using the former ($p < 0.001$ at the voxel level with $k > 200\ mm^3$) the number of additional clusters depends primarily on the total number of experiments entering the ALE. This may be expected because the chance for additional overlap increases if more experiments are present. For voxel-level FDR correction, however, the number of additional "false positive" clusters depends strongly on the number of experiments activating the target location, i.e., the "true" effect. This effect may be explained by Fig. 5, considering that a higher number of significant voxels at the target location allow for more false positive voxels.

may be expected to contain additional, spuriously significant clusters. The more worrisome pattern, however, is found for FDR correction (Laird et al., 2005). Here the number of incidental clusters depends strongly on the number of experiments activating the target location, i.e., the "true" effect. Again, this effect is actually to be expected. As evident in Fig. 5, the number of significant voxels in the target region strongly increases with the number of experiments activating here. In turn, a higher number of significant voxels at the target location allow for more false positive voxels when the false positive rate is controlled (Genovese et al., 2002). Given the spatial smoothness of ALE data, these false positive voxels are then usually aggregated to form spurious clusters. Our simulations are thus in line with previous considerations (Chumbley and Friston, 2009) regarding the inappropriateness of FDR for inference on fMRI data and underlines that FDR correction is also not appropriate for ALE meta-analyses.

Power analysis for inference on the underlying population of experiments

As noted above, the sensitivity of ALE to detect a true effect is positively related to the number of experiments activating the target location but at the same time negatively related to the total number of experiments. This may seem to contradict the idea that a higher number of experiments should result in greater power to detect subtle effects. Here we need to introduce a critical distinction between the sensitivity to detect convergence within a given sample of experiments and the power to reveal an effect present in the underlying population of experiments from which the analyzed ones are sampled. In the latter context, the "effect size" is given by the proportion of the experiments in the underlying population showing activation at a given location. The analyzed experiments are then assumed to represent random samples from this underlying population. Fundamentally, the power to detect a given effect (with an effect size as denoted above) then depends on the probability that x out of N experiments in an ALE analysis show the effect and the sensitivity of the ALE to identify it. The former probability can easily be derived from the "effect size" using a binomial distribution, the latter is provided by the current simulation study. Whereas the simulation results above are computed for a fixed and known number of truly active studies, we can extend these results to a random and unknown number of truly active studies using a binomial distribution. If $1-\beta_x$ is the power for x truly active studies, we can compute expected power when proportion p studies are expected as

$$\sum_{x=1}^{N}(1-\beta_x)P(X = x; N, p)$$

where $P(X=x;N,p)$ is the binomial probability of observing x out of N counts, each occurring with success probability p.

The results of these power-computations (Fig. 7) reveal three main trends. First, the power is markedly different between the thresholding methods. Given that the likelihood of drawing x out of N experiments that feature an activation effect under the assumption of a certain effect size is fixed, this directly reflects the differences in sensitivity alluded to previously. Second, ALE analyses with less than 10–50 experiments yield a low power to even find very consistent effects. In fact, even ALE analyses with 30 experiments are not very highly powered to reveal low effect-sizes, i.e., effects that are only present in a smaller proportion

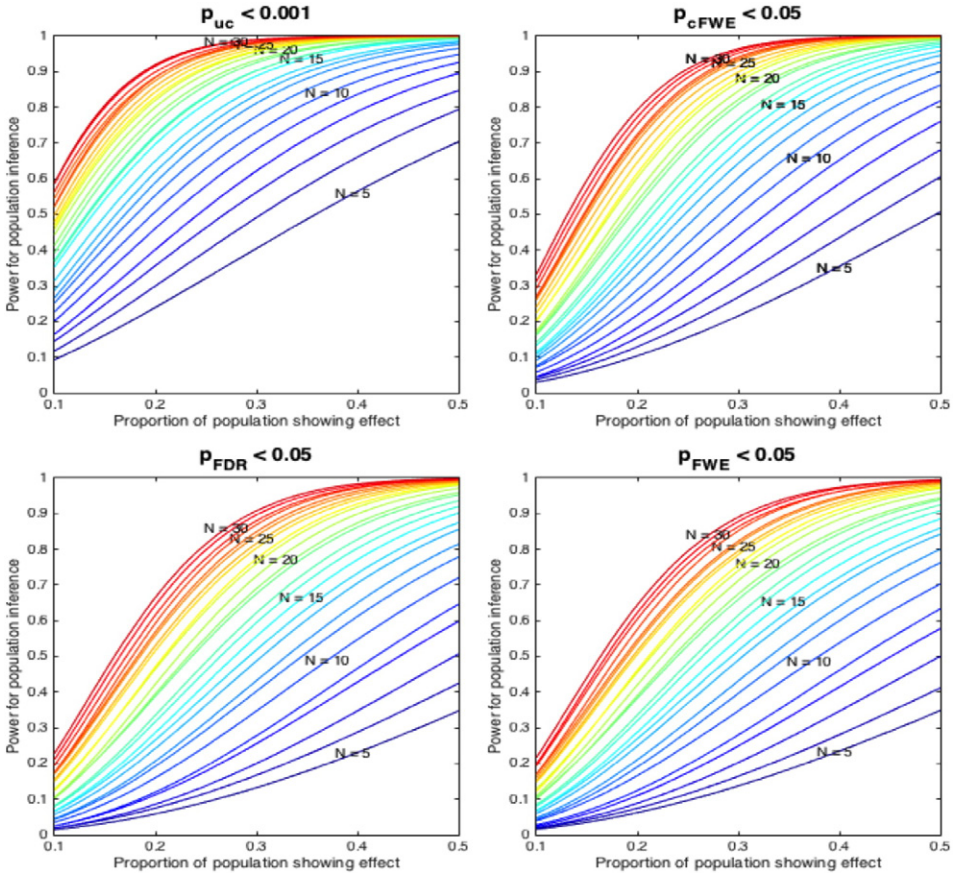

Fig. 7. Power of inference on the underlying population of experiments assuming different "effect sizes" (proportion of the experiments in the underlying population showing an effect at a given location). The power to detect a given effect in the underlying population depends on the probability that x out of N experiments in an ALE analysis (assumed to be random samples from the underlying population) show the effect and the sensitivity of the ALE to identify it (cf. Fig. 4). The total number of experiments in the respective simulated ALE is again coded in a spectral sequence. In spite of the differences in power between the thresholding methods, two trends are noticed. ALE analyses with less than 10–50 experiments yield a low power to find consistent effects. Even ALE analyses with 30 experiments are not very highly powered to reveal rare effects. In addition, we note that voxel-level FDR thresholding combines a low sensitivity (cf. Fig. 4) with a high potential for false positive or spurious findings especially when there is a strong true effect (cf. Fig. 6).

of the underlying population of experiments. Finally, we note that voxel-level FDR thresholding combines low sensitivity (cf. Fig. 4) and hence power with a high potential for false positive or spurious findings when there is a strong true effect (cf. Fig. 6).

