



ELSEVIER

Contents lists available at ScienceDirect

NeuroImage

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

Review

Evaluating the consistency and specificity of neuroimaging data using meta-analysis

Tor D. Wager^{a,*}, Martin A. Lindquist^b, Thomas E. Nichols^{c,d}, Hedy Kober^a, Jared X. Van Snellenberg^a^a Department of Psychology, Columbia University, 1190 Amsterdam Ave, New York, NY, 10027, USA^b Department of Statistics, Columbia University, New York, NY, USA^c Glaxo-Smith-Kline, London, UK^d FMRIB, Oxford University, Oxford, UK

ARTICLE INFO

Article history:

Received 15 September 2008

Revised 22 September 2008

Accepted 15 October 2008

Available online xxxxx

Keywords:

PET

fMRI

Meta-analysis

Neuroimaging

Analysis methods

ABSTRACT

Making sense of a neuroimaging literature that is growing in scope and complexity will require increasingly sophisticated tools for synthesizing findings across studies. Meta-analysis of neuroimaging studies fills a unique niche in this process: It can be used to evaluate the consistency of findings across different laboratories and task variants, and it can be used to evaluate the specificity of findings in brain regions or networks to particular task types. This review discusses examples, implementation, and considerations when choosing meta-analytic techniques. It focuses on the multilevel kernel density analysis (MKDA) framework, which has been used in recent studies to evaluate consistency and specificity of regional activation, identify distributed functional networks from patterns of co-activation, and test hypotheses about functional cortical-subcortical pathways in healthy individuals and patients with mental disorders. Several tests of consistency and specificity are described.

© 2008 Elsevier Inc. All rights reserved.

Contents

27	Introduction	0
28	Why use meta-analysis? Establishing activation consistency	0
29	Why use meta-analysis? Evaluating functional specificity	0
30	Coordinate-based meta-analysis and its many varieties	0
31	Methods	0
32	Section I: The MKDA approach.	0
33	Weighting of study contrast maps and peaks	0
34	Thresholding and multiple comparisons	0
35	Meta-analysis diagnostic plots	0
36	Section II: Analyzing activation specificity	0
37	Comparing two task types using MKDA	0
38	Section III: Testing connectivity	0
39	Section IV: Future directions	0
40	Acknowledgments	0
41	References	0

Introduction

Recent years have seen a rapid increase in the number and variety of investigations of the human brain using neuroimaging techniques. Studies using functional magnetic resonance imaging (fMRI) or

positron emission tomography (PET) have emerged as a major methodology for investigating function in the intact and disordered human brain. Psychological processes under investigation are as diverse as psychology itself, and nearly every major domain of psychology is represented in this growing body of work. Many popular domains—such as cognitive control, working memory, decision-making, language, emotion, and disorders such as attention deficit disorder, schizophrenia, and depression—have been the subject of a large number of neuroimaging studies, whose results can be synthesized

* Corresponding author. Fax: +1 212 854 3609.

E-mail address: tor@psych.columbia.edu (T.D. Wager).

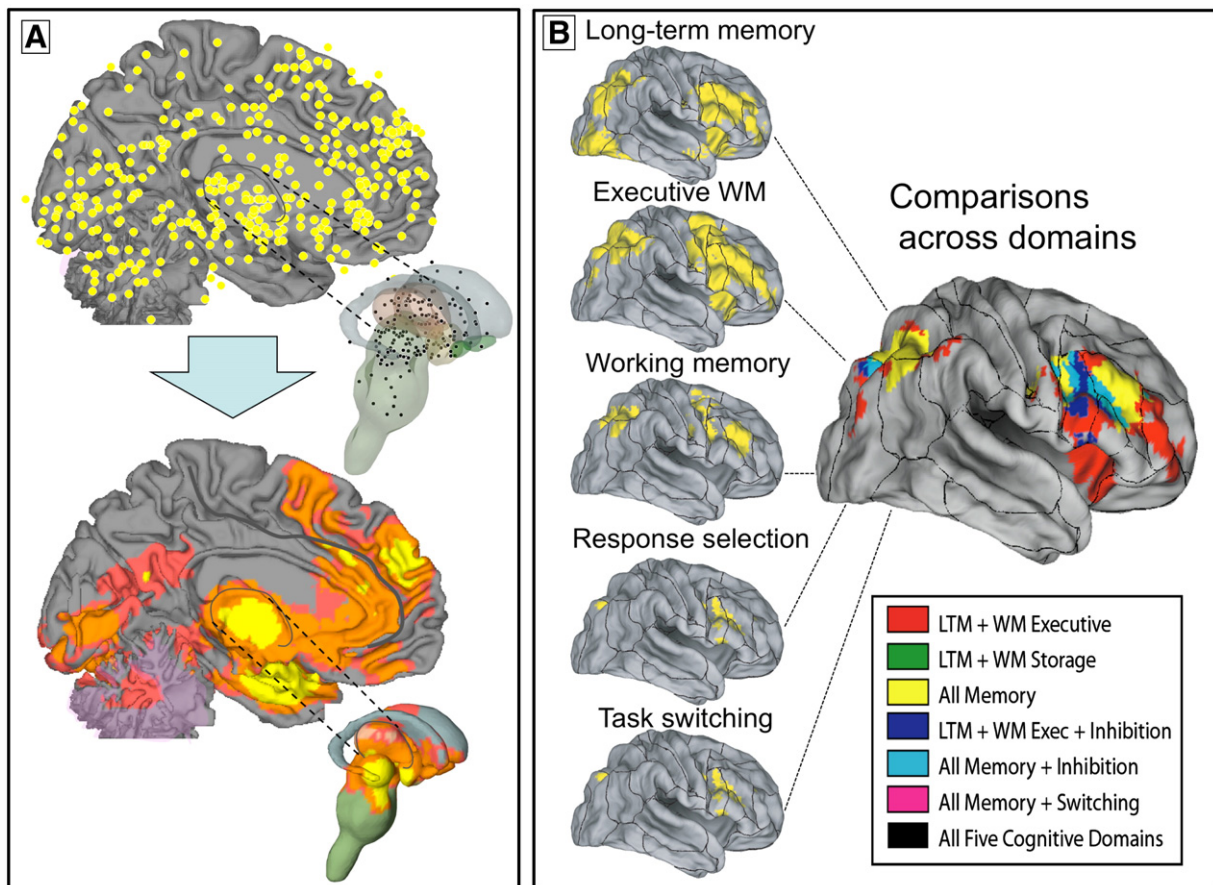


Fig. 1. Examples of results from multilevel kernel density analysis (MKDA). (A) Top panel: Peak activation coordinates from 437 study comparison maps (SCMs; from 163 studies) plotted on medial and subcortical brain surfaces (Wager et al., 2008). Peak locations within 12 mm from the SCM are averaged. Bottom panel: Summary of consistently activated regions across SCMs in the MKDA analysis. Yellow indicates a significant density of SCMs in a local neighborhood, and orange and pink indicate significant density using extent-based thresholding at primary thresholds of 0.001 and 0.01, respectively (see text for details). All results are family-wise error rate corrected at $p < .05$ for search across brain voxels. (B) MKDA results from five published meta-analyses of executive function mapped onto the PALS-B12 atlas (Van Essen, 2005) using Caret software, and the overlap in activations across the five types of executive function, from Van Snellenberg and Wager (2007, p. 71). The results illustrate how meta-analysis can inform about common and differential activations across a variety of psychological processes.

and interpreted in the context of data from lesion studies, electrophysiology, behavioral studies, and related methodologies.

This burgeoning literature places increasing demands on scientists to understand, integrate, and evaluate the neuroimaging work that has been performed in each of these areas. One important set of questions relates to the *consistency*, or replicability across laboratories, scanners, and task variants, of activated regions¹. Which brain regions are consistently associated with domains such as working memory, decision-making, emotional experience, and so on? And where are the boundaries between functional regions that identify studies that do and do not replicate?

Another important set of questions relates to synthesis across areas of study, and in particular the *specificity* of particular activation patterns for particular psychological processes. Is a region typically related to working memory load unique to this domain, or is it shared by a broader set of cognitive demands? For example, some brain regions, such as the left inferior frontal gyrus, are characterized variously as “working memory regions,” “language regions,” “emotion regulation regions,” or “decision making regions”, depending on the functional domain being investigated. Before positing hypotheses about common underlying functions, it is important to establish

¹ We use the term “region” here to refer loosely to an expanse of brain tissue typically considered to be a unit of analysis. What constitutes a functional “region” may vary across disciplines, but the same questions about activation consistency and specificity apply.

whether these different researchers are discussing the same region, or whether nearby activations in different task domains can be reliably discriminated.

Why use meta-analysis? Establishing activation consistency

Meta-analysis fills a unique role in the neuroimaging literature because many of the important, fundamental questions posed above are difficult or impossible to address within individual studies. Therefore, a major use for meta-analysis in imaging is to identify the consistently activated regions across a set of studies. In Fig. 1A, for example, a number of reported activation peaks from many studies (top panel) are summarized in a meta-analysis of consistently reported regions (bottom panel).

Evaluating consistency is important because false positive rates in neuroimaging studies are likely to be higher than in many fields (somewhere in the range of 10–40%; see below). Thus, some of the reported activated locations shown in Fig. 1A are likely to be false positives, and it is important to assess which findings have been replicated and have a higher probability of being real activations.

