NeuroImage 91 (2014) 412–419

Contents lists available at ScienceDirect

NeuroImage

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

Technical Note

Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations

Choong-Wan Woo, Anjali Krishnan, Tor D. Wager *

Department of Psychology and Neuroscience, University of Colorado Boulder, USA
Institute of Cognitive Science, University of Colorado Boulder, USA

ARTICLE INFO

Article history:
Accepted 30 December 2013
Available online 8 January 2014

Keywords:
Cluster-extent thresholding
Multiple comparisons
fMRI
Primary threshold
Family-wise error rate
False discovery rate
SPM
FSL
Gaussian random fields

ABSTRACT

Cluster-extent based thresholding is currently the most popular method for multiple comparisons correction of statistical maps in neuroimaging studies, due to its high sensitivity to weak and diffuse signals. However, cluster-extent based thresholding provides low spatial specificity; researchers can only infer that there is signal *somewhere* within a significant cluster and cannot make inferences about the statistical significance of specific locations within the cluster. This poses a particular problem when one uses a liberal cluster-defining primary threshold (i.e., higher *p*-values), which often produces large clusters spanning multiple anatomical regions. In such cases, it is impossible to reliably infer which anatomical regions show true effects. From a survey of 814 functional magnetic resonance imaging (fMRI) studies published in 2010 and 2011, we show that the use of liberal primary thresholds (e.g., $p < .01$) is endemic, and that the largest determinant of the primary threshold level is the default option in the software used. We illustrate the problems with liberal primary thresholds using an fMRI dataset from our laboratory ($N = 33$), and present simulations demonstrating the detrimental effects of liberal primary thresholds on false positives, localization, and interpretation of fMRI findings. To avoid these pitfalls, we recommend several analysis and reporting procedures, including 1) setting primary $p < .001$ as a default lower limit; 2) using more stringent primary thresholds or voxel-wise correction methods for highly powered studies; and 3) adopting reporting practices that make the level of spatial precision transparent to readers. We also suggest alternative and supplementary analysis methods.

Introduction

Recent advances in the statistical analysis of functional magnetic resonance imaging (fMRI) data have improved the ability of researchers to make meaningful inferences about task-related brain activation. Most statistical analyses of fMRI data are mass univariate approaches, with inferences at a voxel or cluster (of voxels) level. Typical fMRI analyses include >80,000 voxels, resulting in numerous statistical tests, which must be appropriately corrected for multiple comparisons (Bennett et al., 2009; Friston et al., 1994; Genovese et al., 2002; Nichols, 2012; Nichols and Hayasaka, 2003; Nichols and Holmes, 2002).

Among the many approaches to deal with multiple comparisons, cluster-extent based thresholding has become the most popular (Fig. 1A; Friston et al., 1994; also see Carp, 2012). This approach detects statistically significant clusters on the basis of the number of contiguous voxels whose voxel-wise statistic values lie above a pre-determined primary threshold. Tests for statistical significance do not control the estimated false positive probability of each voxel in the contiguous region, but instead control the estimated false positive probability of the region as a whole. Cluster-extent based thresholding generally consists of two stages (Friston et al., 1994; Hayasaka and Nichols, 2003). First, an arbitrary voxel-level *primary threshold* defines clusters by retaining groups of suprathreshold voxels. Second, a cluster-level *extent threshold*, measured in units of contiguous voxels (k), is determined based on the estimated distribution of cluster sizes under the null hypothesis of no activation in any voxel in that cluster. The cluster-level extent threshold that controls family-wise error rate (FWER) can be obtained from the sampling distribution of the largest null hypothesis cluster size among suprathreshold voxels within the search area (e.g., the brain). The sampling distribution of the largest null cluster size under the global null hypotheses of no signal is typically estimated using theoretical methods (e.g., random field theory [RFT]; Worsley et al., 1992), Monte Carlo simulation (Forman et al., 1995), or nonparametric methods (Nichols and Holmes, 2002).

Cluster-extent based thresholding has certain advantages. First, voxel-level corrections for multiple comparisons, such as the Bonferroni and RFT-based corrections, are so stringent that they can dramatically increase Type II errors (i.e., low sensitivity) without extremely large sample sizes (Nichols and Hayasaka, 2003). By contrast, cluster-extent

* Corresponding author at: Department of Psychology and Neuroscience, University of Colorado Boulder, 345 UCB, Boulder, CO 80309-0345, USA.
E-mail address: Tor.Wager@Colorado.Edu (T.D. Wager).