The probably even more important consideration with respect to power is the question regarding the required number of experiments for an ALE analysis. Above we argued that in order to limit the influence of any single experiment on the ensuing results, i.e., to avoid clusters that are predominantly driven by an individual experiment, between 8 (voxel-level FWE) and 17 experiments (cluster-level FWE) are needed. The sample size calculations for ALE meta-analyses assuming a desired power of 80% given different "effect sizes" (proportion of the experiments in the underlying population showing an effect at a given location) displayed in Fig. 8 complement this impression. As may be noted, a sample size of 17 experiments assessed using cluster-level FWE inference yields an 80% power to detect effects that are present in about a third of the underlying population. In other words, only relatively consistent effects may be detected with sufficient power using the recommended minimum number of experiments. In turn, 80% power to detect an effect size of 0.2, i.e., a finding that is present in one out of five experiments, is only present when more than 30 experiments are included in the ALE analysis, given cluster-level FWE thresholding. (See Fig. 9.)

This evidently raises the question as to which effect sizes may reasonably be expected for ALE analyses. While this depends on the subject under investigation, we provide some first insight into this matter in Fig. 8 illustrating the "effect sizes" for the real-life ALE analyses. Here we used the same datasets as in Fig. 1, and combined them with the sample-size calculations for cluster-level FWE inference. It may be noted, that strong effects such as 40% or more of a dataset showing a particular effect are rare, while "effect sizes" of 0.2–0.25 are much more common.

Discussion

Conceptual considerations

The present empirical simulation study based on the BrainMap database provides a quantitative assessment of the statistical properties of ALE analyses. These parameters for the simulations, in particular the assumed spatial spread around a "true" location were based on a large number of stored meta-analyses and moreover converged nicely between hand-coded and automatically generated datasets. Consequently, we feel confident, that the simulations reflect real-world situations that researchers will encounter when performing future ALE analyses in everyday research practice. In that context, we need to acknowledge, however, that the true spread around a hypothetical location of activation must remain unknown, as we can only provide estimates based on databased empirical observations. Moreover, it may be discussed whether indeed something like "the" true location of an effect exists, given recent accounts of distributed coding and representation (Bzdok et al., 2015; Haxby, 2012; Kriegeskorte, 2009; Rissman and Wagner, 2012). It may additionally be argued that different variations of a behavioral task may recruit slightly different locations within a larger brain region, rendering the current estimates a mixture of systematic effects and random noise.

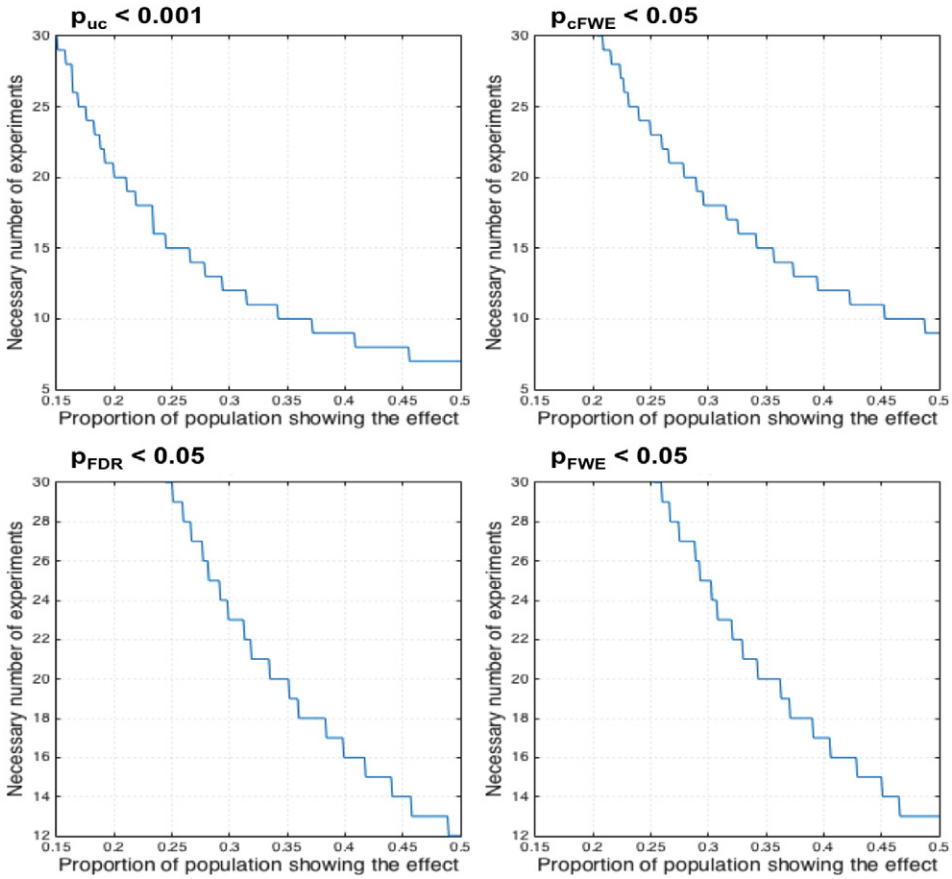

Fig. 8. Sample size calculations for ALE meta-analyses assuming a desired power of 80% given different "effect sizes" (proportion of the experiments in the underlying population showing an effect at a given location) for each of the four assessed thresholding approaches.

While important on the conceptual level, these considerations do not change the fact that such effects should be equally present in the (future) ALE analyses as they are in the simulations. This highlights the advantage of basing the simulations on empirical information, rendering them a faithful reflection of the phenomena under investigation. Moreover, by using a resampled dataset from the BrainMap database to construct the simulations, we ensured that all parameters that otherwise would have to be set manually reflect the distribution thereof in a large sample of published neuroimaging studies. This in particular pertains to the number of subjects and reported number of foci in each experiment as well as the localization of the "noise" foci. We would thus argue, that our results should represent a realistic reflection of situations that are encountered when conducting meta-analyses and should therefore generalize to real-life settings.

Recommendations for statistical inference

From the observations that were made by assessing >120,000 realistically simulated analyses using the four commonly applied inference methods, several recommendations may be derived.

- Cluster-level family-wise error (FWE) thresholding was observed to provide the best compromise between sensitivity and specificity and is hence recommended for inference on ALE analyses. In particular, our analysis showed that cluster-level FWE corrected inference is almost as sensitive to true effects as uncorrected thresholding and consequently provides substantially higher power than voxel-level FWE or FDR thresholding. Importantly, this high sensitivity is not offset by a higher susceptibility to incidental convergence. Here we would like to reiterate, that the latter may not be easily equated with "false positives" in the statistical sense. Rather, these effects represent spurious convergence due to the non-homogeneous likelihood of activating any particular voxel in the brain in neuroimaging (Fox et al., 2014; Laird et al., 2011b; Langner et al., 2014). In the context of a specific ALE meta-analysis, however, such incidental convergence due to the fact that a region may be frequently activated would often be broadly equivalent to a false positive relative to the subject under investigation.
- We would strongly discourage the use of voxel-wise false-discovery rate (FDR) thresholding in the context of ALE meta-analyses. As previously pointed out, voxel-wise FDR correction, the type considered here and almost exclusively used in ALE analysis is not appropriate for inference on topological features such as regions of significant convergence of a smooth dataset (Chumbley and Friston, 2009). The present investigation corroborates this notion by showing that voxel-level FDR correction entails both relatively low sensitivity and a high susceptibility to spurious, false positive findings. Moreover, it also highlights another negative property of FDR thresholding, namely dependence of the latter on the strength of true convergence in other parts of the brain. If there are regions in the image where a strong effect is present, other regions are most likely to be declared significant as well as compared to a situation where the former is not present (Genovese et al., 2002). This is inherently correct in the logic of the false discovery rate as the number of false positives will scale with the number of voxels declared significant which in turn will depend on the presence of true effects. It however implies whether a voxel, such as in the occipital lobe is declared significant or not may depend on the magnitude of convergence in the prefrontal