Inflated false positive rates are a byproduct of the commonly used strategy of making statistical brain maps composed of many tests (“voxel-wise” mapping), combined with the use of small sample sizes (usually 8–25 participants) due to the considerable expense of neuroimaging. Although there is a trend towards larger sample sizes and more rigorous multiple comparisons correction, until recently 100

Table 1

	Database	Studies	Sample size (N)				Reported peaks			
			Total	Median	Min	Max	Total	In	Out	% "replicated"
t1.1	WM storage	26	305	12	5	21	377	225	152	60
t1.2	Executive WM	60	664	10	5	28	1086	867	219	80
t1.3	Emotion	163	2010	11	4	40	2478	2198	280	89
t1.4	Long term memory	166	1877	11	5	33	3265	2950	315	90

"In" refers to peaks within 10 mm of the significant meta-analysis area, and "Out" refers to peaks further than 10 mm from the significant meta-analysis area. WM: Working memory
 "Replicated" peaks are within a consistently activated area in the meta-analysis.

most studies have not corrected for multiple comparisons because they were too under-powered (Nichols and Hayasaka, 2003; Wager et al., 2007a). Many studies that have used corrections used methods whose assumptions are likely to be violated, or *ad hoc* methods that do not control the false positive rate at the nominally specified level (Wager et al., 2007b). Gene arrays in genetic research have the same problem, for the same reasons—though in both fields, the benefits of making whole-brain or whole-genome maps make them preferred choices for many researchers.

Data that illustrate these issues are shown in Table 1, which summarizes the results of four meta-analyses on a total of 415 studies involving 4,856 participants. The meta-analyses were all performed using the same method, multi-level kernel density analysis (MKDA; Kober et al., 2008; Wager et al., 2008; data from Wager and Smith, 2003, was also reanalyzed using MKDA). The median sample size from the included studies range from $N=10$ to $N=12$ across studies of working memory, long-term memory, and emotion. A basic power calculation (see Fig. 12 in Wager et al., in press) shows that with a standard effect size of $d=1$ (Cohen's d , an effect divided by its standard deviation), approximately 45 participants are required to achieve 80% power using Bonferroni correction in a typical whole brain, voxel-wise analysis. Correction methods such as those based on Gaussian Random Field Theory are often just as conservative, but nonparametric correction improves power substantially (Nichols and Hayasaka, 2003). With nonparametric correction, approximately only 30 participants are needed for 80% power (Wager et al., in press), though this sample size is still larger than all but the largest studies in our samples². Thus, performing proper correction is impractical without relatively large sample sizes, but failing to make appropriate corrections leads to increased false positive rates.

The MKDA results can also be used to provide a rough estimate of false positive rates. For each meta-analysis in Table 1, we calculated the number and proportion of peaks reported near (within 10 mm) one of the regions identified as consistently activated in the meta-analysis. The proportion of peaks outside of the consensus regions provides a rough estimate of false positive rates across studies. Table 1 shows an estimated false positive rate around 10% for the larger databases, and 20%–40% for the smaller meta-analyses, which may have been underpowered and therefore failed to find more truly activated regions. Of course, there are a number of reasons why this figure is imprecise; false-positives could contribute to consistently-activated regions, and heterogeneity among studies could result in true positives outside those regions found to be significant in meta-analyses. However, even if imprecise, this figure provides a rough estimate of how big the false-positive problem may be. Using another method based on examining the estimated search space, thresholds, and image smoothness, we previously estimated a false positive rate of roughly 17% (Wager et al., 2007a,b). In sum, there is a need to integrate and validate results across studies.

² With $d=2$, 19 participants yield 80% power with Bonferroni correction, and about 12 participants might be expected to yield 80% power with nonparametric correction. These sample sizes are more in line with those used, and indeed most reported effect sizes.

The simplest goal of a meta-analysis is to provide summaries of the consistency of regional brain activation for a set of studies of a particular task type, providing a consensus about which regions are likely to be truly activated by a given task. In addition, meta-analysis can also be used to extend beyond regional activation to identify groups of consistently co-activated regions that may form spatially distributed functional networks in the brain. We have used this approach to identify distributed groups of functionally related brain regions in emotion (Kober et al., 2008) and anxiety-related disorders (Etkin and Wager, 2007), and other groups have used similar approaches to identifying large-scale functional networks organized around particular cognitive processes (Neumann et al., 2005), or functional co-activation with target brain structures across many tasks (Postuma and Dagher, 2006). Identifying co-activated networks can provide the basis for testing them as units of analysis in individual studies, and can lead to the development of testable hypotheses about functional connectivity in specific tasks.

Why use meta-analysis? Evaluating functional specificity

In addition to establishing consistent activation in one task type, meta-analysis can be used to evaluate the specificity of activation (in regions or 'networks') to one type of task among a set of different tasks. For example, one might identify a set of regions consistently activated by self-referential processes (Northoff et al., 2006), and then ask whether activity in these regions is specific to self-referential processes—that is, that they are not activated by other tasks that do not involve self-reference. This information is critical to using measures of brain activity to predict psychological processes (i.e., making a "reverse inference" that activity in some region implies the involvement of a given psychological process; Poldrack, 2006; Sarter et al., 1996).

Specificity can only be examined across a range of tested alternative tasks: A region that is specific for faces compared with houses may not be specific for faces compared with tools. Likewise, a region that discriminates self-referential word judgments from non-self-referential ones does not imply that the region discriminates self-referential processes from retrieval of semantic knowledge from long-term memory. Unfortunately, different psychological domains are usually studied in isolation, and it is virtually impossible compare a wide range of tasks in a single study. However, meta-analysis provides tools for doing exactly that: Activation patterns can be compared across the entire range of tasks studied using neuroimaging techniques, providing a unique way to evaluate activation specificity across functional domains.

The simplest kind of specificity analysis compares activation patterns among two or more task types, such as positive vs. negative emotion (Phan et al., 2002; Wager et al., 2003), high-conflict vs. low-conflict conditions in cognitive control tasks (Nee et al., 2007), or various types of executive demand in working memory tasks (Wager and Smith, 2003). Many more examples appear in the growing meta-analysis literature, some of which is referenced in Table 2.

However, it is also possible to compare the results of meta-analysis from a number of functional domains, such as the results across 5 different task types shown in Fig. 1. In a recent chapter {Van

Table 2
A sampling of neuroimaging meta analyses

Authors	Year	Method	Psychological focus	
<i>Cognitive control/executive function</i>				
Chein et al.	2002	Density (Gaussian)	Verbal working memory	
Wager et al.	2003	Clustering of peaks, chi-square	Working memory	
Wager et al.	2004	KDA, spatial MANOVA	Attention/task switching	
Buchsbaum et al.	2005	ALE	Wisconsin card sorting	Q1
Chein and Schneider	2005	Density (Gaussian)	Practice effects in cognitive control	Q2
Laird et al.	2005	ALE	Stroop interference	
Neumann et al.	2005	ALE, co-activation "replicator dynamics"	Stroop interference	
Costrafeda et al.	2006	Spatial location	Verbal fluency in left IFG	Q3
Gilbert et al.	2006	Spatial location/Chi-square/classifier	Episodic memory, multitasking mentalizing in BA 10	
Nee et al.	2007	KDA, logistic regression	Cognitive control/interference	
Van Snellenberg and Wager	2008 ^a	MKDA/KDA	Cognitive control and memory	Q4
<i>Emotion and motivation</i>				
Phan et al.	2002	Chi-square within regions	Emotion	
Murphy et al.	2003	Spatial location (K-S test)	Emotion	Q5
Wager et al.	2003	KDA, Chi-square	Emotion	
Kringelbach et al.	2004	Spatial location	Reinforcers in OFC	Q6
Phan et al.	2004	Qualitative	Emotion	Q7
Baas et al.	2004	Chi-square	Amygdala lateralization	Q8
Northoff et al.	2005	Clustering of peaks	Self-referential processes	Q9
Krain et al.	2006	ALE	Decision-making	Q10
Wager et al.	2008	MKDA, Chi-square	Emotion	
Kober et al.	2008 ^a	MKDA, co-activation	Emotion	
<i>Disorder</i>				
Zakzanis et al.	2000	Effect sizes	Schizophrenia	Q11
Zakzanis et al.	2003	Effect sizes	Alzheimer's disease	Q12
Whiteside et al.	2004	Effect sizes	Obsessive-compulsive disorder	Q13
Glahn et al.	2005	ALE	Working memory in schizophrenia	Q14
Fitzgerald et al.	2006	ALE	Depression, DLPFC	Q15
Dickstein et al.	2006	ALE	ADHD	Q16
Van Snellenberg et al.	2006	Effect sizes	Schizophrenia and working memory	
Steele et al.	2007	Spatial location ("unwarped")	Depression, frontal cortex	Q17
Valera et al.	2007	Effect sizes	Brain structure in ADHD	Q18
Etkin and Wager	2007	MKDA, Co-activation	Anxiety disorder	
Hoekert et al.	2007	Effect sizes	Emotional prosody in schizophrenia	Q19
<i>Language</i>				
Turkeltaub et al.	2002	ALE	Single-word reading	
Jobard et al.	2003	Clustering of peaks	Word reading	Q20
Brown et al.	2005	ALE	Speech production	Q21
Vigneau et al.	2006	Clustering peaks	Language, left cortical hemisphere	Q22
Ferstl et al.	2008	ALE, Co-activation "replicator dynamics"	Text comprehension	Q23
<i>Others</i>				
Joseph	2001	Spatial location	Object recognition: category specificity	
Grezes and Decety	2001	Qualitative	Action	Q24
Kosslyn and Thompson	2003	Logistic regression	Visual imagery	
Nielsen et al.	2004	Kernel density/multivariate	Cognitive function	
Gottfried and Zald	2005	Spatial location	Olfaction in OFC	
Nickel and Seitz	2005	Clustering of peaks	Parietal cortex	
Petacchi et al.	2005	ALE	Auditory function, cerebellum	Q25
Lewis	2006	Average maps in CARET	Tool use	
Postuma and Dagher	2006	Co-activation	Basal ganglia	
Zacks	2008		Mental rotation	Q26

A sample of published neuroimaging meta-analyses. See text for abbreviations.

^a Results discussed in relative detail in this paper.

