NeuroImage 45 (2009) 810–823

Contents lists available at ScienceDirect

NeuroImage

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

Meta-analysis of neuroimaging data: A comparison of image-based and coordinate-based pooling of studies

Gholamreza Salimi-Khorshidi [a,*], Stephen M. Smith [a], John R. Keltner [a], Tor D. Wager [b], Thomas E. Nichols [c]

[a] Centre for Functional MRI of the Brain (FMRIB), University of Oxford, Oxford, UK
[b] Department of Psychology, Columbia University, New York, USA
[c] GlaxoSmithKline Clinical Imaging Centre, London, UK

ARTICLE INFO

Article history:
Received 26 June 2008
Revised 28 November 2008
Accepted 13 December 2008
Available online 31 December 2008

ABSTRACT

With the rapid growth of neuroimaging research and accumulation of neuroinformatic databases the synthesis of consensus findings using meta-analysis is becoming increasingly important. Meta-analyses pool data across many studies to identify reliable experimental effects and characterize the degree of agreement across studies. Coordinate-based meta-analysis (CBMA) methods are the standard approach, where each study entered into the meta-analysis has been summarized using only the (x, y, z) locations of peak activations (with or without activation magnitude) reported in published reports. Image-based meta-analysis (IBMA) methods use the full statistic images, and allow the use of hierarchical mixed effects models that account for differing intra-study variance and modeling of random inter-study variation. The purpose of this work is to compare image-based and coordinate-based meta-analysis methods applied to the same dataset, a group of 15 fMRI studies of pain, and to quantify the information lost by working only with the coordinates of peak activations instead of the full statistic images. We apply a 3-level IBMA mixed model for a "mega-analysis", and highlight important considerations in the specification of each model and contrast. We compare the IBMA result to three CBMA methods: ALE (activation likelihood estimation), KDA (kernel density analysis) and MKDA (multi-level kernel density analysis), for various CBMA smoothing parameters. For the datasets considered, we find that ALE at $\sigma = 15$ mm, KDA at $\rho = 25$–30 mm and MKDA at $\rho = 15$ mm give the greatest similarity to the IBMA result, and that ALE was the most similar for this particular dataset, though only with a Dice similarity coefficient of 0.45 (Dice measure ranges from 0 to 1). Based on this poor similarity, and the greater modeling flexibility afforded by hierarchical mixed models, we suggest that IBMA is preferred over CBMA. To make IBMA analyses practical, however, the neuroimaging field needs to develop an effective mechanism for sharing image data, including whole-brain images of both effect estimates and their standard errors.

Introduction

The number of neuroimaging studies is growing dramatically, with the fMRI literature having grown from 2 publications in 1993 to 1970 publications in 2007, and exponential growth since 2000 predicting a doubling of the yearly publication rate every 3.64 years[1]. However many of these publications contain conflicting results, or are based on only a small number of subjects. Hence there has been increasing interest in using meta-analysis methods to find consistent results for a specific functional task. These same methods can also be used to predict the results of a study that has not been performed directly, by combining studies that intersect on a particular concept.

Many neuroimaging studies are under-powered, with the typical number of subjects ranging from 10 to 20 subjects. The main challenge in performing statistical inference over such small sample sizes is the limited power and, related, the risk that results will not be reproduced in another group of subjects. For example, Thirion et al. (2007) investigate the reproducibility of statistical inference over different numbers of randomly-selected subjects from a pool of 80 subjects performing the same task. They show that the number of subjects needed to give a generalizable result is greater than 20. This suggests that studies in the literature based on small samples are difficult to interpret in isolation and researchers could greatly benefit from pooling evidence from multiple studies.

A statistical meta-analysis combines the results of several studies that address a set of related research hypotheses, thus increasing power and reliability (Sutton et al., 2000). While authors of meta-analyses rarely have the complete original datasets, when they are available, it is natural to perform an image-based meta-analysis (IBMA) which combines whole-brain statistic volumes, rather than just using a summary of them (i.e., a list of local maxima coordinates). Lazar et al. (2002) review a number of ways to combine different

* Corresponding author.
E-mail address: reza@fmrib.ox.ac.uk (G. Salimi-Khorshidi).

[1] Based on a per-year PubMed search of "fMRI" in title or abstract.

1053-8119/$ – see front matter © 2008 Published by Elsevier Inc.
doi:10.1016/j.neuroimage.2008.12.039

subjects' statistic maps, and such methods can equally be applied to combining different studies' maps. In particular, Fisher's p-value combining method and Stouffer's average Z method ($\sqrt{n}\bar{Z}$) have frequently been used in traditional meta-analyses. These methods are fixed effects (FFX) methods, however, and their output does not reflect the consistency of the studies considered.

To account for both within- and between-study variance, a hierarchical mixed effects (MFX) model is a natural approach. In fMRI a generic hierarchical modeling framework is often used where, instead of modeling *all* of the data at all levels simultaneously, summary statistics are passed between levels of the hierarchy (Beckmann et al., 2003; Worsley et al., 2002). While this work has generally been used to combine first-level intra-subject fMRI model results into a second-level group fMRI model, it can be equally well used to combine multiple second-level studies into a third-level meta-analysis. An essential component of the work is that both effect size (contrast of parameter estimate, or COPE) images and their variance (variance of the contrast of parameter estimate, or VARCOPE) are passed up from one level to the next, allowing subjects with poor precision to be down-weighted relative to high precision subjects, and provide MFX inferences that incorporate both within- and between-subject variation. At the third level, such a mega-analysis translates to a method that can down-weight studies with poor precision, and inferences that account for both within- and between-study variation. At the third level, between-study random variation may not be of interest and so FFX may be used instead. For example, if one only wants to obtain the most sensitive pooling of a group of studies, an FFX inference at the 3rd (study) level would be appropriate. If, on the other hand, one wants to find the areas found most reliably in many studies, then a 3rd level MFX inference would be desired.

In common practice, neuroimaging studies rarely provide the full image data, and instead only activation foci magnitude and location are reported in journal papers, or submitted to results databases such as BrainMap[2] (Laird et al., 2005a). Hence most meta-analysis methods are based only on activation foci in a standard space (e.g. MNI152) and we called this the coordinate-based meta-analysis (CBMA) approach. There are several limitations to CBMA methods, one being the information loss due to the relative sparseness of such a representation of the image results, and another being that coordinates are very sensitive to methods adopted in the study, from thresholding to report preparation (e.g., how many foci per cluster are reported) (Wager et al., 2007). For example, from one single dataset, three different sets of foci could be obtained depending on whether just three local maxima per cluster are reported (the default in SPM) above a corrected threshold, all local maxima are reported above a corrected threshold, or all maxima above an uncorrected threshold are reported. Since there is no universal standard for reporting results, CBMA methods should ideally take account of these differences but rarely do.

CBMA approaches were pioneered by Fox et al.'s functional volumes modeling (FVM) method (Fox et al., 1997), though it lacked a formal statistical framework (Fox et al., 1998). The FVM method assumed a Gaussian spatial distribution of activations (Fox et al., 1999), though subsequent authors relaxed this assumption using non-parametric modeling of the distribution of foci (Nielsen and Hansen, 2002).

Currently, there are three widely used CBMA methods: ALE, KDA and MKDA. ALE, or activation likelihood estimation (Turkeltaub et al., 2002), is implemented in software provided by the BrainMap database. In brief, ALE constructs 'likelihood' maps for each activation focus by placing a 3D Gaussian density with specified FWHM at the focus location; these maps are then combined with the addition rule for probabilities, giving the probability that one or more foci are near a given voxel. KDA, or kernel density analysis (Wager et al., 2004) also treats each focus independently, but instead uses a spherical kernel and a simple addition rule to produce a map showing the number of foci within a given radius. MKDA, or multi-level KDA (Wager et al., 2007), does not treat each focus independently, and instead creates a binary map for each study, showing where there is one or more foci within a given radius; these study binary maps are then averaged, giving the proportion of studies having any foci within a given radius from a voxel. Unlike ALE and KDA, MKDA does not treat all foci equally and uses studies as the units of analysis, and thus minimizes the potential for one study with many foci to drive a meta-analytic result.

By definition, the CBMA methods retain less information from each individual study than IBMA methods. However, it is an open question as to *how much* information is lost, and whether the CBMA methods can capture similar patterns of activations that IBMA methods provide. Further, the CBMA methods have spatial tuning parameters (Gaussian FWHM for ALE, and sphere radius for KDA and MKDA) which have no objectively-defined optimal setting. Hence the purpose of this work is to compare CBMA results to IBMA results for a variety of CBMA tuning parameter settings, to understand the relative sensitivity of each method and how performance depends on CBMA parameters.

Materials and methods

In this section we first describe IBMA methods, reviewing the hierarchical MFX model, itemising practical issues and discussing when a MFX vs. FFX model is appropriate. We then describe and compare the three considered CBMA methods. After introducing the collection of pain datasets used, we present the evaluation methods used to compare the different IBMA and CBMA methods.

IBMA analyses

Several preparations must be made before any IBMA. First, all image data or relevant summary images must be warped into a common atlas space. While this is a fundamental pre-processing step, it is important that the atlas is the same for all subjects and all studies, and that the warping methods are as similar as possible between studies. All images should use the same smoothing-kernel size, though if subjects come from different imaging centers a "Smooth to" strategy can be used (Friedman et al., 2006).

Another fundamental issue with IBMA methods is masking. Most standard analysis software will only analyze a voxel if all subjects (or studies) have data. This means that the analysis mask is the intersection of all the masks contributing to the model, which can result in dramatic erosion of the brain volume analysed. In particular, a single subject with some missing data (e.g. due to motion, or a poor anatomical-functional alignment) can dramatically reduce the final analysis mask.[3] It is important to be aware of such effects and ameliorate these through either careful investigation of each session's data, generous intrasubject mask definition, or the use of statistical methods that allow for missing data.

IBMA methods that are only based on Z or T statistic images are unitless, while methods that use effect magnitude images require great care to ensure compatible units between the design matrices and contrasts in each study. For example, if one study has BOLD regressors with a baseline-to-peak height of 1 and a second study has BOLD regressors with a baseline-to-peak height of 2, then first study will have parameter estimate units twice that of the second study. Similar issues arise with respect to compatibility of contrasts, especially those expressing differences between conditions or groups.