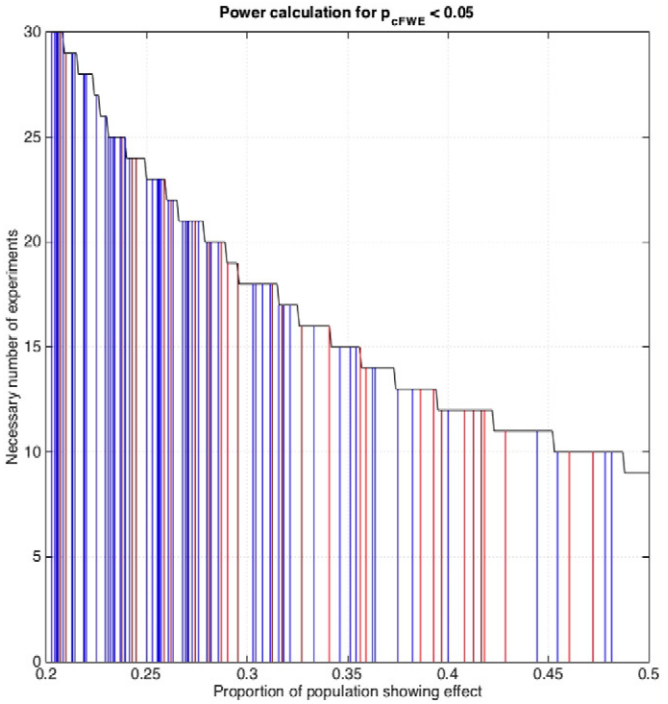

Fig. 9. Illustration of "effect sizes" (proportion of the experiments in the underlying population showing an effect at a given location) found for real-life ALE analyses. Here we used the same datasets as in Fig. 1 and combined them with the sample-size calculations for cluster-level FWE inference. Red lines correspond to peaks from the hand-coded datasets for topic-based ALE meta-analyses and blue lines to the datasets defined by combinations of the Behavioral Domain and Paradigm class meta-data in BrainMap. Note that strong effects such as 40% or more of a dataset showing a particular effect are rare, while "effect sizes" of 0.2–0.25 are much more common.

cortex. In summary, FDR thresholding should thus be avoided for inference on ALE analyses.

- With respect to uncorrected thresholding, our analyses confirmed previous expectations. Namely, uncorrected thresholding, even in combination with an additional extent threshold is very prone to reveal spurious findings. Moreover, results are extremely likely to be driven by only a single dominant experiment. Consequently, uncorrected inference is to be avoided.

Finally, voxel-level FWE corrected thresholding may be too conservative for most applications, apart from maybe those in which hundreds of experiments are analyzed and hence more stringent inference is desired. It does offer an excellent protection from declaring incidental overlap significant and also makes it very unlikely that significant results are predominantly driven by a single experiment. On the other hand, considering that cluster-level FWE thresholding performs similarly well in these aspects and at the same time offers higher sensitivity and, on a pragmatic perspective, also avoids very small and hence hardly interpretable clusters, we would argue that the latter should be preferred. In this context, however, we would like to stress the conceptual differences between cluster- and voxel-level thresholding. In particular, when applying the former, significance is achieved by the cluster as a whole through its spatial extent, which is evaluated against a non-parametric null-distribution thereof. When applying cluster-level correction, one can thus not make claims about any particular voxel within the cluster being (most) significant, even though peak coordinates are often -inappropriately - reported in such situation. This predicament becomes particularly obvious, when a significant cluster spans multiple anatomical areas. To illustrate this point, if a cluster covers both the anterior insula and the inferior frontal gyrus, one cannot necessarily infer that both regions show convergence but rather only that the entire super-threshold cluster covering both regions is larger than expected by chance. In turn, voxel-level correction allows to attribute significance to each voxel above the respective threshold and therefore a somewhat better allocation to a specific brain region.

How many experiments are needed for an ALE analysis?

There is no single right answer to this question. Each topic and hence each dataset will have its idiosyncrasies with respect to many potentially influencing parameters, such as the heterogeneity of the included experiments, the variability of peak locations in these experiments, the average number of subjects and so on. Furthermore, there are two different levels to this question. Namely, how many experiments are needed to avoid results that are largely driven by one experiment and how many experiments are needed to have sufficient statistical power for detecting less consistent effects. The current simulation study addressed both aspects based on a large dataset that should be well representative for situations encountered in current neuroimaging meta-analyses and yielded a rather clear answer to at least the first question.

Based on the obtained results, we would make a strong case for including at least 17 experiments into an ALE meta-analysis to control the influence of any individual experiment. This recommendation is based on the observation that, when controlling for multiple comparisons using cluster-level family wise error correction, from this size on, the average contribution of the most dominant experiment to any above-threshold cluster is less than half and the two most dominant experiments contribute on average less than 80%. While the absolute numbers, in particular the criterion of 50% contribution for the most dominant experiment, may be disputed, this recommendation is well in line with our empirical evidence across many ALE analyses on a large range of topics. In particular, in line with the data summarized in Fig. 3 we observed that in analyses involving less than 10 experiments, it becomes very likely that results are largely driven by a singly experiment. This is in particular the case if there is a marked heterogeneity in the number of subjects involved in the different experiments. More specifically, this is quite likely to happen if one experiment is based on a substantially larger N than the remaining ones and hence features a tighter Gaussian leading to higher voxel-wise probabilities around the reported foci (Eickhoff et al., 2009). In this case, even a relatively minor contribution from a second experiment will almost certainly move the ensuing ALE scores above the significance threshold. From the current simulations, however, it became clear that previous recommendation "at least 10-15 experiments should be included in an ALE meta-analyses" was still somewhat optimistic and needs to be adjusted upwardly closer to 20.