1053-8119/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.12.058

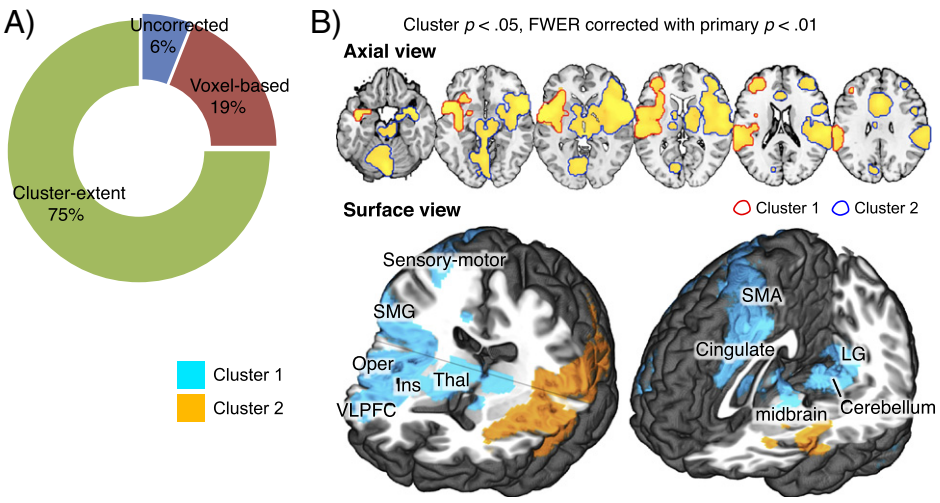

Fig. 1. (A) The proportions of studies using cluster-extent based correction, voxel-based correction, and uncorrected threshold. This result shows that cluster-extent based thresholding is the most popular threshold method among the correction methods. The survey included fMRI studies from Cerebral Cortex, Nature, Nature Neuroscience, NeuroImage, Neuron, PNAS, and Science ($N = 814$). (B) An illustration of potential pitfalls of cluster-extent based thresholding. The presented maps are thresholded at $p < .05$, family-wise error rate (FWER) corrected with cluster-extent based thresholding method with a low cluster-defining primary threshold, $p < .01$. Cluster-extent threshold ($k > 611$) was estimated by Gaussian Random Field method implemented in SPM8. The image shows brain activity induced by experimental thermal pain (from Wager et al., 2013). Two large clusters that contained multiple anatomical regions survived (the red outline [axial]/orange cluster [surface]: Cluster 1, the blue outline [axial]/cyan cluster [surface]: Cluster 2). More than 11 anatomical regions are contained in Cluster 2, including ventrolateral prefrontal cortex (VLPFC), insular cortex (Ins), operculum (Oper), thalamus (Thal), supramarginal gyrus (SMG), sensory–motor cortex, supplementary motor area (SMA), cingulate cortex, cerebellum, lingual gyrus (LG), and multiple midbrain regions. We can only infer that there is true signal "*somewhere*" in this huge cluster and cannot make an inference about specific anatomical regions.

based thresholding has relatively high sensitivity (Friston et al., 1994; Smith and Nichols, 2009). Second, cluster-extent based thresholding accounts for the fact that individual voxel activations are not independent of the activations of their neighboring voxels, especially when the data are spatially smoothed (Friston, 2000; Heller et al., 2006; Wager et al., 2007).

Despite these strengths, cluster-extent based thresholding also has limitations; specifically, low spatial specificity when clusters are large (Friston et al., 1994; Nichols, 2012). The cluster-level *p*-value does not determine the statistical significance of activation at a specific location or voxel(s) within the cluster. Rather, it describes the probability of obtaining a cluster of a given size or greater under the null hypothesis. The logical alternative when this sharp null is rejected is a diffuse family of alternatives: At least *some* signal must be present *somewhere* in the cluster. Therefore, the larger the clusters become, the less spatially specific the inference. Though widely known, we believe the practical implications of this limitation have been largely overlooked.

If cluster sizes are small enough and lie within a single anatomical area of interest, cluster-extent based inferences are reasonably specific. However, if a liberal (i.e., higher *p*-values) primary voxel-level threshold (e.g., $p < .01$) is selected to define clusters, clusters that survive a cluster-extent based threshold for a FWER correction often become large enough to cross anatomical boundaries, particularly in the presence of spatially correlated physiological noise. It is tempting to set a liberal primary threshold in small, underpowered studies, because with more liberal primary thresholds, significant clusters are larger and thus appear more robust and substantial. However, a liberal primary threshold poses a disadvantage in the spatial specificity of claims that can be made. Here, we argue that the use of liberal primary thresholds is both endemic and detrimental to the neuroimaging field.

There are two distinct problems with setting a liberal primary threshold and accepting the reduction in spatial specificity that it entails. First, liberal primary thresholds render the relatively high spatial resolution of fMRI useless, and if significant clusters cross multiple anatomical boundaries, the results yield little useful neuroscientific information. Findings of "activity in the insula *or* the striatum" are not useful in building a cumulative understanding of human brain function. The second, and more pernicious, problem is that results are displayed as colored maps of voxels that pass the primary threshold, with only large-enough clusters retained. These maps invite readers (and authors) to mistakenly believe that significant results are found in *all* the voxels and *all* the anatomical regions depicted as 'significant' in figures. In fact, if a single cluster covers two anatomical regions, the authors cannot in good faith discuss findings in relation to *either* anatomical region, although this is common practice.

In addition to the standard cluster-extent based thresholding methods we discuss extensively here, several recent alternatives have been proposed, including the threshold-free cluster enhancement (TFCE) method (Smith and Nichols, 2009) and hierarchical false discovery rate (FDR) control on clusters (Benjamini and Heller, 2007). TFCE eliminates the need for setting an arbitrary cluster-defining primary threshold by combining voxel-wise statistics with local spatial support underneath the voxel. However, TFCE is also subject to the same limitations of low spatial specificity when significant clusters are large. Benjamini and Heller's (2007) hierarchical FDR method tests clusters first, and then trims locations with no signal within each significant cluster. However, this method heavily depends on a priori information about the data, such as pre-defined clusters or weights, which is generally unavailable in practice.

In this paper, we show a typical example of fMRI results thresholded with a cluster-extent based thresholding method, using an fMRI dataset from our laboratory ($N = 33$), in order to illustrate problems with spatial specificity and inappropriate inferences about anatomical regions. Next, we present findings from a survey of recent fMRI literature ($N = 814$ studies) to demonstrate how researchers currently select the primary threshold levels for their studies. Third, we present results of simulations examining the effects of selection of different primary threshold levels with different levels of signal-to-noise ratio on voxel- and cluster-level false positives (Type I error) and false negatives (Type II error) and on the average anatomical specificity of significant clusters. Finally, we conclude with recommendations for the use of cluster-extent based thresholding in neuroimaging studies.

Methods

Illustration

To illustrate the potential pitfalls of cluster-extent based thresholding, we used fMRI data ($N = 33$) from a study conducted in our laboratory

(Wager et al., 2013). The data include voxel-wise mapping of the positive effects of heat intensity causing acute experimental pain. For more details about the data, please refer to the Methods section of Study 2 in Wager et al. (2013). The results we report here were thresholded with a primary threshold of voxel-wise $p < .01$, which yielded a cluster-extent based threshold of $k > 611$ (cluster-level $p < .05$ FWER corrected). The cluster-extent based threshold was calculated with the Gaussian random field (GRF) method implemented in SPM8 (Wellcome Trust Centre for Neuroimaging, London, UK) using the estimated intrinsic smoothness based on residual images.