Snellenberg and Wager, in press #71} we examined the overlap in meta-analytic results among studies that isolated cognitive control processes (e.g. task switching and speeded response selection) and studies that involved maintenance of information in working memory (WM), including WM storage, the subtraction of [Executive WM - WM storage], and long-term memory encoding and retrieval. Our working hypothesis was that the more complex memory maintenance and manipulation tasks would involve task switching and response selection, and so would activate a super-set of the areas involved in more elementary cognitive control processes. The illustration in Fig. 1B supports this notion, showing that the inferior frontal junction and pre-supplementary motor area are consistently activated across studies within each task type, but that more rostral portions of the

prefrontal cortex were only consistently activated when WM was involved. The most anterior prefrontal regions were activated only when manipulation of information in memory was required.

Whereas the results in Fig. 1 present a qualitative comparison across five task types that summarize commonalities and differences across types, quantitative analyses of specificity can also be performed using several other methods discussed below. These methods include χ^2 (chi-square) and approximations to multinomial exact tests, analysis of reported peak density differences, and pattern classifier systems. In each analysis, formal predictions can be made about task types given patterns of brain activity. For example, in a particularly interesting application using meta-analytic data, Gilbert et al. (2006) used classifier analyses to identify regions within the medial and

229 orbitofrontal cortices that discriminated different cognitive functions
230 of the anterior frontal cortex. This study is an example of how formal
231 predictions about psychological states can be tested across diverse
232 kinds of studies using meta-analysis.

233 Coordinate-based meta-analysis and its many varieties

234 There are now a number of quantitative meta-analyses of
235 neuroimaging data in the literature, as evidenced by the partial list
236 in Table 2. The vast majority use reported peak coordinates from
237 published studies, which are readily available in published papers and
238 stored electronically in databases such as Brainmap (<http://brainmap.org/>). We refer to this as the “coordinate-based meta-analysis”
239 approach. Alternatively, full statistic maps for each study could be
240 used and effect sizes aggregated at each voxel (Lewis, 2006). Though
241 we consider this to be a “gold standard” approach, and advocate its
242 development in future meta-analytic work, this approach is compli-
243 cated by the lack of readily-available statistic images.

244 Collectively, the coordinate-based meta-analysis literature to date
245 covers a cornucopia of innovative techniques. Some meta-analyses
246 evaluate consistency by combining effect size data (Van Snellenberg et
247 al., 2006) or analyzing the frequencies of reported peaks (Phan et al.,
248 2002) within anatomically defined regions of interest. Variants on this
249 theme use multiple logistic regression (Kosslyn and Thompson, 2003;
250 Nee et al., 2007) or summarize co-activations among regions (Etkin
251 and Wager, 2007; Nielsen et al., 2004). A popular approach to
252 examining specificity has been to analyze the locations of coordinates
253 in stereotaxic space, testing for functional gradients or spatial
254 distinctions (Gottfried and Zald, 2005; Joseph, 2001), and sometimes
255 extending these analyses to perform space-based classification of
256 study types using MANOVA (Joseph, 2001; Wager et al., 2004) or
257 cluster analyses using χ^2 tests (Nickel and Seitz, 2005; Northoff et al.,
258 2006; Wager and Smith, 2003).

259 While the procedures above refer to analyses carried out on pre-
260 defined anatomical areas, the most popular approaches for summar-
261 izing reported coordinates from neuroimaging studies involve so-
262 called “voxel-wise” analyses, or the construction of statistical maps
263 summarizing peak coordinates in a neighborhood around each voxel
264 in a standard brain (Chein et al., 2002; Fox et al., 1999). At their heart,
265 these kernel-based methods are related to kernel-based methods for
266 analyzing the multivariate distributions of sparse data, and essentially
267 summarize the evidence for activation in a local neighborhood around
268 each voxel in a standard atlas brain. They are popular because they
269 provide ways of summarizing activations across the brain without
270 imposing rigid prior constraints based on anatomical boundaries,
271 which are currently difficult to specify precisely.

272 Our goal in the remainder of this paper is to describe recent
273 advances and applications using this kernel-based approach. We focus
274 in particular on MKDA, a recently developed extension of voxel-wise
275 meta-analysis approaches, for example activation likelihood estima-
276 tion (ALE; Turkeltaub et al., 2002) and kernel
277 density analysis (KDA; Wager et al., 2007b). The essence of the
278 approach is to reconstruct a map of significant regions for each study
279 (or statistical contrast map within study), and analyze the consistency
280 and specificity across studies in the neighborhood of each voxel.

281 In Section 1, we describe how MKDA can be used to evaluate the
282 consistency of activations. We consider issues of level of analysis
283 (peak vs. study contrast map), weighting, thresholding, and multiple
284 comparisons, and show the results of simulations comparing ALE,
285 KDA, and MKDA methods. We also show how this approach lends
286 itself to the construction of analogues to some meta-analysis plots in
287 the traditional meta-analytic literature, in particular logistic funnel
288 plots. In Section 2, we consider how MKDA can be used to analyze
289 specificity. We consider a) density difference maps to compare
290 activations in two types of studies, and b) A multinomial permutation
291 test—an alternative to the χ^2 test with several desirable properties—for

292 comparing two or more study types. Finally, in Section 3, we describe
293 extensions of the MKDA approach to analyze co-activations across
294 regions, including clustering and mediation analysis on co-activation
295 data to develop models of functional pathways. 296

297 Methods

298 Section 1. The MKDA approach

299 The MKDA method analyzes the distribution of peak coordinates
300 from published studies across the brain. The technique, used in several
301 recent published analyses (Etkin and Wager, 2007; Kober et al., 2008;
302 Wager et al., 2008, 2007b) is summarized in Fig. 2. Essentially, the
303 reported x (left–right), y (posterior–anterior), and z (inferior–superior)
304 coordinates in a standard stereotaxic space (i.e., Montreal Neurologi-
305 cal Institute space) are treated as a sparse representation of activated
306 locations. In the literature, peak coordinates are reported in reference
307 to a particular statistical contrast map (SCM); for example, a study
308 might compare high memory load vs. low memory load. Studies may
309 report results from multiple contrast maps (e.g., load effects for verbal
310 stimuli and load effects for object stimuli), so we refer to the maps as
311 SCMs rather than as study maps.

312 To integrate peaks across space, the peaks obtained from each SCM
313 are convolved with a spherical kernel of radius r , which is user-
314 defined, and thresholded at a maximum value of 1 so that multiple
315 nearby peaks are not counted as multiple activations (left side of Fig.
316 2). Formally, this amounts to the construction of an indicator map for
317 each SCM, where a voxel value of 1 indicates a peak in the neighbor-
318 hood, while 0 indicates the absence of a peak, i.e. for each voxel k :

$$I_k = \begin{cases} 1 & \text{if } \sqrt{\sum_{i=1}^3 (v_k - \bar{x}_i)^2} \leq r \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

319 where v_k is the $[x, y, z]$ triplet in mm for voxel k 's location in MNI
320 space, and \bar{x} is the $[x, y, z]$ triplet for the nearest reported peak. The
321 choice of r is somewhat arbitrary, but should be related to the degree
322 of consistency found across studies. Better inter-study consistency
323 would allow for meaningful neighborhood summaries using smaller
324 values of r and would thus allow for higher-resolution meta-analytic
325 results. In practice, $r=10$ mm is commonly used, which provides a
326 good balance between sensitivity and spatial resolution (Wager et al.,
327 2004). 328

329 A weighted average of the resulting indicator maps provides a
330 summary map with an interpretable metric: The (weighted) number
331 of nominally independent SCM indicator maps that activate in the
332 neighborhood of each voxel. The weights relate to measures of study
333 quality and are described below. The convenient interpretation of the
334 statistic (an SCM activation count) motivates the use of the spherical
335 kernel, though in principle other kernels (such as a Gaussian kernel)
336 could be used. Information about the extent and shape of the
337 activation summarized by each peak could be incorporated as well,
338 but in practice, inconsistency in reporting this information across
339 studies has prevented it from being used.

340 The final step is to establish a statistical threshold for determining
341 what constitutes a significant number of activating SCMs in a local
342 area. The threshold is determined using a Monte Carlo procedure, and
343 a natural null hypothesis is that the ‘activated’ regions in the SCM
344 indicator maps are not spatially consistent; that is, they are distributed
345 randomly throughout the brain. The procedure is described in detail
346 below.