We acknowledge that a robust approach to ensure that results are not driven by any one experiment may be provided by the use of jackknife analyses, that is, re-computation of the ALE analysis while in turn leaving out each experiment from the dataset. There is no question that results obtained throughout all realizations of the jackknife procedure must be considered robustly present in the dataset and not dependent on any one experiment (Amanzio et al., 2013; Palaniyappan et al., 2012). There are, however, two downsides to using this statistical procedure as primary inferential tool, which became most apparent when considering the two extreme cases, i.e., meta-analyses involving very few and very many experiments. In the latter case, repeating the full analysis, including the calculation of the cluster-size null-distribution as many times as there are experiments will be extremely computationally expensive and hence impractical. Moreover, our simulations also indicated, that in these cases jackknife analyses will not really be necessary, given that already in the case of 30 experiments the influence of any one or two experiments becomes quite limited. The other case is conceptually more interesting, though. When only including, e.g., less than ten experiments, jackknife analyses may indeed establish that an effect is not driven by any individual study. However,

such analyses will quickly turn out very conservative while on the other hand not offering any protection against results driven by two experiments, given that in the even smaller sample it becomes likely that the non-removed one may drive the results by itself. Finally, as discussed below, we would argue the idea of meta-analyses as a quantitative assessment of convergence in the current literature is undermined when only a handful of experiments may actually be integrated. One of the reasons is that such analyses would be really ill-powered, which brings us to the second aspect of the question of how many experiments are needed for an ALE meta-analyses.

The concept of statistical power is intimately linked to the notion of effect size given the two fundamental questions of any power analysis, i.e., "what is the effect size that is reasonably likely to be detected with the current sample size" and "what sample size is needed for a sufficient likelihood of finding a particular effect size". The notion of effect-size, however, is somewhat tricky in ALE meta-analyses, as in contrast to classical behavioral meta-analyses neuroimaging meta-analyses assess spatial convergence (Eickhoff et al., 2009; Kober and Wager, 2010; Radua and Mataix-Cols, 2012). In fact, ALE is in stark contrast to behavioral "effect-size" meta-analyses by not considering the size of the reported effects (Cheung et al., 2012; Jones, 1995) but rather their location and associated spatial uncertainty. Nevertheless, there is a rather straightforward notion of effect-size in ALE meta-analyses, namely the fraction of experiments that activate at a particular location (or rather in that region, as the individual experiments will show some spread around the hypothetical true location). Here we present power-estimates for ALE meta-analyses based on this concept of effect-size for the first time.

We would like to point out, that in the context of power-calculations effect-size pertains to the (unknown) underlying population of experiments of which the included ones are considered a random sample. That is, the effect size is the proportion of experiments in the underlying population that show an effect at a given location. For any desired effect-size the power of an ALE analysis may then be computed as the likelihood of drawing x experiments that show the effect among N total experiments multiplied by the likelihood to find an effect present in x/N experiments, summed over all values of x. From the respective power-curves (Fig. 7) and the ensuing sample-size calculations needed for 80% power (Fig. 8), it becomes obvious, that samples sizes needed to control the influence of any individual experiment as discussed above only provide adequate power for rather strong effects, i.e., those present in about a third of all experiments. Even worse, ALE analyses with only <10 experiments only have good power to likely detect effects that are present in every second experiment. In other words, ALE meta-analyses including fewer than close to 20 experiments but in particular those based on less than ten run a high danger to obtain results that are driven by an individual experiment and only have sufficient power to detect effects that are extremely obvious.

Considering the sample-size calculations presented in Fig. 8, we would suspect that a majority of the present ALE literature may actually be underpowered, given that only sample sizes of more than 30 experiments yield sufficient power to detect an effect that is present in 1/5 of all experiments in the underlying population. Conversely, many published meta-analyses may have only been likely to detect very consistent effects. While this is clearly not a desirable situation, we would not consider the presence of low-powered ALE analyses by itself to be a major problem. Rather, we would argue that the combination of low power with the increased susceptibility to effects driven by a single experiment and the increased temptation to use less adequate thresholding (namely, uncorrected thresholds or voxel-level FDR) to be a potentially dangerous constellation leading to both false negative and false positive findings. Finally, returning to the original question of how many experiments are needed for ALE meta-analysis, we recommend that when using cluster-level FWE thresholding, approximately 20 experiments should be considered the lower bound for valid and decently powered analyses.

Are the results transferable to other approaches for coordinate-based meta-analyses?

While one of the most widely used methods for coordinate-based meta-analyses, ALE is by far not the only approach towards this goal. Rather, multi-level kernel density analysis (MKDA (Kober and Wager, 2010; Nee et al., 2007; Wager et al., 2009)) and signed difference map analysis (SDM (Palaniyappan et al., 2012; Radua and Mataix-Cols, 2012; Radua et al., 2010)) have likewise been frequently used to investigate convergence across functional and structural neuroimaging experiments.

While all algorithms (ALE, MKDA and SDM) are based on the same fundamental idea of delineating those locations in the brain where the coordinates reported for a particular paradigm or comparison show an above-chance convergence, there are several distinctions in the implementation of this aim. Whereas ALE and SDM investigate where the location probabilities reflecting the spatial uncertainty associated with the foci overlap, MKDA tests how many foci are reported close to any individual voxel. All approaches avoid excessive summation through neighboring foci from a same experiment by limiting maximum values and involve some form of permutation/relocation procedure for establishing (corrected) significance, though the exact concepts and implementations vary considerably across methods. Moreover, MKDA and SDM allow emphasizing, i.e., up-weighting, foci derived from conservatively corrected analyses. This feature is not present in the probabilistic approach taken by ALE (Kober and Wager, 2010; Radua et al., 2010). A distinct feature of SDM relative to MKDA and ALE, finally, is the possibility to integrate positive and negative effects in a same map in order to cancel out regions in which both are present (Radua and Mataix-Cols, 2012). While cursory, this overview hopefully illustrates that the concepts and machinery differ considerably between ALE, MKDA and SDM. This is in spite of the fact that all three approaches are widely used and thoroughly validated methods for coordinate-based meta-analyses and often yields comparable results to each other and image-based meta-analyses (Salimi-Khorshidi et al., 2009).

Consequently, we would urge caution when trying to extrapolate the current observations on sensitivity and power for ALE meta-analyses to MKDA and SDM. This is in particular true for SDM: if statistical parametric maps rather than (only) foci are included or effect size maps are estimated from t-values reported in individual studies, behavior and power of such analyses should be fundamentally different from the case investigated here.

Nevertheless, we would consider the advice to strive for sufficiently high numbers of included experiments in any meta-analysis to be valid independently of the employed algorithm, based on two considerations. On the algorithmic level, meta-analyses based on a lower number of experiments easily run into the problem that significant "convergence" may be strongly or even exclusively driven by a single experiment. Perhaps even more importantly, however, on the conceptual level the value of a meta-analysis, as integration across a broader set of previous findings in order to identify robust convergence, may be driven ad absurdum when only a handful of experiments are assessed. However, we have no indication whether the estimated lower bound of ~20 experiments, the observation that the sensitivity of ALE is primarily related to the absolute number of converging foci in spite of the right-shift of this curve by the number of "noise" experiments also holds for MKDA and SDM. Likewise, the power-calculations for different proportions of true effects in the underlying (unknown) population of experiments, which directly depend on the sensitivity-curves, should be specific to ALE.

In summary, we would thus consider the recommendation of striving for sample sizes >20 experiments useful for all types of coordinate-based meta-analyses, even though we have no quantitative data to substantiate this claim with respect to MKDA and SDM.