Survey

We surveyed original fMRI research papers published between January 2010 and November 2011 from several selected journals (Cerebral Cortex, Nature, Nature Neuroscience, NeuroImage, Neuron, PNAS, and Science). We used "fMRI" and "threshold" as keywords to search for research papers. Exclusion criteria were: (1) non-human studies, (2) lesion studies, (3) studies in which a threshold or correction method could not be clarified, (4) voxel-based morphometry studies, (5) studies primarily about methodology, and (6) machine-learning based studies. We surveyed over 1500 papers and included 814 studies (coded by author CWW). In papers that used several thresholds, we chose the most stringent threshold that was used for whole-brain analyses. For example, if a paper used both voxel-wise Bonferroni correction and cluster-extent thresholding, we counted the paper to use a voxel-wise correction; if a paper used different levels of primary thresholds (e.g., both $p < .001$ and $p < .01$), we counted the paper as using the more stringent (i.e., lower p-value) primary threshold ($p < .001$). To analyze the proportions of primary threshold levels among 607 studies that used cluster-extent based thresholding, we included 484 studies in which we could determine the primary threshold levels (we excluded 123 studies, including papers that provided t-values without degrees of freedom or provided only corrected threshold levels without specifying primary threshold levels).

Simulation 1

In order to show the effects of primary threshold levels on cluster extent sizes (k) for FWER correction of $p < .05$, we estimated k using the GRF method implemented in SPM8 with a range of primary threshold and intrinsic smoothness. Larger cluster thresholds are more likely to yield significant clusters that cross anatomical boundaries and are difficult to interpret. Intrinsic smoothness ranged from 5.0 to 12.0 (full-width at half-maximum [FWHM] in voxels), which are common in fMRI research (see the horizontal dashed lines in Fig. 2A; the dashed lines indicate the voxel-size adjusted intrinsic smoothness levels of 9 existing fMRI studies reported in Nichols and Hayasaka (2003)). Primary threshold level inputs (z) ranged from 2.3 to 3.1, corresponding to $p < .01$ and $p < .001$ (from left to right), which include three most common primary threshold levels (see the arrows in Fig. 2A). For the simulation space, we used a brain mask that contained 328,798 voxels of $2 \times 2 \times 2$ mm^3 voxel size.

Simulation 2

To show the effects of primary threshold levels on voxel- and cluster-level false positives and false negatives, we conducted another simulation in which we added simulated true signals to a real fMRI dataset of 215 time series images from 23 participants (from Wager et al., 2009), which served as noise. True events were generated with a uniform random distribution across time and convolved with SPM's canonical double-gamma hemodynamic response function, independent of the task effects of the original study (speech preparation) in order to simulate a typical event-related design with realistic fMRI noise (see Fig. 3). fMRI images used to provide noise in the simulations

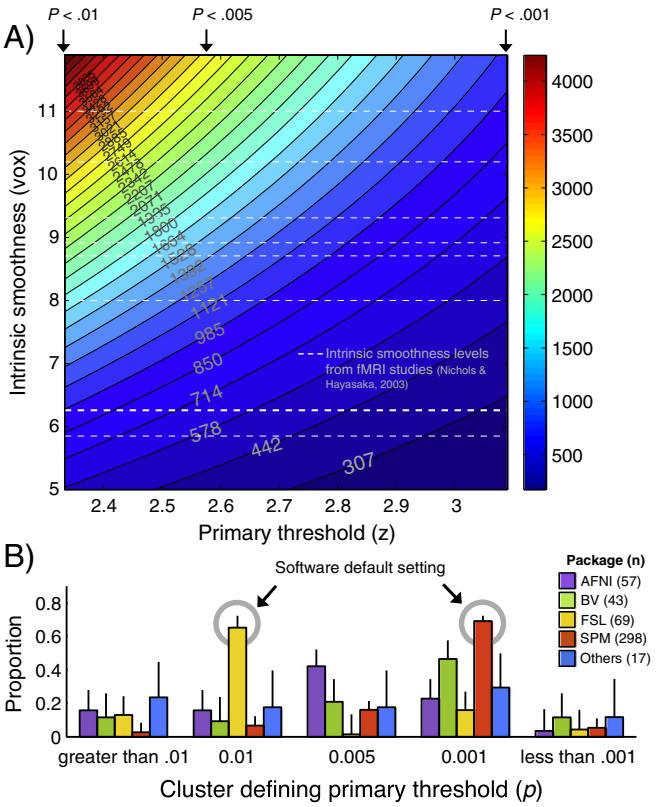

Fig. 2. (A) A contour map of cluster extent size (k) for FWER corrected $p < .05$ (based on cluster-extent thresholding using Gaussian random field method implemented in SPM) as a function of primary threshold (z) and the intrinsic smoothness level (FWHM in voxels). Arrows indicate the three most popular primary threshold levels, $p < .01$ ($z = 2.33$), $p < .005$ ($z = 2.58$), and $p < .001$ ($z = 3.09$), from left to right. Horizontal dashed lines indicate intrinsic smoothness levels from 9 fMRI studies from Nichols & Hayasaka (2003). (B) The proportions of studies using each level of cluster-defining primary p-value across fMRI analysis software packages. The error bars represent standard error of the mean (S.E.M.) using a binomial distribution. The default setting of primary threshold is .001 in SPM and .01 in FSL. BV = BrainVoyager. Others includes Freesurfer, fMRISTAT, LIPSIA, XBAM, and papers that do not specify the software. $n_{AFNI} = 57$, $n_{BV} = 43$, $n_{FSL} = 69$, $n_{SPM} = 298$, and $n_{Others} = 17$.

were preprocessed in a standard way using SPM8 software (i.e., slice-timing correction, realignment, normalization, and smoothing; see Wager et al., 2009). In addition, covariates related to visual stimulation, vascular blood flow-related signal, and head movement were removed prior to the simulations.

We created a mask of 29 seed regions for true signal within different anatomical regions (Table 1) and randomly assigned 5 different cluster sizes to each seed center using spheres with radii from 6 mm to 10 mm (see Fig. 3, the true signal seeds). Seed regions were chosen to approximate "network-like" patterns of bilateral activation typically found in many neuroimaging studies. Because some sphere regions extended into white matter and/or ventricles, we only included voxels within a gray-matter mask. The total number of voxels within the mask was 2113, which was 2.34% of the simulated whole brain space (90,347 voxels with $3.125 \times 3.125 \times 3$ mm^3 voxel size).

