NeuroImage 117 (2015) 56–66

Contents lists available at ScienceDirect

NeuroImage

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

Test-retest reliability of dynamic causal modeling for fMRI

Stefan Frässle [a,b,*], Klaas Enno Stephan [c,d], Karl John Friston [d], Marlena Steup [a], Sören Krach [e], Frieder Michel Paulus [e,1], Andreas Jansen [a,1]

[a] Section of Brainimaging, Department of Psychiatry, University of Marburg, 35039 Marburg, Germany
[b] Department of Child and Adolescent Psychiatry, University of Marburg, 35039 Marburg, Germany
[c] Translational Neuromodeling Unit (TNU), Institute of Biomedical Engineering, University of Zurich & ETH Zurich, Zurich, Switzerland
[d] Wellcome Trust Centre for Neuroimaging, University College London, London, UK
[e] Social Neuroscience Lab | SNL, Department of Psychiatry and Psychotherapy, University of Lübeck, 23538 Lübeck, Germany

ARTICLE INFO

Article history:
Accepted 15 May 2015
Available online 22 May 2015

Keywords:
DCM
Test-retest reliability
fMRI
Motor
Priors
Hyperpriors
Conditional dependencies
Empirical Bayes

ABSTRACT

Dynamic causal modeling (DCM) is a Bayesian framework for inferring effective connectivity among brain regions from neuroimaging data. While the validity of DCM has been investigated in various previous studies, the reliability of DCM parameter estimates across sessions has been examined less systematically. Here, we report results of a software comparison with regard to test-retest reliability of DCM for fMRI, using a challenging scenario where complex models with many parameters were applied to relatively few data points. Specifically, we examined the reliability of different DCM implementations (in terms of the intra-class correlation coefficient, ICC) based on fMRI data from 35 human subjects performing a simple motor task in two separate sessions, one month apart. We constructed DCMs of motor regions with fair to excellent reliability of conventional activation measures. Using classical DCM (cDCM) in SPM5, we found that the test-retest reliability of DCM results was high, both concerning the model evidence (ICC = 0.94) and the model parameter estimates (median ICC = 0.47). However, when using a more recent DCM version (DCM10 in SPM8), test-retest reliability was reduced notably. Analyses indicated that, in our particular case, the prior distributions played a crucial role in this change in reliability across software versions. Specifically, when using cDCM priors for model inversion in DCM10, this not only restored reliability but yielded even better results than in cDCM. Analyzing each component of the objective function in DCM, we found a selective change in the reliability of posterior mean estimates. This suggests that tighter regularization afforded by cDCM priors reduces the possibility of local extrema in the objective function. We conclude this paper with an outlook to ongoing developments for overcoming the software-dependency of reliability observed in this study, including global optimization and empirical Bayesian procedures.

Introduction

While early neuroimaging studies focused on the functional specialization of brain regions (i.e., the localization of task-dependent neuronal activation), it is now generally accepted that any cognitive process rests on multiple brain regions acting in concert (i.e., functional integration). Hence, assessing the functional integration among regions is essential for understanding a particular brain function. Researchers have therefore addressed various aspects of brain connectivity, including graph theoretical approaches, resting-state networks, the concept of the connectome, or biophysical modelling (for a review on fMRI-based inference on connectivity, see Smith, 2012). Whereas all of these approaches have their advantages and drawbacks, a common requirement is the definition of network structure in terms of nodes (neuronal populations or brain regions) and edges (anatomical connections among the nodes). Given such a structural model, connectivity can be characterized in terms of functional connectivity (defined by mere statistical relationships between the nodes of the network) or effective connectivity, which refers to directed interactions between nodes and typically rests on mechanistic models of brain responses (for a review of different methods, see Valdes-Sosa et al., 2011). One frequently used method to infer effective connectivity from fMRI data is dynamic causal modeling (DCM; Friston et al., 2003). Specifically, DCM focuses on how directed interactions among brain regions are perturbed by experimental manipulations (e.g., sensory stimulation, task demands). Using DCM, researchers have gained deeper insight into the mechanisms underlying cognitive tasks, such as visuospatial attention (Kellermann et al., 2012; Siman-Tov et al., 2007), face perception (Fairhall and Ishai, 2007; Li et al., 2010; Nguyen et al., 2014), working memory (Dima et al., 2014), decision making (Stephan et al., 2007a; Summerfield et al., 2006; Summerfield and Koechlin, 2008), motor processes (Grefkes et al., 2008; Grol et al., 2007), and the "resting state" (Goulden et al., 2014; Di and Biswal, 2014; Friston et al., 2014). In this