[2] http://www.brainmap.org.

[3] Note that the FSL 4.0 software introduced a new masking approach that resulted in a smaller mask than in previous versions, sometimes resulting in severely contracted group masks. FSL 4.1 uses a more generous masking scheme and is recommended for IMBA or any study using a large number of subjects.

The best strategy is to ensure that all intrasubject model predictors have the same scaling, and that all contrasts preserve that scale. To ensure that a contrast preserves the units of the predictors, it is usually sufficient to require that all positive contrast elements sum to 1.0 and all negative elements (if any) sum to −1.0.

There are a variety of possible IBMA analysis methods (described next), but each can be classified as providing either FFX or MFX inferences. A FFX meta-analysis measures evidence for a non-zero effect relative to the inter-subject variability pooled over studies, while a MFX meta-analysis measures an effect relative to the combination of inter-subject and inter-study variability. As meta-analysis is often used to increase power with less concern given to inter-study consistency, a FFX may well be the most appropriate type of inference, whereas a MFX inference should only be required if a strong statement about inter-study consistency is needed.

Combining methods

One approach to IBMA is the generic combining of statistic images, one per study. For a thorough review of this approach see Lazar et al. (2002), which discusses many of the well-known methods in the meta-analysis literature. In this work we consider only Fisher's *p*-value combining method ($-2\times\sum_i \ln P_i$, where P_i is the uncorrected *p*-value of the *i*th study) and Stouffer's *Z*-transform test ($\sum_i Z_i/\sqrt{n}$, where Z_i is the *z*-score for the *i*th study). These two methods are FFX methods, which provide evidence of one or more studies possessing an effect. One limitation of Fisher's method is that it can give significant results even when the signs of one-sided tests input to it are highly discordant, while conflicting signs will cancel with the Stouffer's method.

Single-level regression

The simplest model for the effect magnitude data is a single regression model for all data. If all first-level time series data are modeled at once, the resulting (giant) regression model would yield FFX inferences. This requires massive computing resources to simultaneously access gigabytes of data, therefore it is not really practical and thus we do not consider it further.

A regression of the study-level data, in contrast, is very practical. It consists of an ordinary least square (OLS) regression, a simple unweighted analysis of mean effect magnitude data (one per study), as is typically done in SPM and as is available in many other packages such as FSL and AFNI. This produces MFX inferences, but does not weight studies according to their sample size or standard errors. Hence we prefer a multi-level model which weights each study according to study-level precision and which can produce either FFX or MFX inferences.

(A reviewer proposed an OLS analysis of study-level MFX *z*-score data, instead of effect magnitude data. While this provides a kind of MFX inference, as between study variance is considered and the *z*-scores themselves convey group-level significance, the fitted model is difficult to interpret as it is modeling average *significance* rather than average effect magnitude. However, we include in our results as "Stouffer's-MFX" for completeness.)

Hierarchical model for fixed- or mixed-effects inferences

A multi-level hierarchical model (Beckmann et al., 2003; Woolrich et al., 2004; Worsley et al., 2002) fits data of any kind that is grouped within levels, for example time-series data within subjects, or subject data within studies. We first describe it in terms of combining subjects for a single group analysis[4]. First, each subject is modeled individually, producing effect estimates and standard errors. Next these intrasubject effects and standard errors are modeled together, producing group-level effect estimates and separate within- and between-subject variance estimates. For a MFX inference (FLAME-MFX), each subject is individually weighed according to the balance of their within-subject and the between-subject variance, producing an optimal estimate of the population effect. For a FFX inference (FLAME-FFX), the between-subject variance is ignored, but subjects are still individually weighed (unlike OLS) using just the within-subject variance.

Hierarchical model for image-based meta-analysis

We use this same multi-level hierarchical framework to fit a three-level "mega-analysis" model: Level 1 is the intra-subject modeling of each subject's fMRI time series data, level 2 is the inter-subject analysis for each study, and level 3 is the inter-study meta-analysis. For details of the FSL's FLAME method used we refer the reader to the original citations (Beckmann et al., 2003; Woolrich et al., 2004), but in brief: At level 1, temporal autocorrelation is modeled voxel- wise, providing efficient estimates of each subject's effect estimates; at level 2, after alignment into standard space, each subject's effect estimates and standard errors are combined to give a mean group effect size estimate and MFX variance; at level 3, the study-level effect sizes and variances are again jointly modeled to provide either MFX or FFX inference. The 3rd level model will typically be very simple (e.g. a column of ones to estimate the mean effect over studies), but can have any form. For example, a 3rd level model could be used to test for differences between studies or account for study-level covariates.

Note that a potential source of confusion is how, at both levels 2 and 3, either MFX or FFX inferences can be produced. We are not advocating the use of FFX standard errors at the second (study) level. In both single-study and multiple-study analyses, it is crucial that the second-level standard errors incorporate the between-subject variation. Otherwise the final meta-analysis will not reflect population variation in response magnitudes and will have a very limited interpretation. Hence, the only inference choice is whether to use MFX or FFX at the 3rd level.

CBMA analyses

While there have been a wide variety of methods proposed for CBMA (Fox et al., 1997, 1998, 1999; Nielsen and Hansen, 2002; Chien et al., 2002; Neumann et al., 2005), we have limited our evaluations to three: ALE, KDA and MKDA. In all three methods a map of the evidence for activations is created based on a set of foci coordinates. A qualitative 1D example is shown in Fig. 1. All of the methods assess significance using a Monte Carlo resampling approach where, under the null hypothesis of no coherent activation, the foci are randomly distributed across space. At each voxel an uncorrected *p*-value is obtained by counting the number of Monte Carlo realisations that equal or exceed the original value. Familywise error corrected *p*-values can similarly be obtained by counting the number of realisations where the maximal (image-wise) value exceeds the original value.

ALE

For each focus, ALE scores each voxel as a function of its distance from that focus using a Gaussian kernel of size σ (Turkeltaub et al., 2002). After this step, each voxel has a vector of "activation likelihood" probability values whose elements correspond to foci (one probability per foci). These values are assumed to be independent (the occurrence of one focus is assumed to give no information about whether or not the other foci will occur) and then combined with the addition rule for probabilities to yield the final activation likelihood, or ALE statistic value. This statistic indicates the probability of having at least one peak lying in that particular location, based on the Gaussian model for each focus. The procedure is repeated with Monte Carlo realisations of the data (the same number of foci randomly distributed over the brain) building up a null distribution of ALE maps. The significance test

[4] A multi-level hierarchical model is implemented in the FSL software's "FMRIB's local analysis of mixed effects" or FLAME package, http://www.fmrib.ox.ac.uk/fsl.

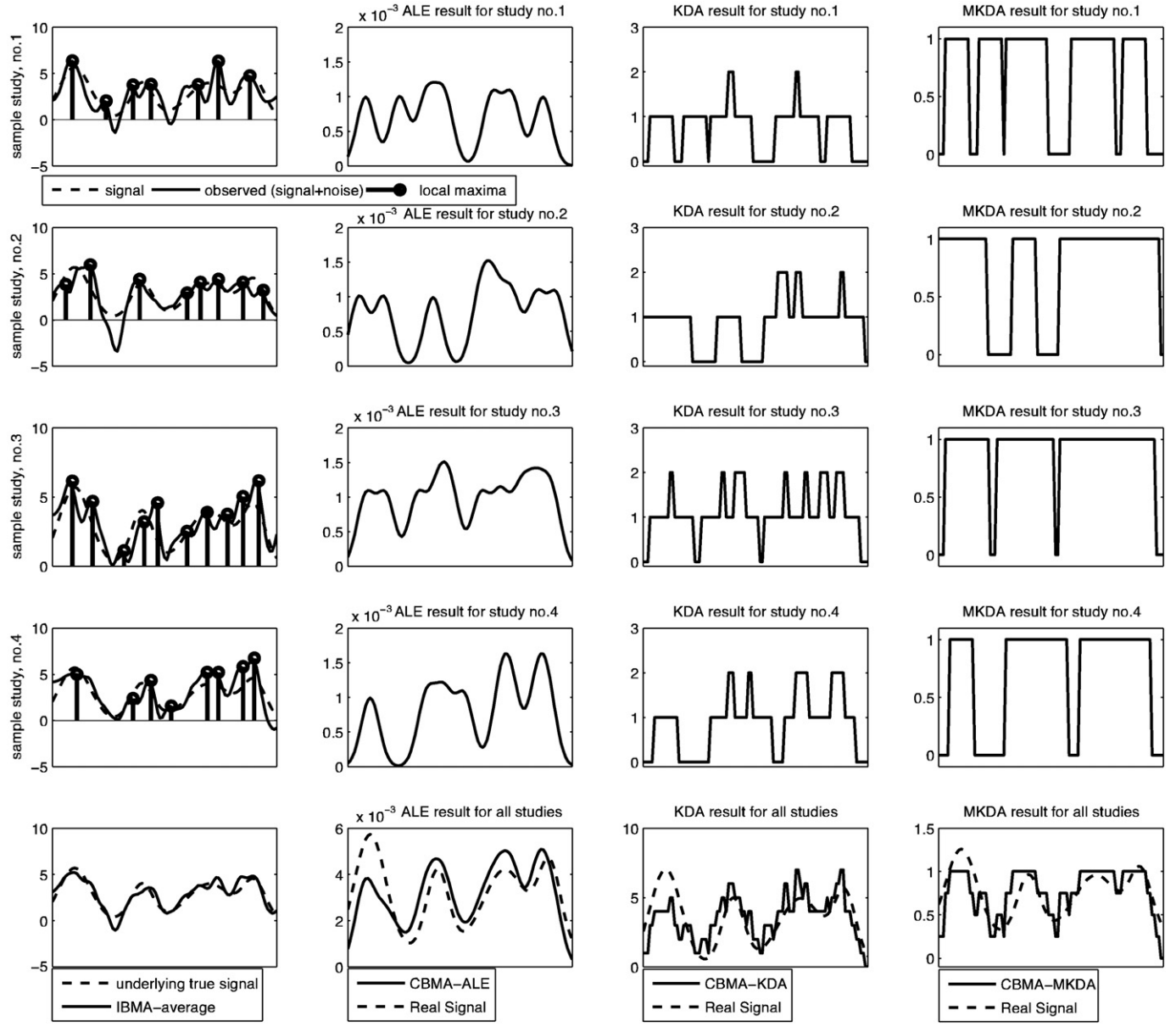

Fig. 1. Illustration of a 4-study, 1-dimensional meta-analysis. A true signal (dashed line) is created and four simulated statistic "images" are created by adding smoothed white noise to the true signal (bold lines in the first column of the first four rows). To apply CBMA to these simulated 1D studies, local maxima (foci) are extracted from each observed signal (circles on the bold lines). Next, the locations of these foci are fed into each CBMA technique. In the last row, the results of each method in reproducing the true signal using the foci are shown. As can be seen, averaging over the complete signals (as IBMA does) yields a better estimate of truth compared to using local maxima (CBMA). ALE results in a smooth estimate (due to its Gaussian kernel) which KDA and MKDA are rougher (due to its spherical kernel). Note that the "true" profile is generated as a sum of Gaussian densities, which is most consistent with the ALE method.

formally tests the global null hypothesis of no coherent activation, but rejecting this null hypothesis voxel-wise should provide evidence of consistent activation at a particular location. Pseudo-code for ALE is shown in Appendix A and a 1D simulated example is shown in the second column of Fig. 1.