We varied the amplitude of true signal (0.3, 0.6, 0.9, 1.2, and $1.5 \times$ the average standard deviation of noise data across the brain) to examine the effects of different levels of signal-to-noise ratio (SNR), which indicates, in this case, the true signal mean divided by the standard deviation of the noise. We added within- and between-subject Gaussian noise (σ_{within} and $\sigma_{between} = 0.5$ and 0.5) to the true signals. Using these images, we carried out the first-level general linear model analyses and obtained a beta (regression parameter estimate) image for each subject. Then, we conducted second-level analyses using 23 beta images, after smoothing the images with Gaussian kernel of a

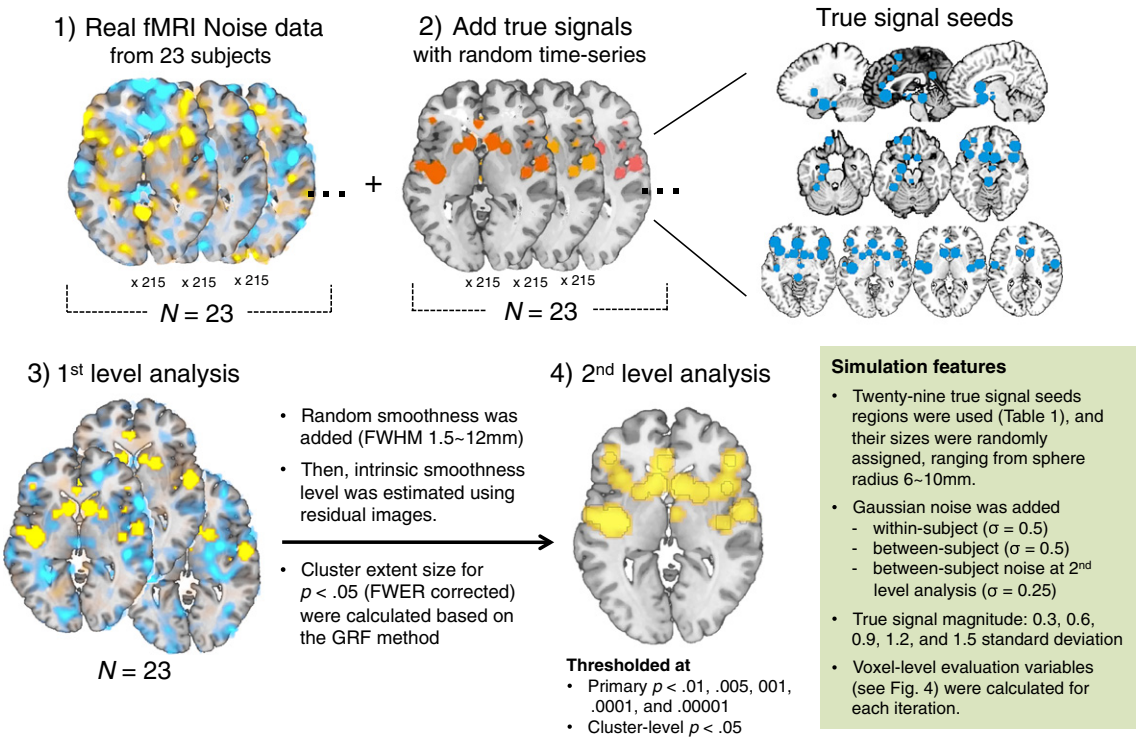

Fig. 3. The Simulation 2 procedure (Simulation 2 section). We conducted 100 iterations of second level analysis for each of 100 iterations of first level analysis (total 10,000 iterations). Real noise data are from Wager et al. (2009); see Simulation 2 section for details. FWER = family-wise error rate; FWHM = full width at half maximum; GRF = Gaussian random field.

randomly chosen size—ranging between 1.5 and 12 mm FWHM to make smoothness levels variable—and added additive white Gaussian noise (mean σ = .25) across the brain. Then, we carried out second-level *t*-tests and thresholded the second-level images using the cluster-extent based thresholding method implemented in SPM8 with five different primary threshold levels ($p < .01$, $p < .005$, $p < .001$, $p < .0001$, and $p < .00001$). For the second-level analyses, we conducted one-tailed *t*-tests (whether a mean is greater than zero) because SPM's GRF method has been developed for the one-tailed test results. We conducted this whole simulation process 10,000 times (i.e., 100 second level iterations for each of 100 first level iterations).

Table 1
Seed regions for true signal in Simulation 2.

	x	y	z
Caudate (L)	−10	16	6
Caudate (R)	10	16	8
Dorsal anterior cingulate cortex (ACC)	−4	16	40
Hippocampus	−20	−6	−22
Inferior parietal lobe	−48	−42	44
Fusiform gyrus	−34	−40	−20
Inferior frontal gyrus (L)	−44	36	−8
Anterior insular cortex (L)	−36	20	−8
Dorsal parietal insular cortex (L)	−42	−14	2
Putamen (L)	−22	10	−2
Supramarginal gyrus (L)	−56	−44	28
Orbitofrontal cortex	−10	44	−20
Orbitofrontal cortex	−28	44	−16
Parahippocampal cortex	−24	−24	−20
Periaqueductal gray	−2	−30	−10
Posterior cingulate cortex	−4	−48	28
Inferior frontal gyrus (R)	46	34	−8
Anterior insular cortex (R)	44	12	−8
Dorsal parietal insular cortex (R)	40	−14	8
Putamen (R)	24	10	−2
Subgenual ACC	−2	32	−6
Rostral ACC	−4	38	10
Rostral–dorsal ACC	−4	24	22
SII (L)	−56	−6	8
SII (R)	56	−4	8
Striatum (L)	−10	14	−10
Striatum (R)	10	12	−12
Supplementary motor area	−4	10	58
Thalamus	4	−4	−8

Note: For Simulation 2 (Simulation 2 section), we created a mask of 29 seed regions for true signal and randomly assigned 5 different cluster sizes, ranging from sphere radii 6–10 mm (Fig. 3).