Relevance for meta-analytic co-activation modeling

In addition to topic-based meta-analyses, i.e., studies integrating previous neuroimaging findings on a particular type of mental process or task, there is a growing interest in using meta-analytic methods for the assessment of task-based interaction patterns (Eickhoff et al., 2010; Fox et al., 2014; Laird et al., 2013). The fundamental idea behind these types of location-based meta-analyses is to assess which regions are co-active above chance given activation in a predefined region of interest as a marker of functional connectivity between them (Robinson et al., 2010; Rottschy et al., 2013). Practically, this is usually implemented by filtering a database such as BrainMap or Neurosynth (Yarkoni et al., 2011) to only retrieve experiments that feature at least one focus of activation in the seed region. In the next step, an (ALE) meta-analysis is then performed over the identified experiments to quantify across-experiment convergence. The highest convergence will probably be found in the seed region, but significant convergence outside the seed indicates significant co-activation. Meta-Analytic Co-Activation Modeling (MACM) thus provides a complementary aspect of long-range integration relative to measures such as resting-state functional connectivity or structural covariance (Clos et al., 2014; Hardwick et al., 2015).

It is important to stress that topic- and location-based meta-analyses use the same underlying algorithms to assess above-chance convergence between experiments and differ only with respect to the way that these are defined (either based on the experimental context that is assessed or the fact that they activate a particular region of interest). Consequently, we argue that the same considerations with respect to inference-approaches, sample-size and power should also hold for MACM studies. We would therefore amend the above recommendations only by two specific aspects. First, in the case of a small seed region and/or a seed that is located in a part of the brain that is not very densely covered by neuroimaging foci, it may be necessary to include a spatial fudge-factor. That is, rather than considering only those experiments that activate within the seed region, it may be advantageous to additionally include those that are within, e.g., 5 mm of the seed in order to obtain a sufficient number of experiments. In the opposite case, i.e., when dealing with a region of interest that yields a very high number of experiments, it may be helpful to use voxel-rather than cluster-level FWE in order to render the analyses more conservative and avoid over-powered inference.

Conclusions

Coordinate-based meta-analyses by means of activation likelihood estimation meta-analyses have enjoyed considerable success over the last decade. Particularly ALE has been applied to a wide range of different topics from cognitive (Chase et al., 2015; Molenberghs et al., 2012), affective (Kohn et al., 2014; Lamm et al., 2011) and motor (King et al., 2014; Yuan and Brown, 2015) neuroscience to the neurobiology of neurological (Herz et al., 2014; Rehme et al., 2012) and psychiatric (Bludau et al., 2015; Chan et al., 2011; Goodkind et al., 2015) disorders. Given the ever-growing expansion of ALE meta-analyses into smaller research areas in which the eligible literature is not as abundant as for, e.g., working memory or social cognitive tasks, the question regarding the number of experiments that is necessary to perform a valid ALE analyses has become increasingly important over the last years. Likewise, several different methods for statistical inference, including cluster- and voxel-level FWE correction, false-discovery rate thresholding and also uncorrected inference combined with extent thresholds, have been proposed and frequently used in the context of ALE analyses without informed recommendation on which one is most appropriate.

In the present paper, we characterized the behavior of ALE analyses and addressed the two questions raised in the last paragraph by means of >120,000 datasets simulated using realistic parameters and distributions and assessed using the current ALE algorithm. From the analyses of these simulated datasets, which should be well representative of those that may be encountered in future ALE projects, we formulate two main recommendations. First, cluster-level family wise-error thresholding represents the inference approach of choice, while voxel-level FWE thresholding is adequate but very conservative. In turn, uncorrected or voxel-level FDR inference should be avoided. Second, in order to avoid results that are dominated by one or two individual experiments and to have sufficient power to detect moderately sized effects, ALE analyses should at least be based on ~20 experiments, preferentially more.

Acknowledgments

This study was supported by the Deutsche Forschungsgemeinschaft (DFG, EI 816/4-1, LA 3071/3-1; EI 816/6-1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain" and the European Union Seventh Framework Program (FP7/2007–2013) under grant agreement no. 604102).

Appendix A. Behavioral domains and paradigm classes for the BrainMap generated datasets

BD: Emotion/PC: Face Monitor/Discrimination → 90 Experiments.
BD: Emotion/PC: Cued Explicit Recognition → 35 Experiments.
BD: Emotion/PC: Passive Viewing → 80 Experiments.
BD: Emotion/PC: Visuospatial Attention → 29 Experiments.
BD: Emotion/PC: Emotion Induction → 136 Experiments.
BD: Emotion/PC: Encoding → 38 Experiments.
BD: Emotion/PC: Film Viewing → 46 Experiments.
BD: Emotion/PC: Task Switching → 48 Experiments.
BD: Emotion/PC: Music Comprehension/Production → 22 Experiments.
BD: Emotion/PC: Theory of Mind → 26 Experiments.
BD: Emotion/PC: Affective Pictures → 40 Experiments.
BD: Emotion/PC: Deception → 46 Experiments.
BD: Emotion/PC: Delay Discounting → 39 Experiments.
BD: Cognition.Attention/PC: Reward → 46 Experiments.
BD: Cognition.Attention/PC: Finger Tapping/Button Press → 25 Experiments.
BD: Cognition.Attention/PC: Cued Explicit Recognition → 22 Experiments.
BD: Cognition.Attention/PC: Visuospatial Attention → 71 Experiments.
BD: Cognition.Attention/PC: Go/No-Go → 175 Experiments.
BD: Cognition.Attention/PC: Encoding → 40 Experiments.
BD: Cognition.Attention/PC: Stroop → 142 Experiments.
BD: Cognition.Attention/PC: Task Switching → 111 Experiments.
BD: Cognition.Attention/PC: Classical Conditioning → 73 Experiments.
BD: Cognition.Attention/PC: Wisconsin Card Sorting Test → 46 Experiments.
BD: Cognition.Attention/PC: Affective Pictures → 39 Experiments.
BD: Cognition.Attention/PC: Flanker → 28 Experiments.
BD: Cognition/PC: Face Monitor/Discrimination → 29 Experiments.
BD: Cognition/PC: Finger Tapping/Button Press → 24 Experiments.
BD: Cognition/PC: Counting/Calculation → 118 Experiments.
BD: Cognition/PC: Film Viewing → 25 Experiments.
BD: Cognition/PC: Task Switching → 39 Experiments.
BD: Cognition/PC: Delay Discounting → 39 Experiments.
BD: Cognition.Language.Semantics/PC: Word Generation (Covert) → 114 Experiments.
BD: Cognition.Language.Semantics/PC: Reasoning → 44 Experiments.