KDA

KDA is similar to ALE, but uses a different kernel and method for combining the statistic maps. KDA creates maps for each focus with a spherical indicator function "kernel", with a radius ρ (Wager et al., 2004, 2007). A statistic map is created by summing, producing a map of the number of peaks activated within radius ρ. Similarly to ALE, a Monte Carlo test is used to reject the global null hypothesis of no coherent activation. Pseudo-code for KDA is shown in Appendix B and its 1D simulation is shown in the third column of Fig. 1.

MKDA

A clear limitation of both ALE and KDA is the independent treatment of each focus. If one study has 100 foci and another only 10, the first study will have an immense impact on the results, even if the increased number of foci is only due to different thresholding. The ALE and KDA Monte Carlo procedures also independently scramble each focus, even though a null study would be expected to generate some clustering of foci, due to the smoothness of the image data.

MKDA (Wager et al. 2007) attempts to address these limitations with two modifications to KDA. First, the convolved images are summed by study and truncated at 1.0, creating study-specific images which indicate the presence of one or more foci within radius ρ. These study images are averaged, creating the mean number of studies that have one or more foci near a given voxel. This provides robustness to possible bias from studies that systematically report more foci per

cluster and produces a more interpretable map. Second, the Monte Carlo procedure scrambles foci as clusters, producing realisations that bear greater resemblance to real data (i.e., have clustered foci) but lack any inter-study coherence. Hence MKDA is testing against a more realistic null hypothesis (no study-level coherence) and, since no single study can contribute disproportionately to the result, it is expected to produce more reliable and reproducible activation results. Pseudo-code for MKDA is shown in Appendix C and its 1D simulation is shown in the fourth column of Fig. 1.

Group comparisons with CBMA methods

While the CBMA methods do not have the flexibility of the hierarchical modeling framework described above, it is possible to make simple tests between groups of studies. As presented in (Laird et al., 2005b), if two groups of studies are separately analyzed for creating their corresponding whole-brain statistic maps, subtracting these two maps gives a measure of the difference contrast. Statistical significance of this difference map is assessed with respect to a null distribution of no coherence in either maps, created by taking null maps from each analysis and computing the difference. The final result provides evidence for difference in activation, though this approach has several caveats (detailed in the Discussion section).

Data

The aim was to pool results of 15 pain studies to find regions of activation induced by painful stimuli. In spite of some differences, all studies concentrated on pain as the main effect of interest. In three of these studies, a pain stimulus is combined with some language-related explanatory variables (EVs, or covariates). In two other studies, a painful stimulus is combined with some cues that warn or deceive subjects about an upcoming painful stimulus. Another group of six studies considers the effect of treatment on subjects' pain perception. In the other four studies, a pain stimulus is modulated to obtain different perceived pain levels. All studies have at least one pain EV, which allows us to form a simple "pain" contrast for each subject at the first level (and consequently at second and third levels) (Iannetti et al., 2005; Leknes et al., in preparation(a), in preparation(b), preparation(c); Lee et al., in preparation).

Despite having a pain covariate in all studies, the pain delivery mechanisms are different across the studies. For example, six of the studies used a mechanical pain stimulus, while the other nine studies used a thermal pain stimulus. We investigate a differential response to the two forms of pain delivery in 3rd level (meta) analysis. The result of this analysis will be areas of the brain that show more or less thermal-induced pain activation relative to mechanical-induced pain.

Processing of functional images at the first level was performed using FSL (Smith et al., 2001). Functional images were motion corrected (Jenkinson et al., 2002) and spatially smoothed (full width half maximum=5 mm) prior to temporal model fitting (Beckmann et al., 2003; Woolrich et al., 2004, 2001). Co-registration to the MNI152 standard brain space was performed in 2 stages: (1) the fMRI data from a given subject was registered to that subject's T1 structural using linear registration (Jenkinson and Smith, 2001; Jenkinson et al., 2002) and (2) the subject's structural image was registered to the MNI standard brain using nonlinear registration.

In the second-level analyses (Woolrich et al., 2004) MFX activation maps corresponding to the main pain effect were created. Third-level cross-study analyses were carried out using all studies, with a one-group model or a two-group model split by mechanical vs. thermal stimulus study type. Both fixed (FLAME-FFX) and mixed (FLAME-MFX) activation maps were created at the third level.

We used the results of the 15 pain studies to create the foci lists for the CBMA analyses. For each study, the second-level analysis produced a list of foci, the locations of local maxima in the statistic image. A constraint is imposed to find local maxima that are not closer than 8 mm to each other, which matches the default behavior of SPM's results. Based on this framework, a list of 231 foci is extracted from all 15 available studies. This foci list is the main input to all following CBMA (ALE, KDA and MKDA).

Map comparison

We use the results of the IBMA FLAME-FFX model to define a gold standard result against which the other methods are compared. This choice of standard result follows from a sequence of three assessments: IBMA is preferred over CBMA, as the image data are a strict superset of the information in CBMA analyses; FFX is preferred over MFX, as the typical meta-analysis goal is aggregation of evidence for an effect, not MFX's inference on inter-study concordance; and, for the choice of IBMA analysis method, FLAME's hierarchical model is preferred over other traditional meta-analytic measures, due to its statistical optimality and flexibility for dealing with group differences and covariates.

We compare CBMA maps to the IBMA gold standard with one symmetric and two asymmetric measures. The Dice similarity measure (DSM) (Dice, 1945) is a symmetric measure of the resemblance of two binary images:

$$\mathrm{DSM} = \frac{2|I \cap C|}{|I| + |C|} \tag{1}$$

where $|I|$ and $|C|$ are the number of non-zero voxels in a thresholded IBMA (reference) image and a thresholded CBMA image, and $|I \cap C|$ is the number of non-zero voxels in their intersection. DSM ranges from 0 (no overlap), to 1 (perfect overlap).

If the gold standard is taken as "truth", we can compute the traditional (asymmetric) similarity measures, the true positive rate (TPR), and the false positive rate (FPR):

$$\mathrm{TPR} = \frac{|C \cap I|}{|I|} \tag{2}$$

$$\mathrm{FPR} = \frac{|C \cap (\neg I)|}{|\neg I|} \tag{3}$$

where $|\neg I|$ are the number of zeroed voxels in the thresholded reference image. The interpretation of TPR is the probability of a CBMA method correctly labelling a voxel as "active", averaged over all truly active voxels, where "true activation" is defined by a threshold applied to the gold standard (see below). Likewise, the FPR is the probability of a CBMA method falsely labelling a voxel as "active", averaged over all truly inactive voxels.

To evaluate CBMA methods with respect to selected IBMA methods, two thresholding schemes are utilized. In the first scheme, three uncorrected p-values (0.001, 0.01 and 0.05) are used to threshold both IBMA and CBMA output maps, providing equal (nominal) false positive rates for each method, and yielding equivalent thresholded images to compute DSM.

In the second thresholding strategy, maps from IBMA and CBMA are each thresholded differently. For IBMA, a 0.05 false discovery rate (FDR) (Nichols and Hayasaka, 2003) corrected threshold is used to create a gold standard map with high sensitivity. For CBMA, the same set of uncorrected p-values as before (0.001, 0.01 and 0.05) is used. With this strategy, the TPR and FPR measures can be computed, while keeping the gold-standard fixed (i.e. it does not change with the CBMA uncorrected p-value threshold).

For each thresholding scheme, each CBMA method is tested over a range of kernel parameters. ALE's kernel parameter is the value of

the Gaussian kernel's standard deviation (σ), and MKDA/KDA's kernel parameter is its indicator kernel's radius (ρ). The aim is to find the optimal setting for each method (for this dataset and a 5 mm FWHM Gaussian smoothing), while comparing CBMA with IBMA. σ values compared are {5, 10, 15, 20, 25, 30, 35, 40, 45, 50}, and ρ values are {5, 10, 15, 20, 25, 30, 35, 40, 45, 50}. For each CBMA method, each threshold (0.001, 0.01 and 0.05) and each kernel parameter, the binary resulting map is compared with two binary