We adopt the notational convention of Nichols and Hayasaka (2003), to define measures of Type I and II errors for voxels and clusters. As shown in Table 2, V denotes the number of voxels that are tested and C denotes the number of clusters that are tested. $V_{\cdot|0}$ is the number of truly inactive voxels (i.e., voxels with a true null hypothesis), and $V_{\cdot|1}$ is the number of truly active voxels (i.e., voxels with a false null hypothesis). $V_{1|0}$ is the number of false positive voxels (i.e., truly inactive voxels that are falsely rejected), and $V_{1|1}$ is the number of truly active voxels that are correctly rejected. $V_{1|\cdot}$ is the total number of rejected voxels. For the cluster level, $C_{1|0}$ is the number of truly inactive clusters that are falsely rejected, where a truly inactive cluster is defined as one that contains no truly active voxels. In addition, as defined in Nichols and Hayasaka (2003), $I_{\{A\}}$ is the indicator function for an event A such that $I_{\{V_{1|\cdot}>0\}}$ is 1 when $V_{1|\cdot} > 0$, and 0 when $V_{1|\cdot} = 0$.

Using these notations, we define the following three measures to evaluate the effects of primary threshold levels (Table 3). First, we define the voxel-level expected false discovery rate (vFDR) as the expected value of the false discovery proportion, which is the proportion of falsely rejected voxels (i.e., false voxel discoveries) among all rejected voxels. Second, we define the voxel-level sensitivity as the expected value of the proportion of truly active voxels that have been correctly rejected. Third, we define the cluster-level family-wise error rate (cFWER) as the probability of observing a family-wise error, which occurs when there are one or more false positive clusters per map (i.e., $C_{1|0} > 0$). In our simulations, cFWER was estimated by calculating the proportion of the maps that contain at least one false positive cluster over simulation iterations. Importantly, cFWER is what is controlled by the cluster-

Table 2
Classification of voxels (V) and clusters (C) in a thresholded map with cluster-extent thresholding.

Hypothesis	Fail to reject null (non-significant)	Reject null (significant)	Total
Truly inactive (null true)	$V_{0\|0}$, $C_{0\|0}$	$V_{1\|0}$, $C_{1\|0}$	$V_{\cdot\|0}$, $C_{\cdot\|0}$
Truly active (alternative true)	$V_{0\|1}$, $C_{0\|1}$	$V_{1\|1}$, $C_{1\|1}$	$V_{\cdot\|1}$, $C_{\cdot\|1}$
	$V_{0\|\cdot}$, $C_{0\|\cdot}$	$V_{1\|\cdot}$, $C_{1\|\cdot}$	V, C

extent based thresholding procedure; thus, a valid correction method at $p < .05$ cluster-extent corrected should yield the estimated cFWER < .05.

Results

An illustration of potential pitfalls of cluster-extent based thresholding: low spatial specificity and inappropriate inferences (Illustration and Survey sections)

As Fig. 1A shows, cluster-extent based thresholding has been the most popular thresholding method for multiple comparisons correction in recent years. However, there are potential pitfalls of cluster-extent based thresholding, as illustrated in Fig. 1B. The presented map is thresholded at $p < .05$, FWER corrected using cluster-extent based thresholding ($k > 611$) with primary threshold of $p < .01$. As expected, voxels in multiple anatomical regions that have been implicated in pain processing showed significant positive activations, including the insular cortex, anterior cingulate cortex, secondary somatosensory cortex, thalamus, and midbrain. Given this map, authors and readers could easily infer that thermal pain activates all these pain-processing regions. However, all these individual anatomical regions are contained in two large clusters, and one of them (the *blue* cluster in Fig. 1B) contains more than 11 distinct anatomical regions ("regions" here means large-scale divisions typically respected in neuroimaging studies, e.g., putamen vs. anterior insula). With cluster-extent based thresholding methods, we cannot accurately make inferences about any of these specific anatomical regions, but can only conclude that there is true signal *somewhere* within the large cluster. The low anatomical specificity caused by large activation clusters renders this map neuroscientifically ambiguous and could potentially mislead readers.

The level of primary threshold is crucial to the size of significant clusters (Survey and Simulation 1 sections)

Cluster extent size (k) for FWER corrected $p < .05$ is determined by the primary threshold and intrinsic smoothness (Friston et al., 1994). The primary threshold defines clusters of suprathreshold voxels, and the intrinsic smoothness determines the distribution of suprathreshold cluster sizes under the null hypothesis. As Fig. 2A indicates, more liberal primary thresholds and higher smoothness increase the cluster extent size needed to pass the threshold for reporting. Particularly, when the primary threshold is liberal, cluster extent size steeply increases as smoothness increases. A smoothness of 8.3 voxel FWHM (in $2 \times 2 \times 2$ mm^3 voxels) is common in neuroimaging studies (this is the average estimated smoothness of 9 fMRI studies reported in Nichols and Hayasaka (2003)). At this smoothness level and a liberal primary threshold ($p < .01$), the required cluster extent size is very large ($k > 1{,}899$), and therefore most suprathreshold clusters are expected to span multiple anatomical regions. With the same smoothness and a more stringent primary threshold ($p < .001$), the threshold is $k > 520$, which is more anatomically constrained. For example, the average size of anatomical regions in the Harvard–Oxford atlas (Desikan et al., 2006) is 995 (at a 50% probability threshold, in $2 \times 2 \times 2$ mm^3 voxels), and the average size of functional parcels from Craddock's 200-region parcellation data (Craddock et al., 2012) is 735 (in $2 \times 2 \times 2$ mm^3 voxels). Thus, with liberal primary thresholds, most reported clusters are expected to span multiple regions and suffer from problems with neuroscientific interpretability. A threshold of $p < .001$ will not guarantee anatomical specificity in all studies, of course, but will be sufficient to identify regions that are localized enough to be anatomically interpretable in many studies.

Our survey results show that the use of a liberal primary threshold of $p < .01$ is endemic (17% of 484 papers where we could determine the primary threshold level) especially in studies using certain software packages that have a liberal primary threshold as a default (65% of 69 papers that used FSL, in which the default primary threshold is $p < .01$, compared to 7% of 298 papers that used SPM, in which the default primary threshold is $p < .001$; Fig. 2B). These survey results suggest that the choice of primary thresholds depends strongly on the defaults of the software packages used for analysis.