BD: Cognition.Language.Semantics/PC: Reading (Covert) → 37 Experiments.
BD: Cognition.Language.Semantics/PC: Word Generation (Overt) → 65 Experiments.
BD: Cognition.Language.Semantics/PC: Naming (Overt) → 94 Experiments.
BD: Cognition.Language.Semantics/PC: Naming (Covert) → 74 Experiments.
BD: Action.Execution/PC: Face Monitor/Discrimination → 23 Experiments.
BD: Action.Execution/PC: Visual Distractor/Visual Attention → 24 Experiments.
BD: Action.Execution/PC: Go/No-Go → 22 Experiments.
BD: Action.Execution/PC: Flexion/Extension → 126 Experiments.
BD: Action.Execution/PC: Saccades → 100 Experiments.
BD: Action.Execution/PC: Sequence Recall/Learning → 24 Experiments.
BD: Action.Execution/PC: Grasping → 33 Experiments.
BD: Action.Execution/PC: Chewing/Swallowing → 34 Experiments.
BD: Action.Execution/PC: Pointing → 27 Experiments.
BD: Cognition.Language.Speech/PC: Word Generation (Covert) → 140 Experiments.
BD: Cognition.Language.Speech/PC: Reading (Covert) → 20 Experiments.
BD: Cognition.Language.Speech/PC: Reading (Overt) → 87 Experiments.
BD: Cognition.Language.Speech/PC: Word Generation (Overt) → 84 Experiments.
BD: Cognition.Language.Speech/PC: Naming (Overt) → 74 Experiments.
BD: Cognition.Language.Speech/PC: Pitch Monitor/Discrimination → 20 Experiments.
BD: Cognition.Language.Speech/PC: Naming (Covert) → 73 Experiments.
BD: Cognition.Language.Speech/PC: Recitation/Repetition (Overt) → 24 Experiments.
BD: Cognition.Memory.Explicit/PC: Semantic Monitor/Discrimination → 44 Experiments.
BD: Cognition.Memory.Explicit/PC: Encoding → 92 Experiments.
BD: Cognition.Memory.Explicit/PC: Paired Associate Recall → 125 Experiments.
BD: Cognition.Memory.Explicit/PC: Episodic Recall → 61 Experiments.
BD: Cognition.Memory.Explicit/PC: Imagined Objects/Scenes → 24 Experiments.
BD: Cognition.Memory.Working/PC: Face Monitor/Discrimination → 20 Experiments.
BD: Cognition.Memory.Working/PC: Encoding → 55 Experiments.
BD: Perception.Vision.Shape/PC: Face Monitor/Discrimination → 61 Experiments.
BD: Perception.Vision.Shape/PC: Passive Viewing → 38 Experiments.
BD: Perception.Vision.Shape/PC: Visuospatial Attention → 47 Experiments.
BD: Perception.Vision.Shape/PC: Mental Rotation → 121 Experiments.
BD: Perception.Audition/PC: Passive Listening → 105 Experiments.
BD: Perception.Audition/PC: Phonological Discrimination → 26 Experiments.
BD: Perception.Audition/PC: Tone Monitor/Discrimination → 69 Experiments.
BD: Perception.Audition/PC: Pitch Monitor/Discrimination → 47 Experiments.
BD: Perception.Somesthesis/PC: Tactile Monitor/Discrimination → 115 Experiments.
BD: Perception.Somesthesis/PC: Transcranial Magnetic Stimulation → 44 Experiments.
BD: Perception.Somesthesis/PC: Acupuncture → 31 Experiments.
BD: Perception.Vision/PC: Visual Distractor/Visual Attention → 97 Experiments.
BD: Perception.Vision/PC: Visuospatial Attention → 26 Experiments.
BD: Action.Inhibition/PC: Go/No-Go → 155 Experiments.
BD: Action.Inhibition/PC: Anti-Saccades → 27 Experiments.
BD: Cognition.Reasoning/PC: Semantic Monitor/Discrimination → 45 45 Experiments.
BD: Cognition.Reasoning/PC: Reasoning → 134 Experiments.
BD: Cognition.Reasoning/PC: Wisconsin Card Sorting Test → 39 Experiments.
BD: Perception.Vision.Motion/PC: Saccades → 107 Experiments.
BD: Perception.Vision.Motion/PC: Anti-Saccades → 27 Experiments.
BD: Cognition.Social Cognition/PC: Theory of Mind → 82 Experiments.
BD: Cognition.Social Cognition/PC: Deception → 50 Experiments.
BD: Cognition.Language.Phonology/PC: Phonological Discrimination → 112 Experiments.
BD: Cognition.Space/PC: Visuospatial Attention → 43 Experiments.
BD: Cognition.Space/PC: Mental Rotation → 108 Experiments.
BD: Action.Execution.Speech/PC: Reading (Overt) → 56 Experiments.
BD: Action.Execution.Speech/PC: Word Generation (Overt) → 35 Experiments.
BD: Action.Execution.Speech/PC: Recitation/Repetition (Overt) → 48 Experiments.
BD: Perception.Gustation/PC: Taste → 85 Experiments.
BD: Cognition.Language.Orthography/PC: Reading (Covert) → 57 Experiments.
BD: Cognition.Language.Orthography/PC: Orthographic Discrimination → 50 Experiments.
BD: Emotion.Fear/PC: Face Monitor/Discrimination → 69 Experiments.
BD: Action.Imagination/PC: Imagined Movement → 81 Experiments.
BD: Emotion.Happiness/PC: Face Monitor/Discrimination → 60 Experiments.
BD: Cognition.Language.Syntax/PC: Semantic Monitor/Discrimination → 46 Experiments.
BD: Action.Observation/PC: Face Monitor/Discrimination → 26 Experiments.
BD: Action.Observation/PC: Film Viewing → 36 Experiments.
BD: Action.Observation/PC: Action Observation → 47 Experiments.
BD: Cognition.Music/PC: Music Comprehension/Production → 71 Experiments.
BD: Interoception.Sexuality/PC: Passive Viewing → 20 Experiments.
BD: Interoception.Sexuality/PC: Film Viewing → 37 Experiments.
BD: Emotion.Disgust/PC: Passive Viewing → 31 Experiments.
BD: Cognition.Soma/PC: Mental Rotation → 28 Experiments.
BD: Perception.Olfaction/PC: Olfactory Monitor/Discrimination → 61 Experiments.
BD: Cognition.Memory/PC: Emotion Induction → 32 Experiments.
BD: Cognition.Memory/PC: Encoding → 38 Experiments.
BD: Cognition.Memory/PC: Affective Pictures → 40 Experiments.
BD: Emotion.Anger/PC: Face Monitor/Discrimination → 30 Experiments.
BD: Interoception.Bladder/PC: Micturition → 30 Experiments.
BD: Action.Motor Learning/PC: Sequence Recall/Learning → 20 Experiments.