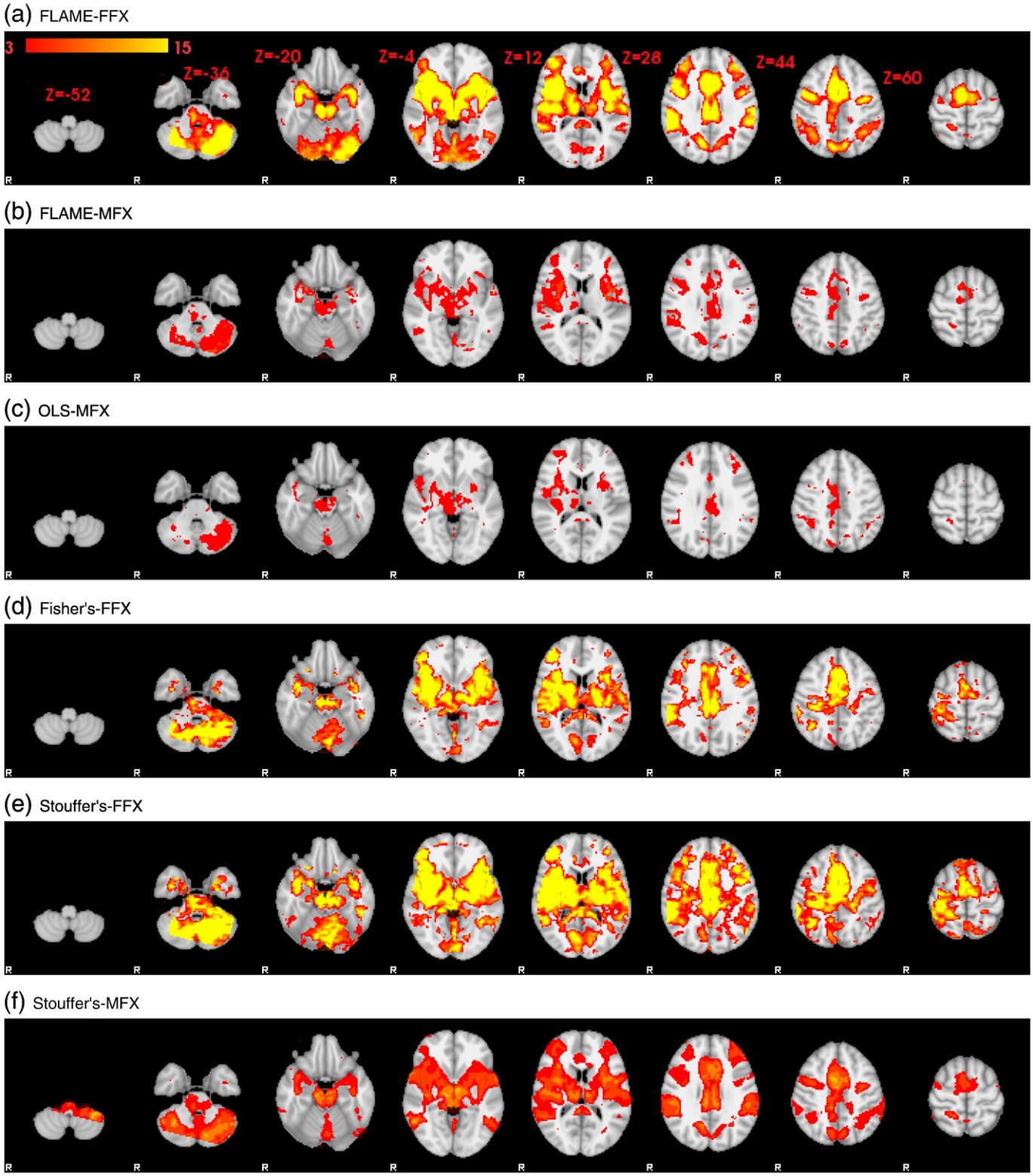

Fig. 2. IBMA maps resulting from MFX at second level and different IBMA methods at third level. *Z*-stat maps are converted to their corresponding *p*-value maps and then, to give clearer visualization, the $-\log_{10} p$ map is shown (with min–max of 3–15). As can be seen, FLAME-MFX and OLS show less extended activations, in areas of consistency across studies. Slice locations are in mm in MNI space.

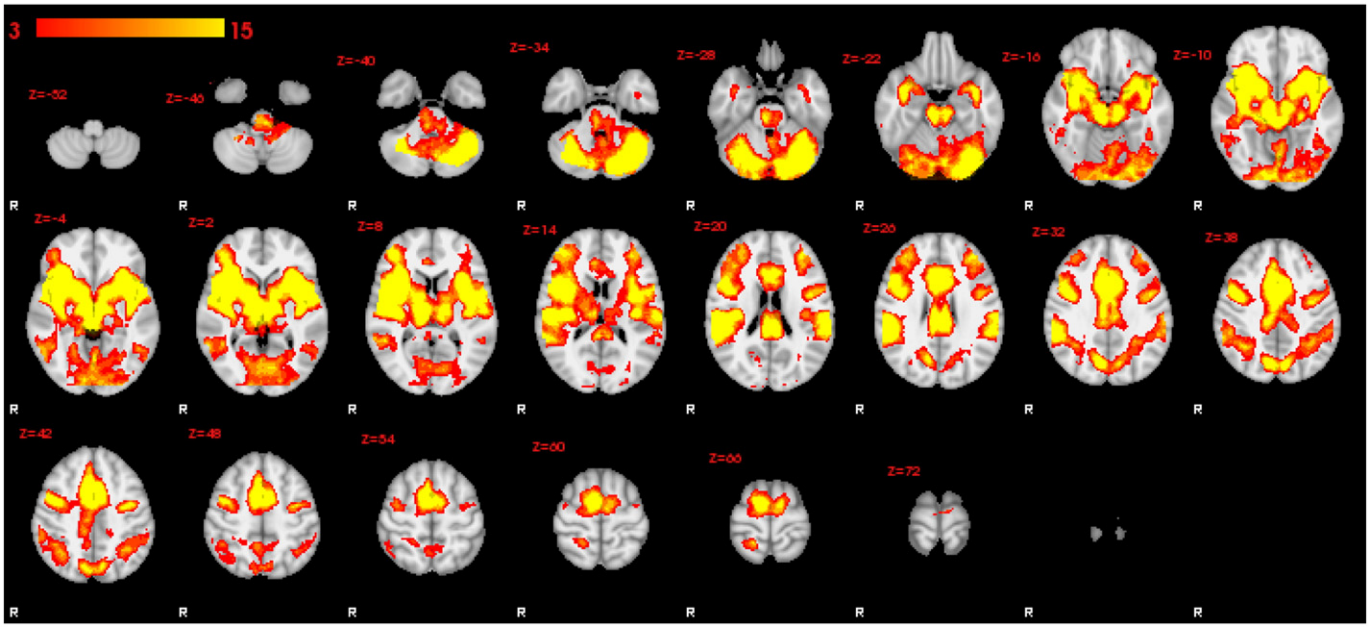

Fig. 3. Gold standard against which CBMA methods are compared. This is the resulting map from a three-level hierarchical analysis, with FLAME-FFX at the third level. Color overlays show $-\log_{10} p$ map (with min–max of 3–15). Slice locations are in mm in MNI space.

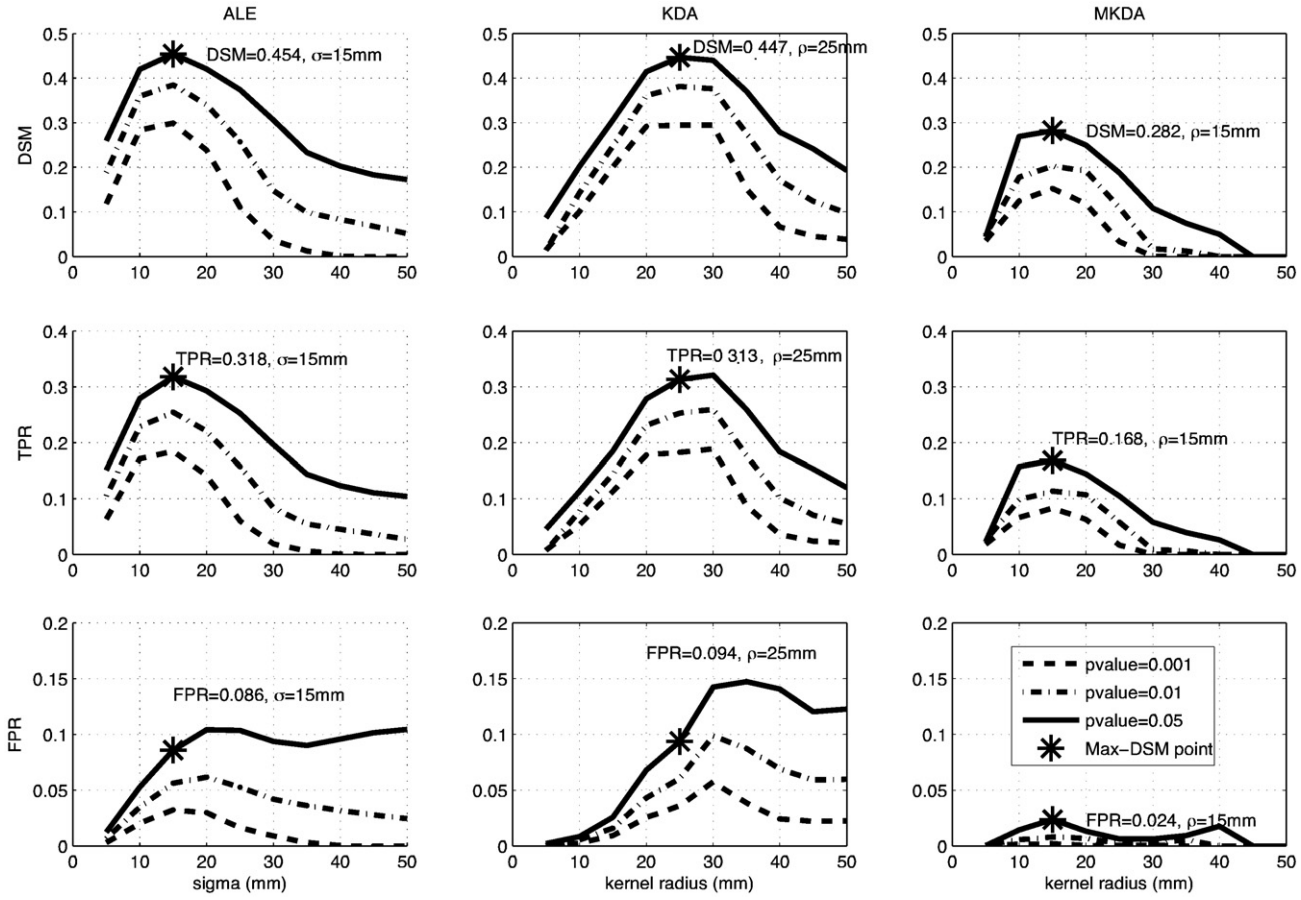

Fig. 4. Evaluating CBMA methods for different kernel parameter values with respect to the gold-standard map. In the first column, DSM, TPR and FPR of the ALE method are shown. In the second and third columns, the same performance measures are shown for KDA and MKDA. In all plots, the x-axis is the kernel parameter (σ for ALE and ρ for KDA and MKDA). To estimate the DSM value, images are thresholded at different uncorrected p-values (shown in the legend) and then binary images are compared. Note that in this plot all scores are for the first thresholding scheme. For the second thresholding scheme, plots are very similar, but with smaller DSM scores overall. More liberal thresholding yields higher DSMs (* indicates the coordinate corresponding to maximum DSM).