* Corresponding author at: University of Marburg, Section of Brainimaging, Department of Psychiatry, Rudolf-Bultmann-Straße 8, 35039 Marburg, Germany.
E-mail address: fraessle@med.uni-marburg.de (S. Frässle).
[1] Contributed equally to this work.

http://dx.doi.org/10.1016/j.neuroimage.2015.05.040
1053-8119/

regard, research questions have become increasingly complex, leading to ongoing refinements and extensions of DCM for fMRI, such as (i) two-state DCM (Marreiros et al., 2008), allowing for two neuronal states per brain region to model the activity of inhibitory and excitatory populations; (ii) nonlinear DCM (nlDCM; Stephan et al., 2008), which accounts for synaptic gating; i.e., the modulatory influence of a neuronal population on the connection between two other populations; and (iii) stochastic DCM (sDCM; Daunizeau et al., 2009; Friston et al., 2008, 2010), that allows for endogenous (stochastic) fluctuations at the neuronal level. More recently, DCM studies have been conducted in clinical settings in order to gain a better understanding of pathophysiological mechanisms of psychiatric disorders such as major depression (Almeida et al., 2009; Schlösser et al., 2008), autism (Grèzes et al., 2009; Radulescu et al., 2013) or schizophrenia (Deserno et al., 2012; Brodersen et al., 2014; Roiser et al., 2013).

So far, numerous studies have examined different aspects of validity of DCM for fMRI, including face validity (Friston et al., 2003; Stephan et al., 2008), construct validity in relation to other models (Lee et al., 2006; Penny et al., 2004), cross-validation against other data modalities (Dima et al., 2009, 2010) of neuroimaging, and predictive validity (Brodersen et al., 2011; David et al., 2008; Reyt et al., 2010). In contrast to validity, another test-theoretical property, reliability, has received comparatively less attention. This refers to the stability of estimates obtained when applying the model to multiple datasets over time, acquired under the same condition in the same subject.

While reliability has been investigated frequently in the context of conventional fMRI activation (e.g., Aron et al., 2006; Brandt et al., 2013; Fliessbach et al., 2010; Friedman et al., 2008; Loubinoux et al., 2001; Plichta et al., 2012; Raemaekers et al., 2007) and functional connectivity studies (e.g., Braun et al., 2012; Birn et al., 2014), the reliability of DCM estimates has received less attention. To date, only two studies have addressed test-retest reliability in the context of DCM for fMRI. Schuyler et al. (2010) examined DCM connectivity parameter estimates for visual and auditory tasks, finding fair to excellent reliability. However, the authors did not report the reliability of another important feature of DCM: Bayesian model selection (BMS), which evaluates the plausibility of competing models with respect to the (log) evidence, a principled measure of the trade-off between model accuracy and model complexity. Rowe et al. (2010) investigated the reliability of DCM using a motor paradigm involving free vs. constrained action selection. They found excellent reliability of model selection, but low test-retest reliability for connectivity parameter estimates, a result they suspected to arise from parameter interdependencies in their particular model.