The detrimental effects of liberal primary thresholds on the spatial localization of true-signal regions (Simulation 2 section)

Our simulation results show the effects of the primary threshold level in more detail (Fig. 4). In this section, the estimated values are represented with a hat. Consistent with the simulation results in the previous section, as shown in Fig. 4A, the sizes of significant clusters for liberal thresholds (e.g., $p < .01$) are large enough to span multiple anatomical regions (the dashed lines in Fig. 4A present the average cluster sizes of anatomical regions from the Harvard–Oxford atlas [Desikan et al., 2006] and functional parcels from Craddock's 200 parcellation solution [Craddock et al., 2012] in $3.125 \times 3.125 \times 3$ mm^3 voxels). Particularly, when the SNR is high, liberal primary thresholds yield very large clusters.

As presented in Fig. 4B(a), the estimated voxel-level FDR ($\widehat{\text{vFDR}}$) was always unacceptably high across all primary thresholds, indicating that a large proportion of voxels in significant clusters are false positives. This result confirms that cluster extent-corrected maps should not be interpreted as voxel-wise maps. Particularly, $\widehat{\text{vFDR}}$ was highest (ranging from .44 to .71) at the most liberal primary threshold ($p < .01$), and it decreased as primary thresholds became more stringent. The estimated voxel-level sensitivity ($\widehat{\text{vSens}}$; Fig. 4B(b)) increased as primary thresholds became more stringent until $p < .001$ (for SNR = .3) or $p < .0001$ (for SNR = .6, .9, and 1.2). This demonstrates that a more liberal primary

Table 3
Definition of evaluation measures of Simulation 2 (Simulation 2 section).

Level	Measure	Notation	Definition	Description
Voxel	False discovery rate	vFDR	$E([V_{1\|0} / V_{1\cdot}] I_{\{V_{1\cdot}>0\}})$	Expected value of the proportion of false positive voxels among all rejected voxels
	Sensitivity	vSens	$E(V_{1\|1}) / V_{\cdot\|1}$	Expected value of the proportion of truly active voxels that have been correctly rejected
Cluster	Family-wise error rate	cFWER	$P(C_{1\|0} > 0)$	Probability of observing a family-wise error, which occurs when there are one or more false positive clusters per map

Note: We adopted the notational convention of Nichols and Hayasaka (2003) to define the measures (see Table 2 and Simulation 2 section). As defined in Nichols and Hayasaka (2003), $I_{\{A\}}$ is the indicator function for an event A.

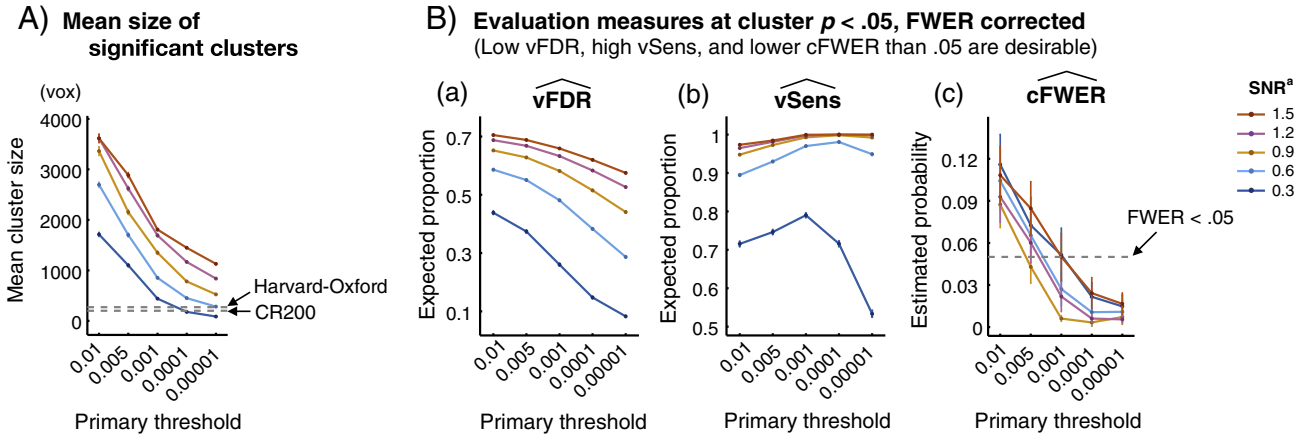

Fig. 4. Simulation results. (A) Mean size of significant clusters at cluster-level $p < .05$ FWER corrected. The horizontal dashed lines indicate the average cluster sizes of anatomical regions of the Harvard–Oxford atlas (>50% probability) and functional parcels of Craddock's 200 parcellation solution (CR200; Craddock et al., 2012). (B) We present (a) the estimated voxel-level false discovery rate ($\widehat{vFDR}$), (b) the estimated voxel-level sensitivity ($\widehat{vSens}$), and (c) the estimated cluster-level family-wise error rate ($\widehat{cFWER}$) from thresholded brain maps at cluster $p < .05$ FWER corrected. For detailed simulation procedures and definitions of the evaluation measures, please refer to Fig. 3, Table 3, and Simulation 2 section. The horizontal dashed line for the ($\widehat{cFWER}$) plot indicates the level that the cluster-extent based thresholding procedure controls. The error bars represent standard error of the mean (S.E.M.). If the error bars are smaller than the markers, they would not be visible. The estimated values are represented with a hat. [a]SNR = signal-to-noise ratio, which indicates the true signal mean divided by the standard deviation of the noise.

threshold is not always beneficial for the detection of signals, even for low SNR data. However, $\widehat{vSens}$ decreased when primary thresholds were more stringent than $p < .001$ or .0001, especially for low SNR data. The decreases of both $\widehat{vFDR}$ and $\widehat{vSens}$ at more stringent primary thresholds (e.g., $p < .00001$) are mainly because stringent primary thresholds reduce the cluster-extent threshold (k), permitting the reporting of smaller, more localized clusters, but causing more true signal voxels to go unreported (i.e., more Type II errors). These results suggest that there could be an optimal primary threshold that depends in part on the SNR and sample size, and higher SNR or larger sample sizes can permit more stringent primary thresholds, which identify focal activations more precisely.

Cluster-level results presented in Fig. 4B(c) show that the ($\widehat{cFWER}$) exceeds .05 (the nominal family-wise error rate) for liberal primary thresholds, indicating an anti-conservative bias. For example, with the primary thereshold $p < .01$, the $\widehat{cFWER}$s for all SNR levels were significantly higher than .05 (range of $\widehat{cFWER}$ = .09–.12, t (99) = 2.3–3.1, p = .001–.013, one-tailed), and with the primary threshold $p < .005$, the $\widehat{cFWER}$ for the highest SNR level (SNR = 1.5) was significantly higher than .05 ($\widehat{cFWER}$ = .085, t (99) = 1.8, p = .037, one-tailed). These cluster-level results suggest that cluster-extent based thresholding does not guarantee $\widehat{cFWER}$ <.05 when liberal primary thresholds are used, especially for high SNR data.