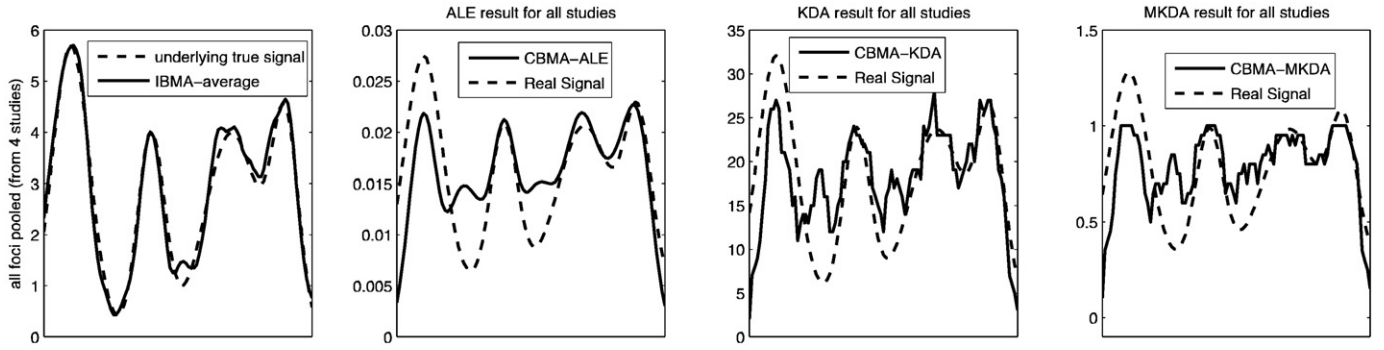

Fig. 5. Illustration of a 20-study 1-dimensional meta-analysis. Using the same setting as in Fig. 1 except with 20 studies, MKDA's estimate is more similar to ALE and KDA's estimate.

references (the first and second thresholding schemes), and each comparison yields a single DSM. The larger the DSM, the better the method–threshold–kernel combination.

Results

Initial IBMA results showed a severely eroded analysis mask. Investigation of each subject's mask (with a cine loop) revealed a small number of subjects with functional-anatomical registration problems. After resolving these problems a full-brain analysis mask was obtained.

Fig. 2 shows statistic maps for the IBMA methods considered, each thresholded at uncorrected $p=0.001$, to give an indication of the differences between the various FFX and MFX image-based combining methods. The figure shows a clear distinction between the FFX methods (Figs. 2a, d, and e) and the MFX methods (Figs. 2b and c), with the FFX showing considerably more activation. The FFX result based on a hierarchical model (Fig. 2a, FLAME-FFX) shows a smoother profile of activation, while the classic meta- analytic statistics (Fig. 2d Fisher's, Fig. 2e Stouffer's) were more irregular, perhaps indicating their greater sensitivity to individual (instead of average) study

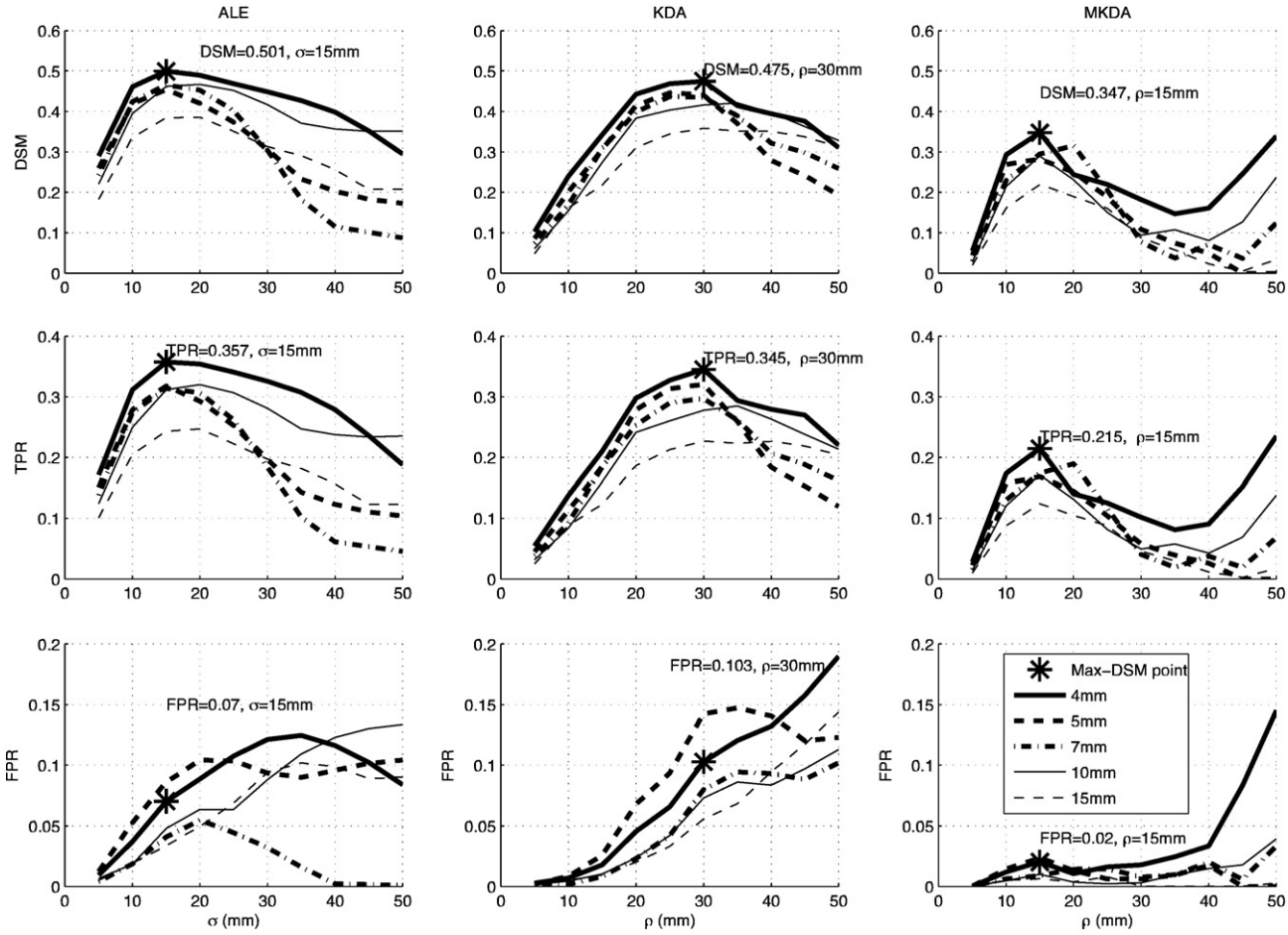

Fig. 6. Evaluating the effect of smoothing extent of studies on optimal CBMA methods' kernel parameter values with respect to their corresponding gold standard map. The first, second and third columns show the results for ALE, KDA and MKDA, respectively. The first, second and third rows show DSM, TPR and FPR for each CBMA method, respectively (* indicates the coordinate corresponding to maximum DSM). Each line in each subplot corresponds to one FWHM from 4, 5, 7, 10 and 15 mm, and all CBMA maps are thresholded at 0.05 ($p_{\text{thresh}}=0.05$ uncorrected).

significance. A complete map of the FLAME-FFX gold standard is shown in Fig. 3.

The evaluation of CBMA methods as a function of kernel parameter is shown in Fig. 4 for the first thresholding scheme (same uncorrected threshold for IBMA and CBMA). Results for the second thresholding scheme had lower DSM scores overall and are qualitatively similar (not shown). For all methods, the best DSM was for the most liberal p-value threshold considered (0.05). For DSM and TPR, the curves generally had the same shape, with an optimal kernel parameter value that was consistent over different p-value thresholds.

Among all CBMA methods, ALE seems to yield the best results overall, with KDA performing similarly and MKDA being more conservative with respect to our gold standard. This conservativeness could be due to MKDA treating studies as independent units, instead of each focus, suggesting that it would require more studies to obtain a similar consistency map to KDA. As can be seen in a 1D illustration in Fig. 5, increasing the number of studies makes MKDA's statistic more similar to the gold standard. The false positive rate for ALE and KDA analyses increases with kernel size, and the "optimal" kernel in both ALE and KDA has a false positive rate close to 0.1. Setting a kernel size of 5 mm for ALE or 20 mm for KDA limits their FPR to .05, which puts them on more similar footing in terms of DSM to MKDA. Thus, the MKDA's DSM is possibly lower because it is more conservative and more similar to a MFX analysis.

The optimal kernel parameter values (shown in Fig. 4) can depend on the amount of first level smoothing applied to studies from which foci are collected. We investigate this by repeating the entire comparative analysis on the basis of 4, 5, 7, 10 and 15 mm FWHM first level smoothing (with the gold standard map re-defined for each smoothing). The DSM results of these comparisons are shown in Fig. 6. As these plots show, the optimal kernel parameter is not very sensitive to smoothing extent; particularly when it varies in the range of 4–10 mm (which is the typical smoothing range used in fMRI studies).

We find the optimal kernel parameter for each CBMA method to be: σ=15 mm for KDA, ρ=25 mm for KDA and ρ=15 mm for MKDA. Fig. 7 compares the gold standard IBMA result to the DSM-optimized CBMA results. Note the dramatic difference in the sensitivity and the overlap pattern of detected regions.

Using the optimal settings we also tested a contrast between two sub-groupings of the 15 studies. There were 9 studies with thermal pain and 6 studies with mechanical pain. We examined the IBMA and CBMA inference for just thermal (Therm), just mechanical (Mech) and their difference, (Therm>Mech) and (Therm<Mech), shown in Figs. 8 a, e and 9 a, e. To generate the Therm/Mech contrast image, foci are collected only from those studies using thermal/mechanical pain stimuli. All of the CBMA analyses are performed using ALE, KDA and MKDA with σ=15 mm, ρ=25 mm and ρ=15 mm, respectively. Results from this analysis are shown in Figs. 8b–d, f–h and 9b–d, f–h.