347 Thus, in MKDA, the peak locations are not analyzed directly. Rather,
348 indicator maps for each SCM are constructed, and the SCM is treated as
349 the unit of analysis. Thus, the metric used to summarize consistency is
350 not directly related to how many peaks were reported near a voxel—
351 after all, the peaks could all have come from one study—but rather, is
352 related to how many SCMs activated near a voxel. 352

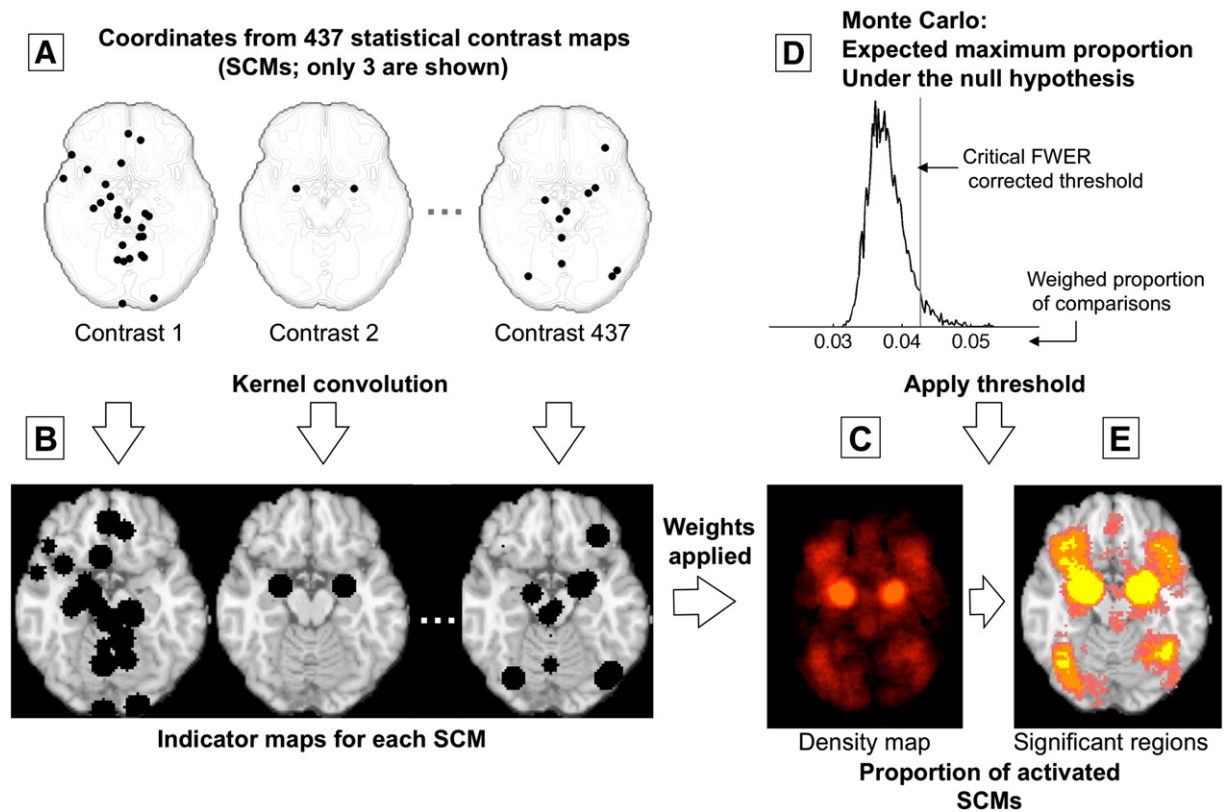


Fig. 2. Example procedures for multilevel kernel density analysis (MKDA). (Adapted from Wager et al. (2007), Fig. 3). (A) Peak coordinates of three of the 437 comparison maps included in a meta-analysis of emotion. Peak coordinates of each map are separately convolved with the kernel, generating (B) indicator maps for each study contrast map (SCM). (C) The weighted average of the indicator maps is compared with (D) the maximum proportion of activated comparison maps expected under the null hypothesis in a Monte Carlo simulation and (E) thresholded to produce a map of significant results. Color key is as in Fig. 1.

Q27

This is the primary difference between MKDA and previously used “voxel-wise” approaches, including KDA and ALE (Fox et al., 1998; Laird et al., 2005; Nielsen et al., 2005; Wager et al., 2003, 2004). The latter methods also summarize peak coordinates using a kernel-based approach, but they do not take into account which SCM or study the peaks came from. Thus, the KDA and ALE measures do not summarize consistency across studies; rather, they summarize consistencies across peak coordinates. Interpreting these methods as reflecting consistency across studies requires the implicit assumption that there are no true inter-study differences in number and location of peaks, smoothness, false positive rates, and statistical power. This assumption is clearly violated in most meta-analyses that integrate data from many laboratories, and the consequence is that a significant KDA or ALE ‘meta-analytic’ result can be driven by a single study. Thus, one cannot conclude from a significant KDA/ALE p -value that a new study on the same topic would be likely to activate similar regions. This issue is analogous to the fixed versus random-effects model issue in individual functional imaging studies, in which fixed-effects models treat observations (time points) as the unit of analysis and ignore inter-subject variability, while random-effects models account for this variability by treating subjects as the unit of analysis. The fixed-effects issue has also received considerable discussion in the traditional meta-analysis literature, and Monte Carlo simulations have demonstrated that when there is true between-study variability, fixed-effects models have inflated Type I error rates, particularly for meta-analysis of a small number of studies (Hedges and Vevea, 1998).

An analogy to a standard clinical study may help clarify this point. Not modeling SCM as a level of analysis is akin to ignoring the fact that different observations in a clinical study came from different participants; thus, the analysis and inference procedures would be identical whether the observations came from a group of participants

or only a single participant. For example, examine the peaks in Fig. 2A, which are 3 representative contrast maps from a set of 437 used in recent meta-analyses of emotion (Kober et al., 2008; Wager et al., 2008). Imagine that we performed a meta-analysis only on the plotted peaks from these three studies. Because study is ignored in the ALE/KDA analysis, information about which study contributed each of the peaks is not preserved, and all the peaks are combined. Contrast 1 contributes 26 peaks, many of them very close together, whereas Contrast 2 contributes only two. When the KDA map is generated and thresholded in this example, three peaks within 10 mm are required to achieve significance in the meta-analysis. Study 1 has enough peaks near the amygdala to generate significant results by itself. This is quite a plausible scenario due to differences in scanning, analysis, and reporting procedures across studies; and, in fact, the data shown are real.

This example illustrates some of the advantages to treating SCM or study, rather than peak, as the unit of analysis. A study may report peaks either very densely or sparsely, depending on reporting standards and the smoothness of statistical images. Smaller studies tend to produce rougher (less smooth) statistic images, because they average over fewer subjects. Rougher statistic images produce a topography consisting of many local peaks; thus, there is a tendency for smaller studies to report more local peaks! Clearly, it is disadvantageous to consider each peak as an independent observation.

In summary, the MKDA procedure has several important advantages over previously used voxel-wise meta-analysis approaches. First, other approaches have typically analyzed the peak locations from a set of studies, ignoring the nesting of peaks within contrasts. MKDA takes account of the multi-level nature of the data. Second, the MKDA statistic has a straightforward interpretation: the proportion of contrasts (P) activating within r mm of each voxel. Third, contrast

415 maps are weighted based on sample size and study quality. And
 416 finally, the procedure controls the family-wise error rate (FWE), or the
 417 chance of observing a false positive anywhere in a meta-analytic brain
 418 map, and so each significant MKDA result can be interpreted. We
 419 elaborate on these latter points of comparison below.

420 Weighting of study contrast maps and peaks

421 Weighting by sample size and/or study quality is typical in meta-
 422 analysis across fields (DerSimonian and Laird, 1986), and incorporat-
 423 ing sample size into the meta-analysis is a key feature of standard
 424 meta-analytic methods, because the precision of a study's estimates
 425 ($1/\text{standard error}$) are proportional to the square root of the study's
 426 sample size. Weighting in meta-analysis ensures that larger and more
 427 rigorously performed studies exert more influence on the meta-
 428 analytic results. However, there are several choices to be made in
 429 deciding how to weight activation peaks from neuroimaging studies.
 430 One choice is whether to weight individual peaks by their reliability
 431 (i.e., Z -scores), individual SCMs, or both. Weighting peaks by their Z -
 432 scores may seem like a good idea at first glance, but there are
 433 significant disadvantages. First, Z -scores from different studies may
 434 not be comparable because of the different analysis methods used. For
 435 example, some (mostly older) studies treat subjects as a fixed effect,
 436 whereas others treat it as a random effect. "Fixed effects" analyses do
 437 not appropriately model inter-subject variance and therefore do not
 438 allow for inferences about a population—a critical part of scientific
 439 inference. Z -scores from fixed-effects studies are systematically higher
 440 than those from random-effects study. Second, the massive multiple
 441 testing framework employed in most neuroimaging studies creates a
 442 situation in which peaks with the highest Z -scores may have occurred
 443 by chance. For an analogy, consider the survivors from the Titanic. On
 444 average, those who survived were better swimmers, but another
 445 major component was luck. Here, reported significant voxels in a
 446 study are the "survivors." This situation causes the well-known
 447 phenomenon of regression to the mean: Z -scores corresponding to
 448 these peaks regress toward their true values in a replication. Thus, it
 449 may not be safe to assume that Z -scores from a group of studies are
 450 comparable.

451 Rather than weighting Z -scores, the current version of MKDA
 452 weights by the square root of the sample size for each SCM. In
 453 addition, we down-weight studies using fixed effects analyses by
 454 a factor of 0.75, an arbitrary value that reduces the impact of fixed-
 455 effects studies. These factors are combined into the following
 456 weighting equation:

$$457 P = \sum_c I_c \left(\frac{\delta_c \sqrt{N_c}}{\sum_c \delta_c \sqrt{N_c}} \right) \quad (2)$$

458 where P is the weighted proportion of activated comparisons (SCM
 459 indicators), c indexes comparison maps I , δ is the fixed effects
 460 discounting factor, and N is the sample size. This approach could
 461 potentially be expanded to weight peaks within study by their relative
 462 activation strength within the study, and thereafter weight studies in
 463 proportion to their sample size. In addition, this weighting scheme
 464 could be used to weight by other study quality measures developed by
 465 the analyst, such as diagnostic criteria or sample-matching procedures
 466 employed in studies of psychiatric or medical populations. While the
 467 precise weight values assigned for various study characteristics are
 468 necessarily somewhat arbitrary, assigning higher weights to higher-
 469 quality studies is generally preferable to ignoring differences in study
 470 quality or excluding some studies altogether. However, because
 471 weighting by study quality involves choices by the analyst that can
 472 be somewhat arbitrary, it is common in the traditional meta-analysis
 473 literature to additionally report the results of an unweighted analysis
 474 (Rosenthal and DiMatteo, 2001) so that the influence of the weighting
 475 procedure on the results can be assessed.