In the present work, we investigated test-retest reliability of both BMS and model parameter estimation for DCM in a simple task probing hemispheric interactions in the motor network. To this end, 35 right-handed subjects performed visually synchronized unimanual or bimanual hand movements in the MR scanner, a paradigm which was previously established for DCM analysis by Grefkes et al. (2008) and subsequently reused in multiple studies, including patient studies (Grefkes et al., 2010). Subjects in our study performed the experiment twice, in two separate sessions approximately one month apart.

In this paper, we focus entirely on a software evaluation with regard to test-retest reliability of model selection and parameter estimation. A separate quantitative analysis of reproducibility (i.e., how well our parameter estimates reproduce the previous results by Grefkes et al., 2008) will be reported in a separate report (Frässle et al., in preparation).

For the reliability analyses in this paper, we contrasted classical DCM (cDCM as implemented in SPM5), which was also used by Grefkes et al. (2008), to a more recent version (DCM10 as implemented in SPM8, v4290). Notably, DCM10 differs from classical DCM in several ways. Most importantly, it allows for constructing models that factorially combine different variations of the neuronal state equation, resulting in eight different principal forms: {bilinear vs. nonlinear} × {one-state vs. two-state} × {deterministic vs. stochastic}. As a consequence, refinements of various technical details, including the priors and the hemodynamic model, were required when unifying all DCM variants in one common implementation framework. Our analyses using both DCM versions indicated a considerable impact of the software implementation on the reliability of DCM results, with DCM10 being less stable over sessions. This motivated a more thorough investigation of some of the above-mentioned technical refinements from cDCM to DCM10 and their potential role in the reduction of test-retest reliability.

Materials and methods

Subjects

Thirty-five subjects (17 female, 18 male, mean age: 23.5 ± 2.8 years, range: 19–31 years) participated in the experiment. All were healthy, with no history of neurological or psychiatric diseases, brain pathology or abnormal brain morphology on T1-weighted MR images. Subjects were native German speakers and right-handed according to the Edinburgh Inventory of Handedness with a cut-off at +30 (Oldfield, 1971). Prior to the study, each gave informed written consent. The study conformed to the Declaration of Helsinki and was approved by the local ethics committee of the Medical Faculty of the University of Marburg. To investigate test-retest reliability, subjects underwent the identical experiment twice, separated by 32 ± 5 days on average (range: 25–44 days).

Experimental procedure

The experimental paradigm was closely matched to the paradigm established for DCM analysis by Grefkes et al. (2008). An fMRI block design was used, asking subjects to perform visually synchronized whole-hand fist-closing movements. Each block started with an instruction period, followed by a time-variable delay period, a hand movement period and a time-variable resting period. First, an instruction text was shown for 1.5 s, informing subjects which hand to use in the upcoming block. Subjects had to perform hand movements (i.e., whole-hand fist closure) with either the left (condition "LH"), right (condition "RH") or both hands (condition "BH"). The instruction was then replaced by an empty red circle (diameter: 7.26 degrees). The circle was shown for either 1.5, 2.0 or 2.5 s before starting to blink at a rate of 1.5 Hz for 15 s. Subjects were asked to close and open their hands synchronized with the rhythm of the blinking. A white screen, indicating that subjects should rest and wait for the next instruction screen, then replaced the red circle. The resting period lasted for either 11, 11.5 or 12 s, depending on the duration of the variable delay period, such that subsequent task and baseline periods summed to 30 s cycle length. All stimuli were displayed on a video screen (visible through a mirror attached to the MR head coil). The experiment consisted of 24 blocks; hence, each hand-movement condition was performed eight times. The order of LH, RH and BH blocks was pseudorandomized and identical in both sessions. Before the experiment started, subjects were trained outside the MR scanner to guarantee accurate performance. Inside the MR scanner, hand movements were visually inspected from the control room through a glass pane.