In sum, our simulation results clearly demonstrate the detrimental effects of liberal primary thresholds on voxel- and cluster-wise inferences. Within the range of parameters simulated here, most significant clusters are large enough to span across multiple anatomical regions, and large portions of voxels reported in cluster-extent corrected maps are false positives, particularly when the SNR is high and the primary threshold is liberal. More importantly, with a liberal primary threshold, cluster-extent based thresholding does not accurately control the family-wise error rate. Therefore, studies with higher SNR or larger sample sizes should use more stringent primary thresholds or voxel-wise correction, and studies with liberal primary thresholds are likely to yield maps of limited neuroscientific utility.

Discussion

The popularity of cluster-extent based thresholding is understandable given its advantages, including generally higher sensitivity in identifying significant regions (Friston et al., 1994) as compared to voxel-level correction methods for multiple comparisons. However, there are potential pitfalls, especially when activation clusters are so large that they cover multiple anatomical brain regions. Researchers can only ascertain that there is true signal *somewhere* within the cluster, and thus cannot accurately explain the neuroscientific significance of the findings. In addition, authors and readers are tempted to make inferences about particular regions within the cluster and are misled to incorrect interpretations.

Our survey and simulations show that large, neuroscientifically uninterpretable activation clusters are mainly caused by the use of liberal primary thresholds and high intrinsic smoothness, and that the choice of primary thresholds depends strongly on the defaults of the software packages used for analysis. The results of our simulation examining the effects of different levels of primary thresholds suggest that liberal primary thresholds do not control the cluster-level FWER adequately; in particular, there is an anti-conservative bias in FWER-corrected results when a primary threshold of $p < .01$ is used. In addition, liberal primary thresholds render the resulting maps—which show all the voxels in each significant cluster—less interpretable and more misleading to readers.

With respect to the anti-conservative bias, it has been shown that cluster-extent based thresholding can become anti-conservative when the data violate the assumptions of GRF theory, such as uniform smoothness and the use of a sufficiently stringent primary threshold (Hayasaka et al., 2004; Silver et al., 2011). Specifically, when smoothness varies across the brain, false positive rates are higher than expected under GRF-based FWER control in regions with high smoothness (Hayasaka et al., 2004). In addition, the expected null-hypothesis cluster size is systematically under-estimated by GRF at liberal primary thresholds, resulting in higher false positive rates (Silver et al., 2011). Our simulations were based on real fMRI noise; thus, while they do not give precise estimates of the degree of bias for all possible fMRI datasets, they illustrate that anti-conservative bias is likely at liberal primary thresholds. Thus, liberal primary thresholds (e.g., $p < .01$ or $p < .005$) are not desirable default options for fMRI analysis.

Practical recommendations

Based on these findings, we recommend several analysis and reporting procedures to avoid the issues of low spatial specificity and inappropriate inferences about anatomical regions.

Choice of thresholding method and primary threshold

If studies are sufficiently powered, we recommend not using cluster-extent based thresholding at all, but using voxel-wise correction methods such as FWER and FDR. In addition, if the question is whether two conditions produce overlapping or distinct activation, voxel-wise correction should always be used, as overlap is assessed and interpreted at the voxel level. Furthermore, if the question is whether a specific small anatomical structure (e.g., periaqueductal gray or specific cortical areas like the dorsal posterior insula) is active or not, cluster extent-based thresholding is likely not appropriate, because the significance of the cluster could depend in part on the extent of activation beyond the anatomical region of interest.

For studies with moderate effect sizes and sample sizes (e.g., Cohen's $d < .8$ [Cohen, 1988; d = mean effect across participants divided by standard deviation across participants] and $N < 50$), cluster-extent based thresholding can offer increased sensitivity to detect activations with large spatial extent. In such cases, we recommend using more stringent cluster-defining primary thresholds to reduce the possibility of obtaining false positive clusters and/or large activation clusters, and to improve the degree of confidence in inferences about specific locations/voxels. Based on our simulation results, $p < .001$ is a reasonable default for a range of typical cases. More stringent thresholds may be desirable if more spatial specificity is needed for neuroscientific interpretability, but the primary threshold level should be chosen *a priori* to reduce potential biases towards findings in specific anatomical regions that researchers desire to find. Importantly, primary thresholds more liberal than $p < .001$ (e.g., $p < .01$) is not recommended given the possibility of inaccurate FWER correction.

Visualization of cluster extent and reporting strategies

Many problems with mis-interpretation of cluster-extent thresholded results could be ameliorated by marking the boundaries of each cluster clearly in figures. This can be done by visualizing different clusters in different colors, or with outlines of different colors, when the cluster extent is not clear from the image (e.g., Fig. 1B). In addition, we recommend that figure legends and captions *explicitly* state that the true activation location and extent within each significant cluster cannot be determined.

In addition, the neuroanatomical descriptors used to report and discuss results should be consistent with the level of spatial specificity of the results. If cluster-extent based thresholding (even with a stringent primary threshold) identifies large clusters, general descriptors for the clusters in tables and results (e.g., right forebrain) should be reported, rather than a list of specific regions (anterior insula, claustrum, caudate, etc.). Popular software packages such as SPM, FSL, and AFNI include algorithms for identifying multiple "peak" activations within large clusters and reporting a series of coordinates. However, these "peaks" cannot be used to infer that all of the "peaks" in the table are truly activated. In addition, it cannot be assumed that the coordinates listed in the table are good estimates of the true peak activation locations. Therefore, a *single* descriptor that covers all the suprathreshold voxels in the cluster should be used, as it more accurately reflects the spatial uncertainty inherent in the results.