Thresholding and multiple comparisons

476

The null hypothesis in MKDA, like KDA and ALE analyses, is a
 477 "global" null hypothesis stating that there is no coherent spatial
 478 consistency across SCMs (or reported peaks, for KDA and ALE).
 479 Rejecting the null technically implies that there are one or more
 480 neighborhoods (regions) with consistent reports. However, the test
 481 still provides a test with strong control of FWE, in the sense that under
 482 the null hypothesis, the chances of a false positive *anywhere* in the
 483 brain is α (e.g., $p < .05$, corrected for search across the brain).
 484 Considering an alternative null conditional on one or more consistent
 485 regions, it can be shown that the required density to achieve FWE
 486 control for remaining regions is lower than the required density for
 487 the global null. Thus, the test is over-conservative, and KDA analysis
 488 incorporated a step-down test (Wager et al., 2004) that has not yet
 489 been implemented in MKDA. 490

In practice, MKDA uses a threshold derived from Monte Carlo
 491 simulation of the global null. Contiguous clusters of activated voxels
 492 are identified for each SCM, and the cluster centers are randomized
 493 within gray-matter (plus an 8 mm border) in the standard brain. For
 494 each iteration (although results typically stabilize after about 2000
 495 iterations, we typically use 10,000), the maximum MKDA statistic (P in
 496 Eq. (2)) over the whole brain is saved. As with other nonparametric
 497 FWE correction methods, the $(1-\alpha)$ th percentile of the distribution of
 498 maxima provides a critical statistic value. 499

An advantage to randomizing cluster locations, rather than peak
 500 locations, is that the density of peaks in a particular study will not
 501 have an undue influence on the null hypothesis values in the Monte
 502 Carlo simulation. Even if peaks are reported very densely, the MKDA
 503 Monte Carlo threshold will not be influenced as long as peaks are
 504 reported within the same activated area. This is not true for peak-
 505 coordinate based Monte Carlo simulations (i.e., KDA and ALE), and
 506 thus dense peak reporting will lead to higher thresholds for reporting
 507 significant meta-analytic results and less power. 508

In addition, in MKDA an 'extent-based' thresholding can be used
 509 (Wager et al., 2008), paralleling methods available in the popular
 510 Statistical Parametric Mapping software (Friston et al., 1996). In our
 511 MKDA implementation, we have established primary thresholds at
 512 the average uncorrected $(1-\alpha)$ th percentile of the MKDA statistic
 513 across the brain (with permuted blobs, i.e., under null hypothesis
 514 conditions), where α is by default .001, .01, and .05. The maximum
 515 extent of contiguous voxels at this threshold is saved for each iteration
 516 of the Monte Carlo simulation, and the critical number of contiguous
 517 voxels is calculated from the resulting distribution of maximum null-
 518 hypothesis spatial extents. For example, the yellow regions in Figs. 1A
 519 and 2 are significant at $p < .05$ MKDA-height corrected, whereas
 520 orange and pink regions are significant at $p < .05$ cluster-extent
 521 corrected with primary thresholds of .001 and .01, respectively. 522

Meta-analysis diagnostic plots

523

Traditional meta-analyses often make extensive use of diagnostic
 524 plots to illustrate the sensibility (or lack thereof) of results across a
 525 group of studies. For example, the Galbraith plot shows the relation-
 526 ship between effect size (e.g., Z -scores, y -axis) and study precision
 527 (x -axis) (Egger et al., 1997). Precision is equal to $1/\text{standard error}$ for
 528 each study, which is related to the residual standard deviation and
 529 square root of the study sample size (N). Simple regression is used to
 530 analyze the relationship between precision and effect size. A reliable
 531 non-zero effect across studies should have a positive slope in the
 532 Galbraith plot, because the more precise studies (with lower standard
 533 errors) should have higher Z -scores (Fig. 3). This plot can be used to
 534 detect bias of several types. If there is no bias, the intercept of the plot
 535 should pass through the origin: With a precision of zero (e.g., zero
 536 sample size), the predicted effect size should be zero. A positive
 537 intercept indicates small-sample bias. 538

Executive working memory: Adapted Galbraith plots

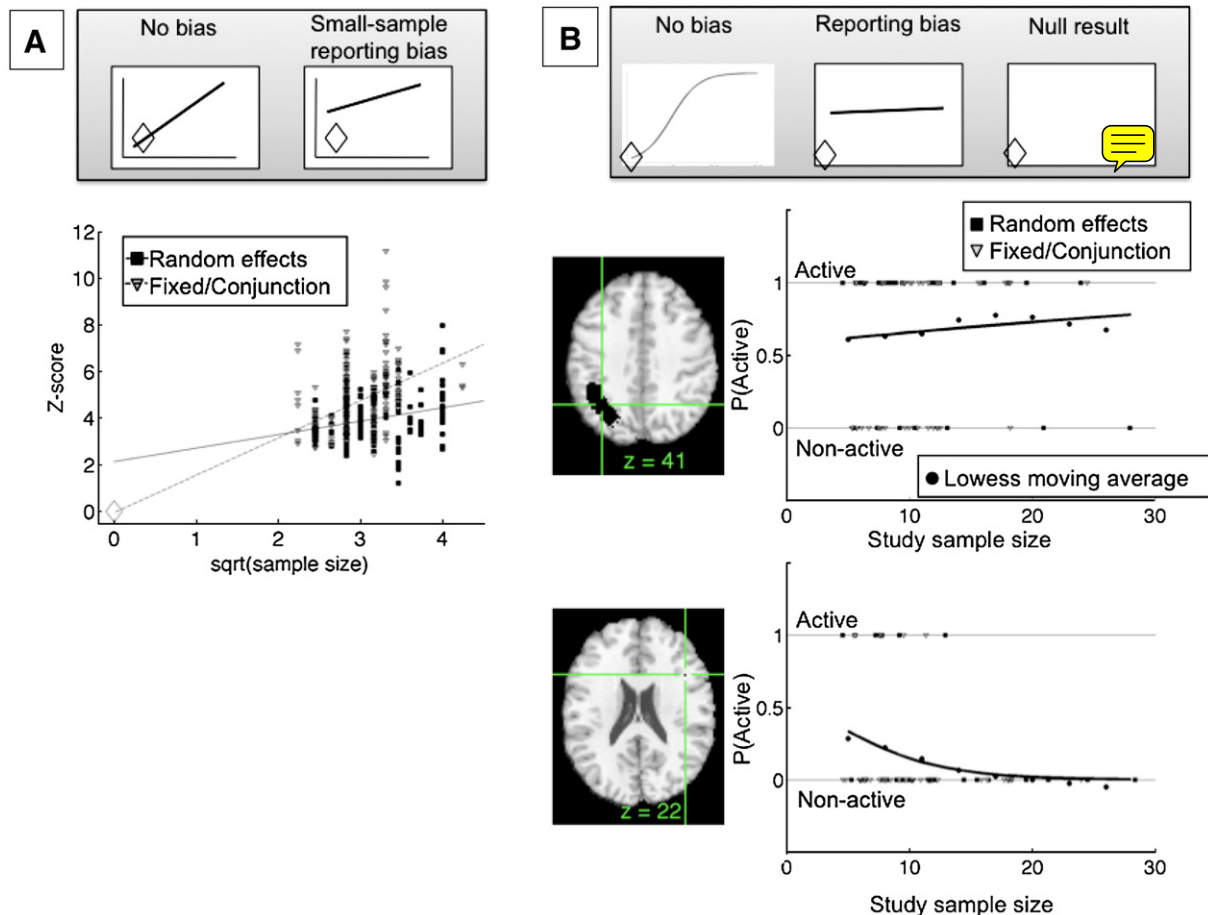


Fig. 3. Adapted Galbraith plots illustrating application to meta-analysis. (A) Plot of Z-scores from available peaks from the executive working-memory (WM) vs. WM storage comparison of a published meta-analysis (Wager and Smith, 2003). Z-scores within significant regions in the multilevel kernel density analysis (MKDA; y-axis) are plotted against the square root of sample size (x-axis). In the absence of bias, the regression line should pass through the intercept (unfilled diamond). This condition holds for fixed-effects studies (light gray triangles), but not for random-effects studies (dark gray squares), indicating small-sample bias in the random-effects studies. See text for additional details. (B) Adapted Galbraith-style graph plotting activations for each study contrast map (SCM; y-axis) as a function of sample size (x-axis) within regions of interest from the MKDA analysis. Individual SCMs are plotted as points (1 = active, 0 = not active), and the solid regression line shows logistic regression fits for the proportion of activated SCMs ($P(\text{Active})$, y-axis) as a function of sample size. The gray circles show estimates of $P(\text{Active})$ using loess smoothing ($\lambda = .75$) and can be used to assess the quality of logistic regression fits. In the absence of bias, the logistic fit should pass through the intercept (see text). The upper plot shows results from a parietal region indicating some small-sample bias. The lower plot shows a small white-matter region in the frontal cortex. Activation was significantly consistent in the MKDA analysis, but the plot shows that it was driven entirely by the small-sample studies, suggesting a lack of true responses to executive WM in this region.