References

Amanzio, M., Benedetti, F., Porro, C.A., Palermo, S., Cauda, F., 2013. Activation likelihood estimation meta-analysis of brain correlates of placebo analgesia in human experimental pain. Hum. Brain Mapp. 34, 738–752.
Amunts, K., Hawrylycz, M.J., Van Essen, D.C., Van Horn, J.D., Harel, N., Poline, J.B., De Martino, F., Bjaalie, J.G., Dehaene-Lambertz, G., Dehaene, S., Valdes-Sosa, P., Thirion,

B., Zilles, K., Hill, S.L., Abrams, M.B., Tass, P.A., Vanduffel, W., Evans, A.C., Eickhoff, S.B., 2014. Interoperable atlases of the human brain. NeuroImage 99, 525–532.
Bandettini, P.A., 2012. Twenty years of functional MRI: the science and the stories. NeuroImage 62, 575–588.
Bludau, S., Bzdok, D., Gruber, O., Kohn, N., Riedl, V., Sorg, C., Palomero-Gallagher, N., Müller, V.I., Hoffstaedter, F., Amunts, K., 2015. Medial prefrontal aberrations in major depressive disorder revealed by Cytoarchitectonically informed voxel-based morphometry. Am. J. Psychiatr.
Bullmore, E., 2012. The future of functional MRI in clinical medicine. NeuroImage 62, 1267–1271.
Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafo, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376.
Bzdok, D., Langner, R., Caspers, S., Kurth, F., Habel, U., Zilles, K., Laird, A., Eickhoff, S.B., 2011. ALE meta-analysis on facial judgments of trustworthiness and attractiveness. Brain Struct. Funct. 215, 209–223.
Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A.R., Langner, R., Eickhoff, S.B., 2012. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. Brain Struct. Funct. 217, 783–796.
Bzdok, D., Eickenberg, M., Grisel, O., Thirion, B., Varoquaux, G., 2015. Semi-supervised factored logistic regression for high-dimensional neuroimaging data. Adv. Neural Inf. Proces. Syst. 3330–3338.
Carp, J., 2012a. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. Front. Neurosci. 6, 149.
Carp, J., 2012b. The secret lives of experiments: methods reporting in the fMRI literature. NeuroImage 63, 289–300.
Chan, R.C., Di, X., McAlonan, G.M., Gong, Q.Y., 2011. Brain anatomical abnormalities in high-risk individuals, first-episode, and chronic schizophrenia: an activation likelihood estimation meta-analysis of illness progression. Schizophr. Bull. 37, 177–188.
Chase, H.W., Kumar, P., Eickhoff, S.B., Dombrovski, A.Y., 2015. Reinforcement learning models and their neural correlates: an activation likelihood estimation meta-analysis. Cogn Affect. Behav. Neurosci. 15, 435–459.
Cheung, M.W., Ho, R.C., Lim, Y., Mak, A., 2012. Conducting a meta-analysis: basics and good practices. Int. J. Rheum. Dis. 15, 129–135.
Chumbley, J.R., Friston, K.J., 2009. False discovery rate revisited: FDR and topological inference using Gaussian random fields. NeuroImage 44, 62–70.
Cieslik, E.C., Mueller, V.I., Eickhoff, C.R., Langner, R., Eickhoff, S.B., 2015. Three key regions for supervisory attentional control: evidence from neuroimaging meta-analyses. Neurosci. Biobehav. Rev. 48, 22–34.
Clos, M., Rottschy, C., Laird, A.R., Fox, P.T., Eickhoff, S.B., 2014. Comparison of structural covariance with functional connectivity approaches exemplified by an investigation of the left anterior insula. NeuroImage 99, 269–280.
Derrfuss, J., Mar, R.A., 2009. Lost in localization: the need for a universal coordinate database. NeuroImage 48, 1–7.
Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC press.
Eickhoff, S.B., Bzdok, D., 2013. Meta-Analyses in Basic and Clinical Neuroscience: State of the Art and Perspective fMRI. Springer, pp. 77–87.
Eickhoff, S.B., Laird, A.R., Grefkes, C., Wang, L.E., Zilles, K., Fox, P.T., 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, S.B., Jbabdi, S., Caspers, S., Laird, A.R., Fox, P.T., Zilles, K., Behrens, T.E., 2010. Anatomical and functional connectivity of cytoarchitectonic areas within the human parietal operculum. J. Neurosci. 30, 6409–6421.
Eickhoff, S.B., Bzdok, D., Laird, A.R., Kurth, F., Fox, P.T., 2012. Activation likelihood estimation meta-analysis revisited. NeuroImage 59, 2349–2361.
Evans, A.C., Janke, A.L., Collins, D.L., Baillet, S., 2012. Brain templates and atlases. NeuroImage 62, 911–922.
Fox, P.T., Lancaster, J.L., Laird, A.R., Eickhoff, S.B., 2014. Meta-analysis in human neuroimaging: computational modeling of large-scale databases. Annu. Rev. Neurosci. 37, 409–434.
Friston, K.J., Holmes, A., Poline, J.-B., Price, C.J., Frith, C.D., 1996. Detecting activations in PET and fMRI: levels of inference and power. NeuroImage 4, 223–235.
Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15, 870–878.
Glatard, T., Lewis, L.B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., Rioux, P., Rousseau, M.E., Sherif, T., Deelman, E., Khalili-Mahani, N., Evans, A.C., 2015. Reproducibility of neuroimaging analyses across operating systems. Front. Neuroinform. 9, 12.
Goodkind, M., Eickhoff, S.B., Oathes, D.J., Jiang, Y., Chang, A., Jones-Hagata, L.B., Ortega, B.N., Zaiko, Y.V., Roach, E.L., Korgaonkar, M.S., Grieve, S.M., Galatzer-Levy, I., Fox, P.T., Etkin, A., 2015. Identification of a common neurobiological substrate for mental illness. JAMA Psychiatry 72, 305–315.
Hardwick, R.M., Rottschy, C., Miall, R.C., Eickhoff, S.B., 2013. A quantitative meta-analysis and review of motor learning in the human brain. NeuroImage 67, 283–297.
Hardwick, R.M., Lesage, E., Eickhoff, C.R., Clos, M., Fox, P., Eickhoff, S.B., 2015. Multimodal connectivity of motor learning-related dorsal premotor cortex. NeuroImage 123, 114–128.
Haxby, J.V., 2012. Multivariate pattern analysis of fMRI: the early beginnings. NeuroImage 62, 852–855.
Herz, D.M., Eickhoff, S.B., Lokkegaard, A., Siebner, H.R., 2014. Functional neuroimaging of motor control in Parkinson's disease: a meta-analysis. Hum. Brain Mapp. 35, 3227–3237.
Jones, D.R., 1995. Meta-analysis: weighing the evidence. Stat. Med. 14, 137–149.
King, M., Rauch, H.G., Stein, D.J., Brooks, S.J., 2014. The handyman's brain: a neuroimaging meta-analysis describing the similarities and differences between grip type and pattern in humans. NeuroImage 102 (Pt 2), 923–937.
Kober, H., Wager, T.D., 2010. Meta-analysis of neuroimaging data. Wiley Interdiscip. Rev. Cogn. Sci. 1, 293–300.
Kohn, N., Eickhoff, S.B., Scheller, M., Laird, A.R., Fox, P.T., Habel, U., 2014. Neural network of cognitive emotion regulation—an ALE meta-analysis and MACM analysis. NeuroImage 87, 345–355.
Kriegeskorte, N., 2009. Relating population-code representations between man, monkey, and computational models. Front. Neurosci. 3, 363–373.