Alternative and supplementary methods

If one wishes to make inferences about the location of "peak" activation within a cluster, it is desirable to conduct an explicit statistical test of the spatial location of the peak. One way to do this is to identify peak locations within the cluster for individual participants, and then to construct 3-D 95% confidence volumes on the mean peak location (e.g., Wager et al., 2003). Such confidence intervals could be visualized and reported along with cluster-extent based results. Tests of differences in spatial location among two or more conditions can be performed using multivariate analysis of variance (MANOVA) with 3-D coordinate locations as a multivariate dependent variable and condition labels as a predictor (e.g., Johnson and Wichern, 2007; Wager et al., 2004), or with other explicitly spatial models (e.g., Kang et al. 2011).

In some cases, widely distributed activation may be apparent even with stringent primary thresholds or FWER correction, and even extend into white matter. Conversely, there may be cases in which significant results can only be found with liberal thresholds that produce large clusters. These situations are likely to arise if the true activity pattern is not anatomically localized, but rather results from activation of diffuse modulatory systems or similar mechanisms. In such cases, we recommend characterizing activity in terms of components, and explicitly stating that localization cannot be inferred. The use of independent component analysis (ICA; Calhoun et al., 2002) and similar data compression methods are suitable in this case; the underlying generative model for ICA involves sources distributed across many voxels. When using ICA, we recommend that researchers focus on the distributed patterns of voxels and/or report general descriptors, rather than looking for and reporting effects in specific locations loading on components, unless voxel-wise thresholding methods are used that can support inferences about specific regions.

Conclusion

Cluster-extent based thresholding has become the most popular correction method for multiple comparisons in fMRI data analysis because it is more sensitive (more powerful) and reflects the spatially correlated nature of fMRI signal. However, when a significant cluster is so large that it spans multiple anatomical regions, we cannot make inferences about a specific anatomical region with confidence, but we can only infer that there is signal *somewhere* within the large cluster. In other words, even when cluster-level false positive rate is well controlled, large true positive clusters are likely to consist of mostly noise and render the positive findings useless because of its low informativeness. Therefore, the cluster size is crucial to make the cluster-extent thresholded findings interpretable and useful in building a cumulative understanding of human brain function.

With an illustration, survey, and simulations, we showed that the primary threshold level is crucial for determining the cluster size and the valid FWER correction. To avoid the pitfalls in cluster-extent based thresholding, we recommend setting $p < .001$ as a lower limit default, and using more stringent primary thresholds or voxel-wise correction methods for highly powered studies. We also recommend displaying discrete clusters in different colors in figures and explicitly stating caveats about low spatial specificity in figure legends and captions. If the signal is diffuse anyway, we recommend reporting general diffuse descriptors rather than a list of sub-regions within clusters. We also suggest alternative and supplementary methods, such as the visualization of 3-D confidence volumes, MANOVA, and ICA.

Conflict of interest

We have no relevant conflicts of interest.

References

Benjamini, Y., Heller, R., 2007. False discovery rates for spatial signals. J. Am. Stat. Assoc. 102, 1271–1281.
Bennett, C.M., Wolford, G.L., Miller, M.B., 2009. The principled control of false positives in neuroimaging. Soc. Cogn. Affect. Neurosci. 4, 417–422.
Calhoun, V.D., Adali, T., Pearlson, G.D., van Zijl, P.C., Pekar, J.J., 2002. Independent component analysis of fMRI data in the complex domain. Magn. Reson. Med. 48, 180–192.
Carp, J., 2012. The secret lives of experiments: methods reporting in the fMRI literature. NeuroImage 63, 289–300.
Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. Erlbaum, Mahwah, NJ.
Craddock, R.C., James, G.A., Holtzheimer III, P.E., Hu, X.P., Mayberg, H.S., 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. Hum. Brain Mapp. 33, 1914–1928.
Desikan, R.S., Segonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980.

Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., Noll, D.C., 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn. Reson. Med. 33, 636–647.
Friston, K.J., 2000. Experimental design and statistical issues. In: Mazziotta, J.C., Toga, A.W., Frackowiak, R.S.J. (Eds.), Brain Mapping: The Disorders. Academic Press, San Diego, pp. 33–59.
Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., Evans, A.C., 1994. Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp. 1, 210–220.
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.
Hayasaka, S., Nichols, T.E., 2003. Validating cluster size inference: random field and permutation methods. NeuroImage 20, 2343–2356.
Hayasaka, S., Phan, K.L., Liberzon, I., Worsley, K.J., Nichols, T.E., 2004. Nonstationary cluster-size inference with random field and permutation methods. NeuroImage 22, 676–687.
Heller, R., Stanley, D., Yekutieli, D., Rubin, N., Benjamini, Y., 2006. Cluster-based analysis of fMRI data. NeuroImage 33, 599–608.
Johnson, R.A., Wichern, D.W., 2007. Applied Multivariate Statistical Analysis, 6th ed. Pearson, New Jersey.
Kang, J., Johnson, T.D., Nichols, T.E., Wager, T.D., 2011. Meta analysis of functional neuroimaging data via bayesian spatial point processes. J. Am. Stat. Assoc. 106, 124–134.
Nichols, T.E., 2012. Multiple testing corrections, nonparametric methods, and random field theory. NeuroImage 62, 811–815.
Nichols, T., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat. Methods Med. Res. 12, 419–446.
Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.
Silver, M., Montana, G., Nichols, T.E., Neuroimaging, A.D., 2011. False positives in neuroimaging genetics using voxel-based morphometry data. NeuroImage 54, 992–1000.
Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage 44, 83–98.
Wager, T.D., Phan, K.L., Liberzon, I., Taylor, S.F., 2003. Valence, gender, and lateralization of functional brain anatomy in emotion: a meta-analysis of findings from neuroimaging. NeuroImage 19, 513–531.
Wager, T.D., Jonides, J., Reading, S., 2004. Neuroimaging studies of shifting attention: a meta-analysis. NeuroImage 22, 1679–1693.
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., van Ast, V.A., Hughes, B.L., Davidson, M.L., Lindquist, M.A., Ochsner, K.N., 2009. Brain mediators of cardiovascular responses to social threat, part II: prefrontal–subcortical pathways and relationship with anxiety. NeuroImage 47, 836–851.
Wager, T.D., Atlas, L.Y., Lindquist, M.A., Roy, M., Woo, C.-W., Kross, E., 2013. An fMRI-based neurologic signature of physical pain. N. Engl. J. Med. 368, 1388–1397.
Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12, 900–918.