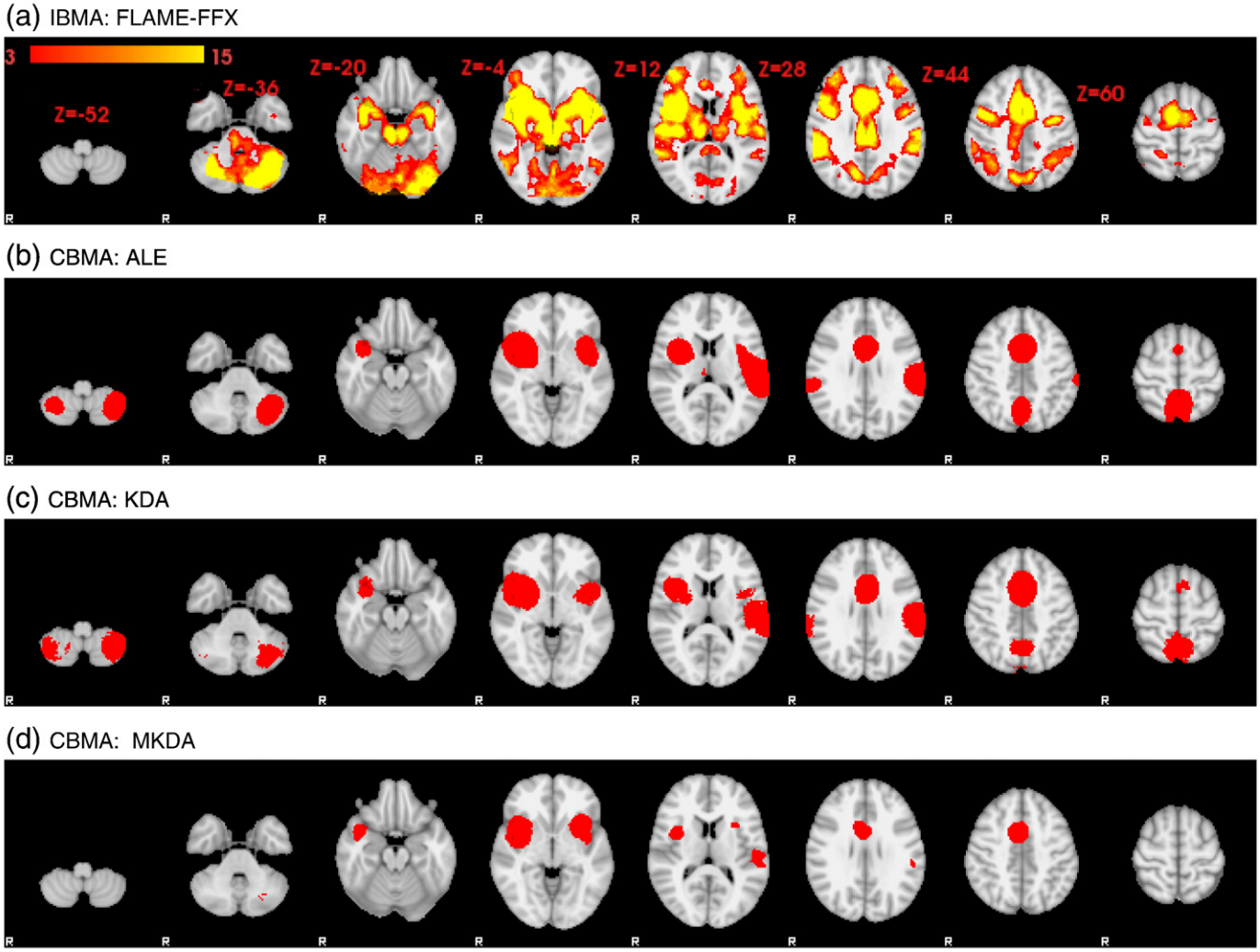

Fig. 7. Gold-standard IBMA map (panel a) shown with CBMA maps. Color overlays show $-\log_{10} p$ values (with min–max of 3–15). Maps in panel b, c and d are resulting from ALE with σ=15 mm, KDA with ρ=25 mm and MKDA with ρ=15 mm, respectively. Slice locations are in mm in MNI space.

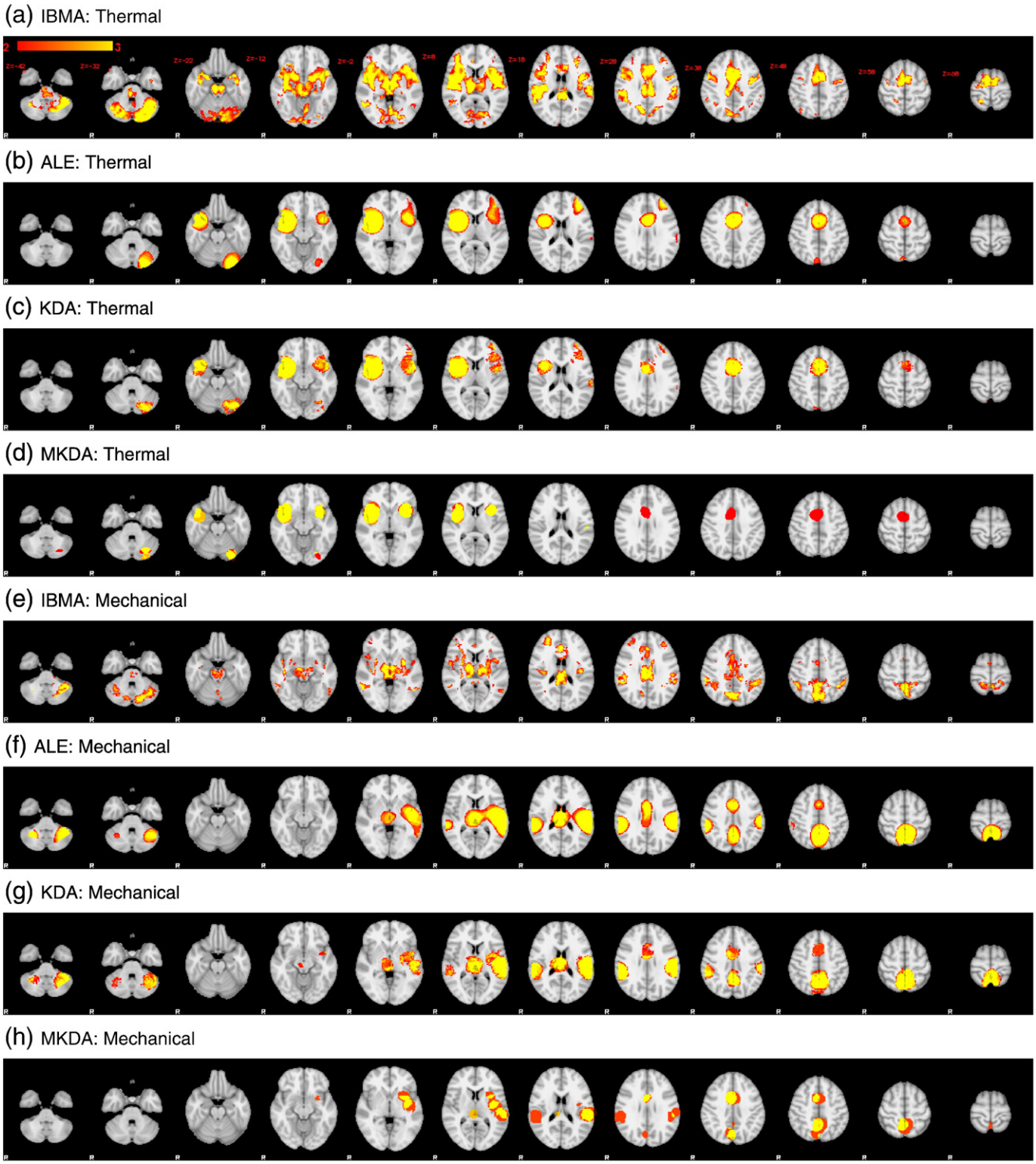

Fig. 8. Using a third-level design to answer cognitive pain questions both in IBMA and CBMA. $-\log_{10} p$ overlay maps are shown (with min–max of 2–3). Slice locations are in mm in MNI space.

In studies using both thermal and mechanical pain stimuli, activity is widely extended across the cortex. For example, both sets of studies activate the attentional network, including the parietal sulcus. Note that we cannot unambiguously attribute Therm vs. Mech effects to differences in pain perception, as there are multiple confounding factors. For example, in most of the mechanical studies, stimuli were delivered to the right foot, while the thermal stimuli were delivered to the left arm. This confounding effect can be seen clearly in activation maps as a lateralization effect, where the thermal stimuli cause activation in right somatosensory cortex, while mechanical stimuli cause more activation on left somatosensory cortex. Also note that mechanical activations are more medial