539 An example is shown in Fig. 4. Panel A shows an adapted Galbraith
 540 plot; \sqrt{N} is plotted on the x-axis for studies of executive working
 541 memory, as the full standard error is not generally available from
 542 published neuroimaging papers. The slope will thus be different from
 543 the standard Galbraith plot, but the expected intercept is still zero in
 544 the absence of bias. Z-scores from the subset of available studies
 545 within the significant regions in the MKDA analysis for [Executive
 546 WM-Storage] (Fig. 1) are shown. As Fig. 4A shows, Z-scores for fixed-
 547 effects studies (light-colored triangles) pass through the intercept
 548 (unfilled diamond), but those from random-effects studies (dark
 549 squares) do not. Thus, there is evidence for bias in the random-effects
 550 studies. One plausible type of bias is the well-known “file drawer”
 551 problem. Smaller studies that did not find effects in these regions may
 552 be unpublished, and thus Z-scores from published studies with small
 553 sample sizes would be inflated relative to the true effect size across all
 554 studies. This problem is exacerbated because small-sample studies
 555 have very little power in a random-effects framework, and thus those
 556 that end up being published are those that happen to have particularly
 557 large Z-scores (either by chance or because of some real difference in
 558 effect magnitude). Fixed-effects studies do not show apparent bias,
 559 perhaps because these studies tend to be older and were publishable

even with relatively low effect sizes. In addition, fixed-effects analysis
 560 is substantially more liberal than random-effects analysis, resulting in
 561 higher Z-scores overall, and thus studies using fixed-effects analyses
 562 are more likely to yield Z-scores high enough to meet publication
 563 standards even with small samples. One issue with these plots is that
 564 Z-score values are not independent from one another, and thus the
 565 statistical significance of the Galbraith plot regression is difficult to
 566 interpret. 567

Fig. 4B shows an analogous plot, but shows the probability of a
 568 nominally independent SCM activating (y-axis) vs. N (x-axis).
 569 Individual studies are plotted with y-axis values of either 1 (active)
 570 or 0 (non-active), and logistic regression is used to create a best-fit
 571 prediction (solid black lines) of the probability of activation ($P(\text{Active})$)
 572 as a function of N . The gray circles show smoothed averages of P
 573 (active) vs. N estimated using loess smoothing, and can be used to
 574 assess the logistic regression model fit. As with the standard Galbraith
 575 plot, if a region is truly activated by the task (executive WM in Fig. 4)
 576 and there is no bias, $P(\text{Active})$ should increase with increasing sample
 577 size and should pass through the origin ($P(\text{Active}) = 0$ when $N = 0$). Bias
 578 in small-sample studies is indicated by a non-zero intercept. Finally, a
 579 negative regression slope would indicate that an effect is driven 580

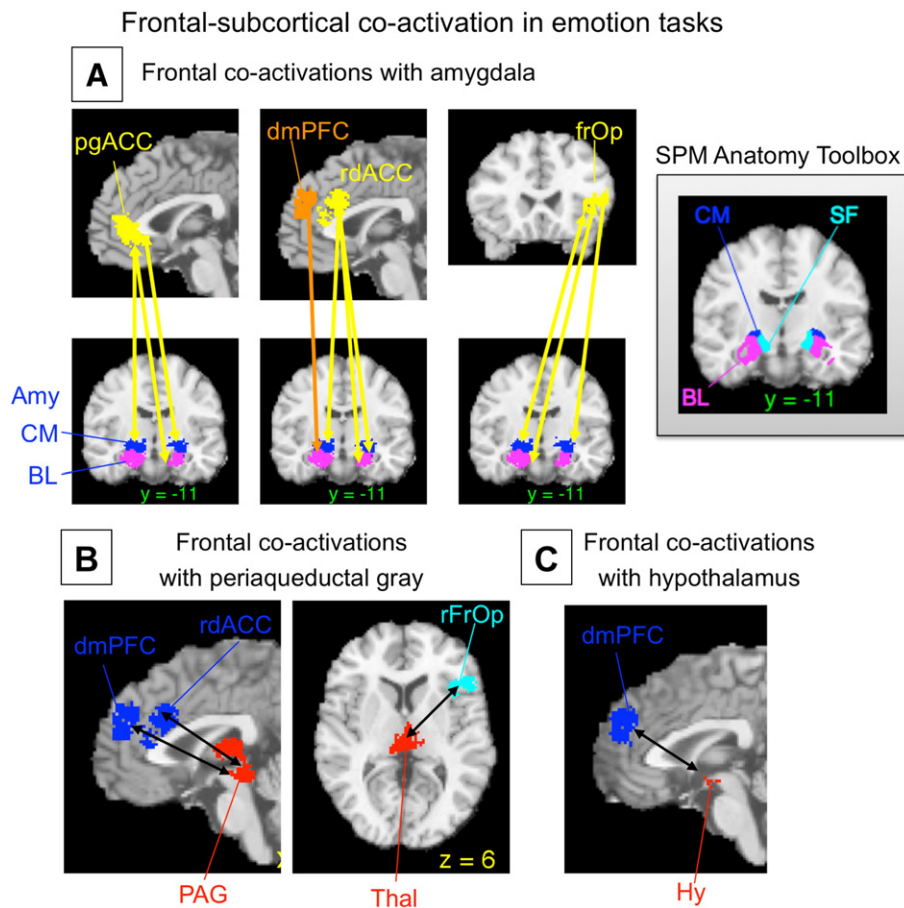


Fig. 4. Example of co-activation analyses from a recent meta-analysis of emotion Adapted from Kober et al. (2008), Figs. 8–9). Co-activated regions show a significant tendency to be activated in the same study contrast maps (SCMs), as assessed with Kendall's tau- b . Arrows show significant co-activation. (A) Frontal regions (yellow/orange) co-activated with amygdala subregions (blue/purple) are a surprisingly circumscribed set of regions limited to the medial prefrontal cortex (mPFC) and the right ventrolateral PFC/frontal operculum. The inset shows regions from the SPM Anatomy Toolbox (V15; (Eickhoff, Heim, Zilles, and Amunts, 2006; Eickhoff et al., 2005). Amy, amygdala; BL, basolateral complex; CM, centromedial complex; dmPFC, dorsomedial prefrontal cortex; pgACC, pregenual anterior cingulate; rdACC, rostral dorsal anterior cingulate; rFrOp, right frontal operculum; SF, superficial amygdala. (B) Frontal regions co-activated with midbrain periaqueductal gray (red, shown including a contiguous region in the thalamus) include a subset of the same frontal regions. (C) The only frontal region co-activated with hypothalamus (red) was the dmPFC. These results suggest locations for functional frontal-limbic and frontal-brainstem pathways related to emotional experience that can be tested in future neuroimaging and lesion studies.

581 predominantly by the small studies, and that $P(\text{active})$ converges on
 582 zero as N increases; thus, it is evidence that a region does not respond
 583 to the task studied. Plots are shown for two regions of contiguous
 584 voxels that were significant in the MKDA analysis shown in Fig. 1. The
 585 first, a region in the left parietal cortex commonly activated in
 586 executive WM tasks, which shows evidence for both a positive slope
 587 and a non-zero intercept, indicating a true effect and a tendency to
 588 over-report by small studies. This bias could be related to the use of
 589 lowered thresholds, or other factors discussed above. The second
 590 region, a small region in white matter in the lateral frontal cortex,
 591 shows evidence for a null result: The consistent activation is produced
 592 exclusively by the small-sample studies, resulting in a negative
 593 regression slope. (We are not arguing here against a role of lateral
 594 prefrontal cortex in executive WM: Other lateral prefrontal regions
 595 showed more well-behaved data). These results illustrate the
 596 usefulness of meta-analytic plots, above and beyond localizing
 597 significant regions using MKDA or a similar analysis.

598 Section II: Analyzing activation specificity

599 Meta-analysis is perhaps the only way to compare neuroimaging
 600 results across a wide variety of tasks, as shown in the example in Fig.
 601 1B (Van Snellenberg and Wager, in press #71). This unique advantage
 602 can be captured quantitatively in analyses that examine the specificity
 603 of regional activation to particular task types.

The most basic type of between-task comparison is between two 604
 conditions (e.g., positive and negative emotion, or executive WM vs. 605
 simple storage). An early approach counted the number of peaks or 606
 studies activating within a pre-specified anatomical area, and used a 607
 χ^2 test to determine whether the proportions of peaks inside vs. 608
 outside the area differed by task type (Phan et al., 2002; Wager et al., 609
 2003; Wager and Smith, 2003). This analysis controls for the marginal 610
 counts of overall peaks within the area and overall frequency of peaks 611
 for each task type, and is valid for comparisons of two or more task 612
 types. However, it has several drawbacks. First, anatomical boundaries 613
 currently cannot be precisely specified. Second, counting peaks suffers 614
 from the same fixed-effects issues discussed above, thereby limiting 615
 generalizability, but study counts are often too low to perform a valid 616
 χ^2 test on studies or SCMs. This is because the χ^2 test is a large-sample 617
 approximation and is not valid if expected counts in any cell in the 618
 contingency table fall below 5 or so. Therefore, large numbers of 619
 studies and large regions are needed. 620

In addition, it is important to note one other consideration. The 621
 Phan et al. χ^2 test provides estimates of relative activation frequency: 622
 if one area is very dense with peak/study activations, it will influence 623
 the overall marginal frequencies of peaks used in tests in every other 624
 region. We return to this issue in more detail below. 625

In recent work, we have employed an alternative to the χ^2 test, a 626
 multinomial permutation test (MPT), which addresses some of these 627
 issues. The MPT is very similar in principle to the χ^2 test, and in fact 628

629 uses the χ^2 statistic as a summary statistic; however, it is a
 630 permutation-based procedure that approximates the multinomial
 631 exact test (Agresti, 2002). Like the χ^2 test, it can be used to make
 632 whole-brain maps of areas showing task-type differences in each local
 633 neighborhood around the brain. For the local area around each voxel, a
 634 “yes/no” by task type contingency table is constructed, where “yes”
 635 and “no” refers to whether the SCM activated within r mm of the
 636 voxel. Exact p -values can be obtained for 2×2 tables using Fisher's
 637 exact test or for larger tables using the multinomial exact test (MET),
 638 but both of these are extremely computationally demanding, and the
 639 MET for even a single voxel of a moderately sized meta-analysis (e.g.,
 640 80 maps) is not feasible with current commonly available computing
 641 resources. However, permutation methods can be used to approx-
 642 imate the MET with much lower computational cost. We permute the
 643 “yes/no” activation indicator, providing a sample from the set of null-
 644 hypothesis tables with the same marginal distributions of activation
 645 counts and task-type counts, as suggested in (Agresti, 2002); p. 98).
 646 We use the χ^2 statistic as a convenient summary of asymmetries
 647 between activation and task type, and threshold the distribution of χ^2
 648 statistics from permuted tables at $1-\alpha$. In practice, 5000
 649 permutations at each voxel provides stable results, is computationally
 650 feasible (2–3 days for a whole-brain map with a large sample of >400
 651 SCMs), and is substantially faster than Fisher's exact test for large (i.e.,
 652 80 or more) numbers of SCMs.