Kurth, F., Zilles, K., Fox, P.T., Laird, A.R., Eickhoff, S.B., 2010. A link between the systems: functional differentiation and integration within the human insula revealed by meta-analysis. Brain Struct. Funct. 214, 519–534.
Laird, A.R., Fox, P.M., Price, C.J., Glahn, D.C., Uecker, A.M., Lancaster, J.L., Turkeltaub, P.E., Kochunov, P., Fox, P.T., 2005. ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts. Hum. Brain Mapp. 25, 155–164.
Laird, A.R., Eickhoff, S.B., Kurth, F., Fox, P.M., Uecker, A.M., Turner, J.A., Robinson, J.L., Lancaster, J.L., Fox, P.T., 2009. ALE meta-analysis workflows via the Brainmap database: progress towards a probabilistic functional brain atlas. Front. Neuroinform. 3, 23.
Laird, A.R., Eickhoff, S.B., Fox, P.M., Uecker, A.M., Ray, K.L., Saenz Jr., J.J., McKay, D.R., Bzdok, D., Laird, R.W., Robinson, J.L., Turner, J.A., Turkeltaub, P.E., Lancaster, J.L., Fox, P.T., 2011a. The BrainMap strategy for standardization, sharing, and meta-analysis of neuroimaging data. BMC Res. Notes 4, 349.
Laird, A.R., Fox, P.M., Eickhoff, S.B., Turner, J.A., Ray, K.L., McKay, D.R., Glahn, D.C., Beckmann, C.F., Smith, S.M., Fox, P.T., 2011b. Behavioral interpretations of intrinsic connectivity networks. J. Cogn. Neurosci. 23, 4022–4037.
Laird, A.R., Eickhoff, S.B., Rottschy, C., Bzdok, D., Ray, K.L., Fox, P.T., 2013. Networks of task co-activations. NeuroImage 80, 505–514.
Lamm, C., Decety, J., Singer, T., 2011. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. NeuroImage 54, 2492–2502.
Langner, R., Rottschy, C., Laird, A.R., Fox, P.T., Eickhoff, S.B., 2014. Meta-analytic connectivity modeling revisited: controlling for activation base rates. NeuroImage 99, 559–570.
Logothetis, N.K., Wandell, B.A., 2004. Interpreting the BOLD signal. Annu. Rev. Physiol. 66, 735–769.
Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., Le Goualher, G., Boomsma, D., Cannon, T., Kawashima, R., Mazoyer, B., 2001. A probabilistic atlas and reference system for the human brain: international consortium for brain mapping (ICBM). Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 356, 1293–1322.
Molenberghs, P., Sale, M.V., Mattingley, J.B., 2012. Is there a critical lesion site for unilateral spatial neglect? A meta-analysis using activation likelihood estimation. Front. Hum. Neurosci. 6, 78.
Muller, V.I., Langner, R., Cieslik, E.C., Rottschy, C., Eickhoff, S.B., 2015. Interindividual differences in cognitive flexibility: influence of gray matter volume, functional connectivity and trait impulsivity. Brain Struct. Funct. 220, 2401–2414.
Nee, D.E., Wager, T.D., Jonides, J., 2007. Interference resolution: insights from a meta-analysis of neuroimaging tasks. Cogn. Affect. Behav. Neurosci. 7, 1–17.
Nickl-Jockschat, T., Janouschek, H., Eickhoff, S.B., Eickhoff, C.R., 2015. Lack of meta-analytic evidence for an impact of COMT Val158Met genotype on brain activation during working memory tasks. Biol. Psychiatry 78, e43–e46.
Palaniyappan, L., Balain, V., Radua, J., Liddle, P.F., 2012. Structural correlates of auditory hallucinations in schizophrenia: a meta-analysis. Schizophr. Res. 137, 169–173.
Poldrack, R.A., 2012. The future of fMRI in cognitive neuroscience. NeuroImage 62, 1216–1220.
Radua, J., Mataix-Cols, D., 2012. Meta-analytic methods for neuroimaging data explained. Biol. Mood Anxiety Disord. 2, 6.
Radua, J., van den Heuvel, O.A., Surguladze, S., Mataix-Cols, D., 2010. Meta-analytical comparison of voxel-based morphometry studies in obsessive–compulsive disorder vs other anxiety disorders. Arch. Gen. Psychiatry 67, 701–711.
Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J., Kahn, R.S., Ramsey, N.F., 2007. Test–retest reliability of fMRI activation during prosaccades and antisaccades. NeuroImage 36, 532–542.
Rehme, A.K., Eickhoff, S.B., Rottschy, C., Fink, G.R., Grefkes, C., 2012. Activation likelihood estimation meta-analysis of motor-related neural activity after stroke. NeuroImage 59, 2771–2782.
Rissman, J., Wagner, A.D., 2012. Distributed representations in memory: insights from functional brain imaging. Annu. Rev. Psychol. 63, 101–128.
Robinson, J.L., Laird, A.R., Glahn, D.C., Lovallo, W.R., Fox, P.T., 2010. Metaanalytic connectivity modeling: delineating the functional connectivity of the human amygdala. Hum. Brain Mapp. 31, 173–184.
Rosen, B.R., Savoy, R.L., 2012. fMRI at 20: has it changed the world? NeuroImage 62, 1316–1324.
Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A.R., Schulz, J.B., Fox, P.T., Eickhoff, S.B., 2012. Modelling neural correlates of working memory: a coordinate-based meta-analysis. NeuroImage 60, 830–846.
Rottschy, C., Caspers, S., Roski, C., Reetz, K., Dogan, I., Schulz, J.B., Zilles, K., Laird, A.R., Fox, P.T., Eickhoff, S.B., 2013. Differentiated parietal connectivity of frontal regions for "what" and "where" memory. Brain Struct. Funct. 218, 1551–1567.
Salimi-Khorshidi, G., Smith, S.M., Keltner, J.R., Wager, T.D., Nichols, T.E., 2009. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. NeuroImage 45, 810–823.
Schilbach, L., Bzdok, D., Timmermans, B., Fox, P.T., Laird, A.R., Vogeley, K., Eickhoff, S.B., 2012. Introspective minds: using ALE meta-analyses to study commonalities in the neural correlates of emotional processing, social & unconstrained cognition. PLoS One 7, e30920.

Turkeltaub, P.E., Eden, G.F., Jones, K.M., Zeffiro, T.A., 2002. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. NeuroImage 16, 765–780.
Turkeltaub, P.E., Eickhoff, S.B., Laird, A.R., Fox, M., Wiener, M., Fox, P., 2012. Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. Hum. Brain Mapp. 33, 1–13.
Wager, T.D., Lindquist, M., Kaplan, L., 2007. Meta-analysis of functional neuroimaging data: current and future directions. Soc. Cogn. Affect. Neurosci. 2, 150–158.
Wager, T.D., Lindquist, M.A., Nichols, T.E., Kober, H., Van Snellenberg, J.X., 2009. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. NeuroImage 45, S210–S221.
Weinberger, D.R., Radulescu, E., 2015. Finding the elusive psychiatric "lesion" with 21st-century neuroanatomy: a note of caution. Am. J. Psychiatry (appiajp201515060753).
Woo, C.W., Krishnan, A., Wager, T.D., 2014. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. NeuroImage 91, 412–419.
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. Nat. Methods 8, 665–670.
Yuan, Y., Brown, S., 2015. Drawing and writing: an ALE meta-analysis of sensorimotor activations. Brain Cogn. 98, 15–26.