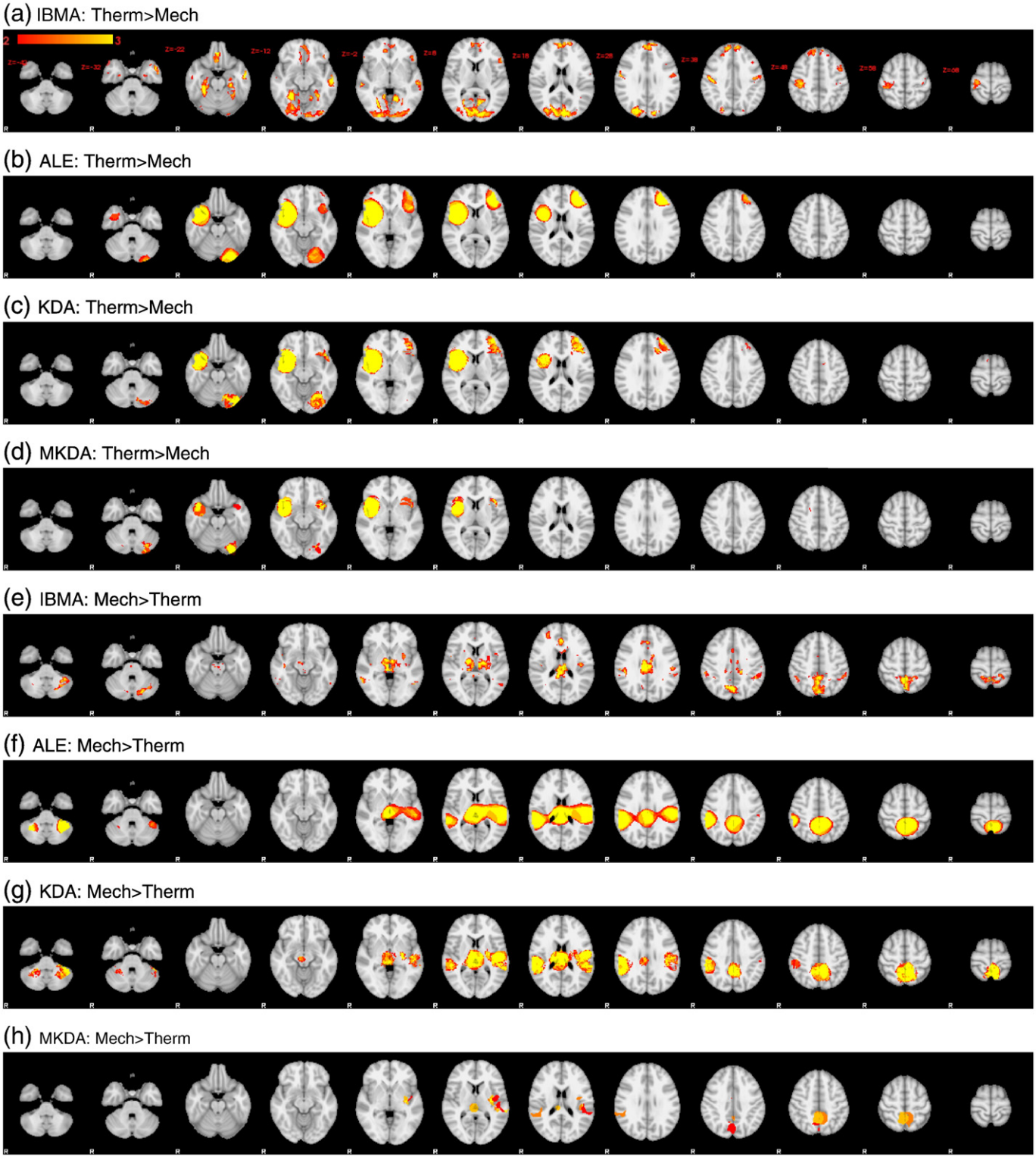

Fig. 9. Using difference contrasts to answer cognitive pain questions both in IBMA and CBMA, for the Therm>Mech and Mech>Therm contrasts. $-\log_{10} p$ maps are shown (with min–max of 2–3). Slice locations are in mm in MNI space.

(Fig. 8e), while the thermal activations are more lateral (Fig. 8b), consistent with typical somatosensory findings (Becerra et al., 2006; Borsook et al., 2008).

The other confound arising from the studies' experimental setups are the difference in visual cortex activity. Most thermal studies used a visual analogue scale (VAS), while studies using mechanical stimuli instructed subjects to close their eyes during the experiment (compare visual cortex in Figs. 8b and e).

Figs. 8b–d, f–h and 9b–d, f–h show the corresponding results for the coordinate-based methods. Note that the CBMA and IBMA results are more similar for the Therm and Mech effects, and less so for the Therm>Mech and Therm<Mech results. This differential perfor-

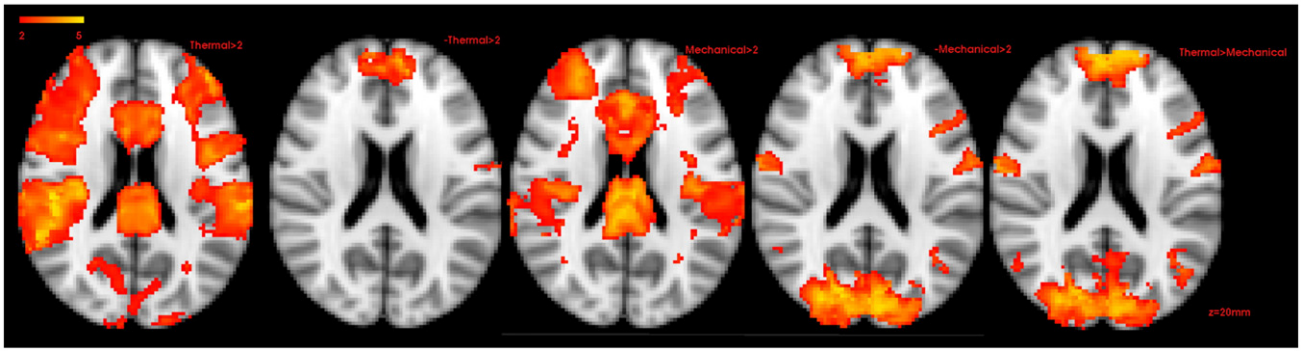

Fig. 10. The effect of deactivation information in difference contrasts for a sample slice ($z=20$ mm in MNI space). This figure shows z-statistic images ($z>2$) for Therm, -Therm, Mech, -Mech, and Therm>Mech, from left to right respectively. It can be seen that the reason for having significant areas corresponding to Therm>Mech that are not significant in Therm is associated with deactivation in Mech.

mance is likely due to the lack of information about activation decreases in the coordinate-based data. For example, while Therm>Mech can be significant if Therm activation is greater than Mech, it can also be significant if the Mech deactivation is greater than Therm deactivation, which cannot be reconstructed by positive activation foci. This highlights a potential danger of omitting deactivation foci from such a comparison (see Fig. 10).

Finally, note the much greater spatial detail available in the IBMA results. These show a greater general richness, as well as greater sensitivity for finding small activation areas, again attributable to IBMA's retention of all data.

Discussion and conclusions

In this paper we have tried to assess the information lost from working only with foci when the voxel-level data are available, as well as highlight the importance of kernel parameters in a typical CBMA. Using a group of 15 pain studies, all analyzed in a similar fashion, we generated a gold-standard map using a three-level hierarchical model, and generated foci to produce the data that would normally be used for a CBMA. While we know of no other work that considers IBMA and CBMA in such a parallel way with a common set of studies, we stress that our findings may depend on the number of studies considered, number of subjects in each, scanning techniques, and chosen foci extraction/reporting style. None-the-less, we believe our collection of pain studies are representative of common practice and are useful for the evaluations considered.

Using DSM plots, it is clear that CBMA methods cannot produce the same map as an IBMA. The best result for CBMA comes from ALE for $\sigma=15$ mm, which is just DSE=0.45. Using this evidence, it seems a necessary future concern for the neuroimaging field to find a way to share datasets (Toga, 2002; Van Horn et al., 2004). Full data need not be transferred, rather just sufficient statistics, that is effect magnitude estimates and their standard errors. However there are a multitude of issues to be taken into consideration for a successful future sharing policy which are beyond the scope of this paper. These issues include description of the experimental design (Miller et al., 2001; Liu and Frank, 2004; Liu, 2004; Smith et al., 2007), image acquisition, and analysis techniques adopted in different research groups and institutes (Friedman et al., 2006, 2007;Zou et al., 2005).

Hierarchical GLM models can be applied simply by using summary statistic maps (like contrasts of effect sizes and their variances) (Beckman et al., 2003; Woolrich et al., 2004). As an appropriate way of doing IBMA, there is no hard rule about whether to use FFX or MFX at the third level. We propose, in general, using FFX at the third level on the basis that individual studies are valid samples from the population and, even if just one is significant, this is valid information to drive a meta-analysis. If there is greater concern about the quality of the constituent studies in a meta-analysis, a MFX approach would be a safer approach, and would only identify consistent results relative to the inter-study variability.

Two important parameters that all three CBMA techniques depend on are kernel parameter and significance threshold. We tried to investigate both of these factors to find the optimal combination of these parameters to maximize the IBMA–CBMA similarity (at least for our datasets). The main reason for adopting a voxel-wise comparison with a fixed, uncorrected p-value threshold is to have comparable thresholds for all methods. For example, using a threshold from FDR would create adaptively-determined thresholds for each result. This is also the reason why we adopted the same protocol in CBMA techniques.

For the other important factor, kernel parameter, we tried to find a good kernel value given typical first-level analyses (FWHM=5 mm first-level smoothing). Our recommendation for these parameters is $\sigma=15$ mm for ALE, $\rho=25$ mm for KDA and $\rho=15$ mm for MKDA. These values are dependent on our 15-study sample, but can provide a guide for other similar data. Although this raises another weak-point of CBMA—that the most optimal setting can vary from one dataset to another—the same could be said with respect to the effect that first-level smoothing has on IBMA approaches.

Comparing results from analysis techniques using DSM is not necessarily the best way for all such comparisons. DSM is a combination of TPR and FPR and in other comparisons/applications it may not be necessary for all variables to have the same weight. For example, in cases where FPR is the most important variable, a method such as MKDA may then appear to perform relatively better, in spite of having a smaller overall DSM. Another issue that might cause MKDA to be more desirable than ALE and KDA is cases where there is a large difference in the number of foci extracted from each study. In such cases, pooling in ALE or KDA style can be highly biased toward studies with higher numbers of foci. As the MKDA technique looks for consistency over studies (by using studies as input units), outlier studies and foci will have less chance of having noticeable effects on the final result.

The obvious primary weak point of CBMA techniques arises from discarding a huge amount of information, simply by using coordinates of maxima. When a comparison is made between different conditions, there will be further loss of accuracy if deactivations are not included. In case of having two contrasts, C_1 and C_2, a difference contrast like C_1-C_2 can be significant if C_1 is more active than C_2, or C_1 is less deactive than C_2. The first case can be *partially* assessed by activation foci, while the second case cannot be assessed in the absence of deactivation foci. In short, the difference between IBMA and CBMA group comparisons can be due to either omission of decreases resulting in CBMA false negatives, or thresholding artifacts resulting in CBMA false positives.

This was highlighted by the differences in our thermal–mechanical comparisons between the CBMA and IBMA results. Based on this, it is strongly recommended to include deactivation foci in CBMA, as well

as activation foci, to have a more accurate and reliable result. However, of course, the greater data reduction implicit in CBMA approaches is considerably more convenient than needing to provide full summary images from all studies; CBMA can even be carried out purely on the basis of activations reported in journal papers.