653 In this way, the other problematic issues raised above are
 654 addressed as well. To avoid ambiguities with imprecisely defined
 655 ROIs, we perform the test voxel-by-voxel over the whole brain (or a
 656 volume of interest). To avoid the complications related to making
 657 inferences about peaks without considering which study they came
 658 from, SCMs rather than individual peaks are counted and analyzed.
 659 This test is different from the χ^2 test described above in another way
 660 as well: It analyzes only the distribution of activating vs. non-
 661 activating SCMs within a given brain region. Therefore, it provides a
 662 direct test of differences among tasks in the probability of activating a
 663 single region, independent of activation frequencies in other regions.
 664 This test is implemented in the current version of the MKDA software.

665 Comparing two task types using MKDA

666 Another means of comparing two conditions uses voxel-wise
 667 analysis within the ALE/KDA/MKDA framework. In this approach,
 668 separate maps are constructed for each of two task types and
 669 subtracted to yield difference maps. The same procedure is employed
 670 in the course of the Monte Carlo randomization: The locations of
 671 contiguous activation blobs (peaks in ALE/KDA) are randomized,
 672 providing simulated null-hypothesis conditions from which we
 673 establish a threshold for significant differences.

674 Like the Phan et al. χ^2 test, the Monte Carlo ALE/KDA/MKDA
 675 difference maps test the relative frequency of activating in a given
 676 region, compared with the overall frequencies in the rest of the brain.
 677 Thus, a very reliable concentration of peaks in one area for one task
 678 type will shift (increase) the marginal activation frequencies for that
 679 task, which will affect the null-hypothesis difference in the Monte
 680 Carlo simulations. Thus, for task types with relatively few peaks, there
 681 need not be a greater absolute probability of activating a region to
 682 achieve a significant density for that region relative to other task
 683 types. Consider the following example: Studies of positive and
 684 negative emotion activate the ventral medial prefrontal cortex
 685 (vmPFC) with about equal frequencies. The MPT test would reveal
 686 no differences. However, negative emotions more reliably activate the
 687 amygdala and many other regions (Wager et al., 2008), resulting in a
 688 greater frequency of activation across the brain. With enough studies,
 689 either the Phan et al. χ^2 test or density-difference analyses will
 690 produce a significant positive >negative effect in vmPFC, even though
 691 the absolute proportion of activating studies is roughly equal for
 692 positive and negative emotion. This is not necessarily a flaw, as a

relative concentration of activity in a condition that produces few
 activations in general can convey meaningful information. For
 example, vmPFC activity may indeed be diagnostic of positive
 emotion. However, it is important to keep these issues in mind
 when interpreting results from these analyses.

698 Section III: Testing connectivity

699 Meta-analysis can also be used to reveal patterns of co-activated
 700 regions. If two regions are co-activated, studies that activate one
 701 region are more likely to activate the other region as well. Co-
 702 activation is thus a meta-analytic analogue to functional connectivity
 703 analyses in individual neuroimaging studies, and can provide
 704 converging evidence on functionally connected regions and hypoth-
 705 eses that can be tested in subsequent studies.

706 As with summaries of consistency, a natural level of analysis is the
 707 SCM (Etkin and Wager, 2007; Kober et al., 2008). In the MKDA-based
 708 approach, the data is an $n \times v$ indicator matrix of which of the n SCMs
 709 activated in the neighborhood of each of the v voxels in the brain. The
 710 resulting connectivity profiles across voxels can be simplified into
 711 connectivity among a smaller set of structurally or functionally
 712 defined regions (groups of voxels). Hypothesis tests can be performed
 713 on connectivity, and relationships among multiple regions can be
 714 summarized and visualized.

715 There are several potential measures of association for bivariate,
 716 binomial data, including Kruskal's Gamma, Kendall's Tau, Fisher's
 717 exact test, and other recent measures of association for binomial data
 718 developed within the neuroimaging literature (Neumann et al., 2005;
 719 Patel et al., 2006; Postuma and Dagher, 2006). We have used Kendall's
 720 Tau- b (τ) because it is appropriate for binomial data and has a clearly
 721 interpretable metric (Gibbons, 1993; Gibbons et al., 2003).

722 Co-activation measures can be used for a number of purposes.
 723 First, they can be used to test for specific relationships among brain
 724 areas of interest. For example, we used a database of 437 SCMs from
 725 emotion tasks to test which frontal regions were co-activated with the
 726 amygdala, periaqueductal gray (PAG), and hypothalamus, three key
 727 subcortical nuclear complexes involved in emotion (Kober et al.,
 728 2008). Only four specific frontal areas showed positive co-activation
 729 with these areas (see Fig. 4). They included several specific regions in
 730 the medial prefrontal cortex (mPFC)—including pregenual anterior
 731 cingulate, rostral dorsal cingulate, and dorsomedial prefrontal cortex—
 732 and one area in the right frontal operculum. These results reveal a
 733 relatively specific pattern of frontal connectivity with these important
 734 subcortical regions. They correspond well with animal studies
 735 showing direct projections to amygdala and PAG mainly from the
 736 MPFC (An et al., 1998; McDonald et al., 1996). In addition, the
 737 homologies between rat or primate and human mPFC are not
 738 currently well understood, and this kind of information in humans
 739 helps to establish homologous regions.

740 Another use for co-activation measures is in functional parcellation
 741 of the brain (Flandin et al., 2002; Thirion et al., 2006), or the
 742 establishment of groups of contiguous voxels that show similar
 743 functional characteristics and may be treated as units of analysis in
 744 future studies. In the Kober et al. study, it would have been
 745 computationally unwieldy to examine co-activation between thou-
 746 sands of voxels in the frontal cortex and thousands of voxels in
 747 subcortical regions of interest. Instead, we calculated co-activation
 748 among parcels: We first used singular value decomposition on a 437
 749 $(\text{SCMs}) \times 18,489$ (voxels) matrix of significant voxels from the MKDA
 750 analysis and identified groups of contiguous voxels that loaded most
 751 highly on the same component. These regions were taken as parcels,
 752 and a new SCM indicator for each parcel was constructed, which
 753 indicated whether each SCM activated in the neighborhood of the
 754 parcel. These parcels corresponded well in many cases with the
 755 locations of known anatomical regions. For example, in Fig. 4, sub-
 756 regions of the amygdala derived from parcellation of the meta-

analysis are shown in comparison with those derived from cytoarchitectural analysis of post-mortem brains (Eickhoff et al., 2005).

The parcel indicators were subjected to two iterations of non-metric multidimensional scaling and clustering to identify functional regions and large-scale networks. The details of this procedure are beyond the scope of this brief discussion, but the end result is that parcels of functionally related brain activity, and networks of co-activated regions at several spatial scales, can be identified and used to guide interpretation and *a priori* testing in future studies.

Co-activation measures can also be used to characterize differences among groups of individuals, including those with psychiatric and neurological disorders. For example, Etkin and Wager (2007), compared frontal-amygdala and frontal-insula co-activation in studies of three types of anxiety-related disorders: Post-traumatic stress disorder, social anxiety disorder, and specific phobias. We tested the hypothesis that medial frontal increases would be consistently associated with a lower incidence of amygdala and insula activity across studies. Co-activation analyses supported this view (See Fig. 4), and we found that this co-activation was driven by studies of PTSD specifically. This is one example of how meta-analysis can be used to test the consistency of functional relationships among brain regions, and also compare functional relationships across different functional domains (in this case, anxiety-related disorders).

Section IV: Future directions

There is tremendous potential for development of meta-analytic techniques and applications to advance the cumulative science of brain imaging. One avenue for development involves increasing integration of meta-analysis results with brain atlases and databases (Dickson, Drury, and Van Essen, 2001; Van Essen et al., 2001) so that consensus results will be immediately available to researchers. Another is the aggregation and analysis of full summary statistic images from each study, rather than analysis of the reported peaks. This would allow effect-size based meta-analyses with full information across the brain, and would greatly enhance the value of meta-analytic maps.

Whether full statistic images or reported coordinates are analyzed, there is ample room for the development and application of both new and traditional meta-analysis techniques. Here we have presented an initial use of graphical meta-analysis plots, which could be very useful in detecting and quantifying bias in future meta-analyses. New applications of techniques for parcellating and evaluating co-activation based on data across studies can provide increasingly precise maps of large-scale functional regions, which can in turn inform increasingly precise anatomical hypotheses in new studies.

In addition, other avenues require development: One is how to model SCMs, which are currently treated as independent, but which are often nested within studies, and whose cohorts sometimes share individuals even if they come from different studies. Another is the application of logistic regression techniques appropriate for low-frequency responses, to analyze task specificity while controlling for confounding variables. The tests for specificity described above analyze activation frequencies as a function of a single psychological variable (e.g., spatial vs. verbal vs. object WM). However, such variables may be correlated with other confounding variables: for example, PET vs. fMRI studies, storage and manipulation vs. pure storage in WM, or other factors may be asymmetrically distributed across levels of WM Content Type. This raises the potential for multicollinearity and, in some cases, for Simpson's Paradox to occur. For example, spatial WM may activate more frequently than object WM overall, but the reverse may be true when comparing within categories of PET and fMRI studies. Only a few meta-analyses have used logistic regression to control for confounding variables because coverage of the possible combinations of variables is too sparse. This approach will become more feasible as the number of studies

increases and samples of studies can be collected that are relatively balanced across levels of potentially confounding factors.

Finally, developing meta-analysis based applications of classifier techniques is a particularly important future direction, as meta-analysis affords a unique opportunity to make quantitative brain-psychology inferences across many task domains. This approach can be extended beyond simple classification to testing functional ontologies. Because many different kinds of task labeling schemes can be applied to study contrasts, meta-analysis provides the means to pit alternative psychological categorization schemes against one another and ask which maps most cleanly onto brain activity. This approach may turn out to be a unique and valuable way of establishing links between psychological and biological levels of analysis.

Acknowledgments

This research and the preparation of this manuscript were supported in part by National Science Foundation grant (SES631637) and National Institute of Mental Health grant (R01MH076136) to Tor D. Wager. We would like to thank Lisa Feldman Barrett, Eliza Bliss-Moreau, John Jonides, Kristen Lindquist, Derek Nee, and Edward Smith, for their contributions to the meta-analysis datasets presented here.

References

- Agresti, A., 2002. *Categorical Data Analysis* (2nd ed.). John Wiley and Sons, Hoboken, NJ.
- An, X., Bandler, R., Ongur, D., Price, J.L., 1998. Prefrontal cortical projections to longitudinal columns in the midbrain periaqueductal gray in macaque monkeys. *J. Comp. Neurol.* 401 (4), 455–479.
- Chein, J.M., Fissell, K., Jacobs, S., Fiez, J.A., 2002. Functional heterogeneity within Broca's area during verbal working memory. *Physiol. Behav.* 77 (4–5), 635–639.
- DerSimonian, R., Laird, N., 1986. Meta-analysis in clinical trials. *Control. Clin. Trials* 7 (3), 177–188.
- Dickson, J., Drury, H., Van Essen, D.C., 2001. The surface management system (SuMS) database: a surface-based database to aid cortical surface reconstruction, visualization and analysis. *Philos. Trans. R. Soc. Ser. B* 356, 1277–1292.
- Egger, M., Smith, G.D., Schneider, M., Minder, C., 1997. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 315, 629–634.
- Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes, C., Fink, G.R., Amunts, K., et al., 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage* 25 (4), 1325–1335.
- Eickhoff, S.B., Heim, S., Zilles, K., Amunts, K., 2006. Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. *NeuroImage* 32 (2), 570–582.
- Etkin, A., Wager, T.D., 2007. Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. *Am. J. Psychiatry* 164 (10), 1476–1488.
- Flandin, G., Kherif, F., Pennec, X., Riviere, D., Ayache, N., Poline, J.B., et al., 2002. Parcellation of brain images with anatomical and functional constraints for fMRI data analysis. *Biomedical Imaging, 2002. Proceedings. 2002 IEEE International Symposium on*, 907–910.
- Fox, P.T., Parsons, L.M., Lancaster, J.L., 1998. Beyond the single study: function/location metanalysis in cognitive neuroimaging. *Curr. Opin. Neurobiol.* 8 (2), 178–187.
- Fox, P.T., Huang, A.Y., Parsons, L.M., Xiong, J.H., Rainey, L., Lancaster, J.L., 1999. Functional volumes modeling: scaling for group size in averaged images. *Hum. Brain Mapp.* 8 (2–3), 143–150.
- 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 (3 Pt 1), 223–235.
- Gibbons, J.D., 1993. *Nonparametric Measures of Association*. Sage Publications Inc.
- Gibbons, J.D., Chakraborti, S., Gibbons, J.G.D., 2003. *Nonparametric Statistical Inference*. Marcel Dekker.
- Gilbert, S.J., Spengler, S., Simons, J.S., Steele, J.D., Lawrie, S.M., Frith, C.D., et al., 2006. Functional specialization within rostral prefrontal cortex (area 10): a meta-analysis. *J. Cogn. Neurosci.* 18 (6), 932–948.
- Gottfried, J.A., Zald, D.H., 2005. On the scent of human olfactory orbitofrontal cortex: meta-analysis and comparison to non-human primates. *Brain Res. Brain Res. Rev.* 50 (2), 287–304.
- Hedges, L.V., Vevea, J.L., 1998. Fixed- and random-effects models in meta-analysis. *Psychol. Methods* 3, 486–504.
- Joseph, J.E., 2001. Functional neuroimaging studies of category specificity in object recognition: a critical review and meta-analysis. *Cogn. Affect. Behav. Neurosci.* 1 (2), 119–136.
- Kober, H., Barrett, L.F., Joseph, J., Bliss-Moreau, E., Lindquist, K., Wager, T.D., 2008. Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *NeuroImage* 42, 998–1031.
- Kosslyn, S.M., Thompson, W.L., 2003. When is early visual cortex activated during visual mental imagery? *Psychol. Bull.* 129 (5), 723–746.

- Laird, A.R., Fox, P.M., Price, C.J., Glahn, D.C., Uecker, A.M., Lancaster, J.L., et al., 2005. ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts. *Hum. Brain Mapp.* 25 (1), 155–164.
- Lewis, J.W., 2006. Cortical networks related to human use of tools. *Neuroscientist* 12 (3), 211–231.
- McDonald, A.J., Mascagni, F., Guo, L., 1996. Projections of the medial and lateral prefrontal cortices to the amygdala: a *Phaseolus vulgaris* leucoagglutinin study in the rat. *Neuroscience* 71 (1), 55–75.
- 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), 1–17.
- Neumann, J., Lohmann, G., Derrfuss, J., von Cramon, D.Y., 2005. Meta-analysis of functional imaging data using replicator dynamics. *Hum. Brain Mapp.* 25 (1), 165–173.
- Nichols, T., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat. Methods Med. Res.* 12 (5), 419–446.
- Nickel, J., Seitz, R.J., 2005. Functional clusters in the human parietal cortex as revealed by an observer-independent meta-analysis of functional activation studies. *Anat. Embryol. (Berl)* 210 (5–6), 463–472.
- Nielsen, F.A., Hansen, L.K., Balslev, D., 2004. Mining for associations between text and brain activation in a functional neuroimaging database. *Neuroinformatics* 2 (4), 369–380.
- Nielsen, F.A., Copenhagen, D., Lyngby, D., 2005. Mass meta-analysis in Talairach space. *Adv. Neural Inf. Process. Syst.* 17, 985–992.
- Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., Panksepp, J., 2006. Self-referential processing in our brain—a meta-analysis of imaging studies on the self. *NeuroImage* 31 (1), 440–457.
- Patel, R.S., Bowman, F.D., Rilling, J.K., 2006. A Bayesian approach to determining connectivity of the human brain. *Hum. Brain Mapp.* 27 (3), 267–276.
- Phan, K.L., Wager, T., Taylor, S.F., Liberzon, I., 2002. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage* 16 (2), 331–348.
- Poldrack, R.A., 2006. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* 10 (2), 59–63.
- Postuma, R.B., Dagher, A., 2006. Basal ganglia functional connectivity based on a meta-analysis of 126 positron emission tomography and functional magnetic resonance imaging publications. *Cereb. Cortex* 16 (10), 1508–1521.
- Rosenthal, R., DiMatteo, M.R., 2001. Meta-analysis: recent developments in quantitative methods for literature reviews. *Annu. Rev. Psychol.* 52, 59–82.
- Sarter, M., Berntson, G.G., Cacioppo, J.T., 1996. Brain imaging and cognitive neuroscience. Toward strong inference in attributing function to structure. *Am. Psychol.* 51 (1), 13–21.
- Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.B., 2006. Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets. *Hum. Brain Mapp.* 27 (8), 678–693.
- 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 (3), 765–780.
- Van Essen, D.C., 2005. A population-average, landmark- and surface-based (PALS) atlas of human cerebral cortex. *NeuroImage* 28 (3), 635–662.
- Van Essen, D.C., Drury, H.A., Dickson, J., Harwell, J., Hanlon, D., Anderson, C.H., 2001. An integrated software suite for surface-based analyses of cerebral cortex. *J. Am. Med. Assoc.* 8 (5), 443–459.
- Van Snellenberg, J.X., and Wager, T.D. (in press). Cognitive and motivational functions of the human prefrontal cortex. In E. Goldberg and D. Bougakov (Eds.), *A tribute to Alexander Luria*.
- Van Snellenberg, J.X., Torres, I.J., Thornton, A.E., 2006. Functional neuroimaging of working memory in schizophrenia: task performance as a moderating variable. *Neuropsychology* 20 (5), 497–510.
- Wager, T.D., Smith, E.E., 2003. Neuroimaging studies of working memory: a meta-analysis. *Cogn. Affect. Behav. Neurosci.* 3 (4), 255–274.
- 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 (3), 513–531.
- Wager, T.D., Reading, S., Jonides, J., 2004. Neuroimaging studies of shifting attention: a meta-analysis. *NeuroImage* 22 (4), 1679–1693.
- Wager, T.D., Hernandez, L., Jonides, J., Lindquist, M., 2007a. Elements of functional neuroimaging. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (Eds.), *Handbook of Psychophysiology*, 4th ed. Cambridge University Press, Cambridge, pp. 19–55.
- Wager, T.D., Lindquist, M., Kaplan, L., 2007b. Meta-analysis of functional neuroimaging data: current and future directions. *Soc. Cogn. Affect. Neurosci.* 2 (2), 150–158.
- Wager, T.D., Barrett, L.F., Bliss-Moreau, E., Lindquist, K., Duncan, S., Kober, H., et al., 2008. The neuroimaging of emotion. In: Lewis, M., Haviland-Jones, J.M., Barrett, L.F. (Eds.), *Handbook of Emotions*, 3rd ed. Guilford Press, New York, pp. 249–271.
- Wager, T.D., Lindquist, M., and Hernandez, L. (in press). Essentials of functional neuroimaging. In J. Cacioppo and G. G. Berntson (Eds.), *Handbook of Neuroscience for the Behavioral Sciences*.