We have offered a three-part justification on why the IBMA FLAME-MFX analysis should be a gold standard. As further evidence, if CBMA were a better choice, and in fact IBMA were less sensitive than CMBA, this would be reflected by CMBA having essentially perfect power relative to IBMA gold standard. Instead, CMBA shows quite poor power relative to the IBMA reference, and thus further justifies the choice of reference method.

Finally, we note that the recommended IBMA mega-analysis method (hierarchical linear modeling) depends on having comparable contrast (and standard error) images for each study, unlike the CBMA methods, and other IBMA methods, which are based only on *t*- or *z*-statistics that are invariant to design matrix or contrast scaling. All of the IBMA methods can be affected by corrupted masks for one or more subjects, resulting in excessive erosion of the analysis mask. These issues simply highlight the importance of careful quality control of the analysed data to maximize the interpretability of the final results.

Acknowledgments

The authors would like to thank Irene Tracey, Merle Fairhurst, Siri Graff Leknes and Mike Lee for advice and for providing the datasets used in this paper. Salimi-Khorshidi is funded by a Dorothy Hodgkin Postgraduate Award, provided by the UK Research Councils and GlaxoSmithKline.

Appendix A. Pseudo-code for ALE method

```
Activation Likelihood Estimation (ALE) Pseudo-code
begin
 for all foci in the list
 begin
  for all voxels in the brain
   calculate the activation likelihood using the Gaussian kernel
 end
 for a necessary number of times (i.e., 10000 times)
 begin
  generate a group of randomly located foci (uniformly distributed over whole gray matter)
  for all these fake foci
  begin
   for all voxels in the brain
    calculate the activation likelihood using the Gaussian kernel
  end
 end
 test the real AL map wrt AL maps from Monte Carlo approach
end
```

Appendix B. Pseudo-code for KDA method

```
Kernel Density Approximation (KDA) pseudo-code
begin
 for all foci in the list
 begin
  for all voxels in the brain closer than a radius to each foci
   increment voxel's intensity(with step of one)
 end
 for a necessary number of times (i.e., 10000 times)
 begin
  generate a group of randomly located foci (uniformly distributed over whole gray matter)
  for all these fake foci
  begin
   for all voxels in the brain
    increment voxel's intensity(with step of one)
  end
 end
 test the real density map wrt density maps from Monte Carlo approach
end
```

Appendix C. Pseudo-code for KDA method

```
Multi-level Kernel Density Analysis (KDA) pseudo-code
begin
 for all studies
 begin
  generate the KDA map
  make comparison indicator maps (CIM) by binarising the map wrt 0.5
 end
 weight and average CIMs to make the proportion of study comparison maps (PSCM)
 for a necessary number of times (i.e., 10000 times)
 begin
  for all CIMs
  begin
   extract the blobs
   move the centers of blobs randomly (Monte Carlo part)
  end
  make the PSCM maps
 end
 test the real PSCM map map wrt distribution of Monte Carlo-generated PSCM maps
end
```

References

Becerra, L., Morris, S., Bazes, S., Gostic, R., Sherman, S., Gostic, J., Pendse, G., Moulton, E., Scrivani, S., Keith, D., Chizh, B., Borsook, D., 2006. Trigeminal neuropathic pain alters responses in cns circuits to mechanical (brush) and thermal (cold and heat) stimuli. J. Neurosci. 26 (42), 10646–10657.

Beckmann, C.F., Jenkinson, M., Smith, S.M., 2003. General multilevel linear modeling for group analysis in FMRI. NeuroImage 20 (2), 1052–1063.

Borsook, D., Moulton, E.A., Tully, S., Schmahmann, J.D., Becerra, L., 2008. Human cerebellar responses to brush and heat stimuli in healthy and neuropathic pain subjects. Cerebellum 1–21.

Chien, 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.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Fox, P.T., Lancaster, J.L., Parsons, L.M., Xiong, J., Zamarripa, F., 1997. Functional volumes modeling: theory and preliminary assessment. Hum. Brain Mapp. 5 (4), 306–311.

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., Rainey, L., Lancaster, J.L., 1999. Functional volumes modeling: scaling for group size in averaged images. Hum. Brain Mapp. 8, 143–150.

Friedman, L., Glover, G.H., Krenz, D., Magnotta, V., 2006. Reducing inter-scanner variability of activation in a multicenter fMRI study: role of smoothness equalization. NeuroImage 32 (4), 1656–1668.

Friedman, L., Stern, H., Brown, G.G., Mathalon, D.H., Turner, J., Glover, G.H., Gollub, R.L., Lauriello, J., Lim, K.O., Cannon, T., Greve, D.N., Bockholt, H.J., Belger, A., Mueller, B., Doty, M.J., He, J., Wells, W., Smyth, P., Pieper, S., Kim, S., Kubicki, M., Vangel, M., Potkin, S.G., 2007. Test-retest and between-site reliability in a multicenter fMRI study. Hum. Brain Mapp. 29 (8), 958–972.

Iannetti, G.D., Zambreanu, L., Wise, R.G., Buchanan, T.J., Huggins, J.P., Smart, T.S., Vennart, W., Tracey, I., 2005. Pharmacological modulation of pain-related brain activity during normal and central sensitization states in humans. Proc. Natl. Acad. Sci. U. S. A. 102 (50), 18195–18200.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17 (2), 825–841.

Laird, A.R., Lancaster, J.L., Fox, P.T., 2005a. Brainmap: the social evolution of a functional neuroimaging database. Neuroinformatics 3, 65–78.

Laird, A.R., Fox, P.M., Price, C.J., Glahn, D.C., Uecker, A.M., Lancaster, J.L., Turkeltaub, P.E., Kochunov, P., Fox, P.T., 2005b. Ale meta-analysis: controlling the false discovery rate and performing statistical contrasts. Hum. Brain Mapp. 25 (1), 155–164.

Lazar, N.A., Luna, B., Sweeney, J.A., Eddy, W.F., 2002. Combining brains: a survey of methods for statistical pooling of information. NeuroImage 16 (2), 538–550.

Lee, M.C., Zambreanu, L., Menon, D.K., Tracey, I. Identify brain activity specifically related to the maintenance and perceptual consequence of central sensitization. manuscript in preparation.

Leknes, S., Berna, C., Tracey, I. Are pessimists ever better off? The case of rewarding 'phew effect' relief. manuscript in preparation.

Leknes, S., Duncan, K., Fairhurst, M., Wiech, K., Tracey, I. Effects of anticipatory pleasure on pain processing. manuscript in preparation.

Leknes, S., Snyder, G.D., Lee, M., Tracey, I. When hedonic contrast renders pain pleasant: an fmri study. manuscript in preparation.

Liu, T.T., 2004. Efficiency, power and entropy in event-related FMRI with multiple trial types. part ii: design of experiments. NeuroImage 21, 400–413.

Liu, T.T., Frank, L.R., 2004. Efficiency, power, and entropy in event-related fMRI with multiple trial types: Part I. Theory. NeuroImage 21, 387–400.

Miller, K.L., Luh, W.M., Liu, T.T., Martinez, A., Obata, T., Wong, E.C., Frank, L.R., Buxton, R.B., 2001. Nonlinear temporal dynamics of the cerebral blood flow response. Hum. Brain Mapp. 13 (1), 1–12.

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.E., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat. Methods Med. Res. 12 (5), 419–446.

Nielsen, F.A., Hansen, L.K., 2002. Modeling of activation data in the brainmap database: detection of outliers. Hum. Brain Mapp. 15 (3), 146–156.

Smith, S.M., Bannister, P.R., Beckman, C., Brady, M., Clare, S., Flitney, D., Hansen, P., Jenkinson, M., Leibovici, D., Ripley, B., Woolrich, M., Zhang, J., 2001. FsI: new tools for functional and structural brain image analysis. NeuroImage 13 (6), 249.

Smith, S., Jenkinson, M., Beckmann, C., Miller, K., Woolrich, M., 2007. Meaningful design and contrast estimability in FMRI. NeuroImage 34 (1), 127–136.

Sutton, A.J., Jones, D.R., Abrams, K.R., Sheldon, T.A., Song, F., 2000. Methods for Meta-Analysis in Medical Research. John Wiley, London.

Thirion, B., Pinel, P., Meriaux, S., Roche, A., Dehaene, S., Poline, J.B., 2007. Analysis of a large fMRI cohort: statistical and methodological issues for group analyses. NeuroImage 35 (1), 105–120.

Toga, A.W., 2002. Neuroimage databases: the good, the bad and the ugly. Nat. Rev. Neurosci. 3 (4), 302–309.

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 (3.1), 765–780.

Van Horn, J.D., Grafton, S.T., Rockstroh, D., Gazzaniga, M.S., 2004. Sharing neuroimaging studies of human cognition. Nat. Neurosci. 7 (5), 473–481.

Wager, T.D., Jonides, J., Reading, S., 2004. Neuroimaging studies of shifting attention: a meta-analysis. NeuroImage 22 (4), 1679–1693.

Wager, T.D., Lindquist, M., Kapla, L., 2007. Meta-analysis of functional neuroimaging data: current and future directions. Soc. Cogn. Affect. Neurosci. 2 (2), 150–158.

Woolrich, M.W., Ripley, B.D., Brady, M., Smith, S.M., 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. NeuroImage 14 (6), 1370–1386.

Woolrich, M.W., Behrens, T.E., Beckmann, C.F., Jenkinson, M., Smith, S.M., 2004. Multilevel linear modelling for FMRI group analysis using Bayesian inference. NeuroImage 21 (4), 1732–1747.

Worsley, K.J., Liao, C.H., Aston, J., Petre, V., Duncan, G.H., Morales, F., Evans, A.C., 2002. A general statistical analysis for fMRI data. NeuroImage 15 (1), 1–15.

Zou, K.H., Greve, D.N., Wang, M., Pieper, S.D., Warfield, S.K., White, N.S., Manandhar, S., Brown, G.G., Vangel, M.G., Kikinis, R., Wells III, W.M., 2005. Reproducibility of functional MR imaging: preliminary results of prospective multi-institutional study performed by Biomedical Informatics Research Network. Radiology 237 (3), 781–789.