Image acquisition

Time courses of subjects' brain activity were acquired using a 3-Tesla MR scanner (Siemens TIM Trio, Erlangen, Germany) with a 12 channel head matrix receive coil at the Department of Psychiatry, University of Marburg. Functional images were obtained using a T_2^*-weighted gradient-echo echo-planar imaging sequence (EPI) sensitive to the Blood Oxygen Level Dependent (BOLD) contrast (33 slices, TR = 2000 ms, TE = 30 ms, matrix size 64 × 64 voxels, voxel size 3.6 × 3.6 × 3.6 mm, gap size 0.4 mm, flip angle 90°). Slices covered the

whole brain and were positioned transaxially parallel to the intercommissural (AC-PC) plane. In each session, 380 functional images were collected; both sessions were 1 month apart from each other. For each subject, an additional high-resolution anatomical image was acquired using a T1-weighted magnetization-prepared rapid gradient-echo (3d MP-RAGE) sequence in sagittal plane (176 slices, TR = 1900 ms, TE = 2.52 ms, matrix size 256 × 256 voxels, voxel size 1 × 1 × 1 mm, flip angle 9°).

Functional imaging data analysis

Preprocessing and analysis of the functional images were conducted using the SPM8 software package (Statistical Parametric Mapping, Wellcome Trust Center for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk) and Matlab (MathWorks). The first four images were discarded from the analysis. To control for small head movements, functional images from both sessions were realigned to the mean image of each subject. Realigned images were coregistered with the high-resolution anatomical image and then spatially normalized into the Montreal Neurological Institute (MNI) standard space using the unified segmentation-normalization of the anatomical image (Ashburner and Friston, 2005). Normalized functional images were spatially smoothed using an isotropic 8 mm full width at half maximum Gaussian kernel.

Voxel-wise BOLD activity was modelled by means of a first-level General Linear Model (GLM; Friston et al., 1995; Worsley and Friston, 1995). The three hand-movement conditions (i.e., LH, RH and BH) were included as block regressors with the above-mentioned stimulus duration. The regressors for the two sessions were entered into two separate sessions in one GLM (i.e., were not concatenated) to allow for conjunction analyses. This resulted in six task regressors (3 per session), which were convolved with SPM8's standard canonical hemodynamic response function (HRF). In addition, the instruction periods and the six realignment parameters of each session were entered into the first-level GLM as nuisance regressors to control for task-independent activation and movement-related artifacts not accounted for by the realignment during preprocessing, respectively. A high-pass filter (cut-off frequency: 1/128 Hz) was used to account for low-frequency noise. Individual BOLD activity related to each of the three hand-movement conditions for each session was identified from the six linear baseline contrasts (i.e., LH, RH, and BH each for both sessions).

The individual baseline contrasts were then entered into a random effects group-level analysis (3 × 2 within-subject ANOVA) to assess BOLD activity for each hand-movement condition and session. Anatomical localization and characterization of the activated brain regions were achieved using the Anatomy toolbox extension within SPM8 (Eickhoff et al., 2005).

Time series extraction

Given the group-level results, eight regions of interest (ROIs) were selected for DCM analyses, in line with previous approaches (Grefkes et al., 2008). These eight ROIs were located bilaterally in the primary motor cortex (M1), the lateral premotor cortex (PMC), the supplementary motor area (SMA) and the motion-sensitive visual area (V5). While bilateral M1, PMC and SMA are key components of the cortical motor system (Rizzolatti and Luppino, 2001), activation in V5 was attributed to the blinking of the red circle (Zeki et al., 1991) and bilateral V5 thus served as input regions in the DCMs. The coordinates of the ROIs were determined for each subject individually, anatomically constrained by masks that were defined by the group-level activations as follows: First, the group-level peak activation coordinates of the conjunction analysis of hand-specific contrasts of both sessions served to identify motor areas. That is, left M1/PMC/SMA were determined from the conjunction of the right-hand baseline contrasts (RH_session1 ∩ RH_session2), and right M1/PMC/SMA were determined from the equivalent left-hand conjunction (LH_session1 ∩ LH_session2). Bilateral visual areas V5 were defined from the conjunction of all three contrasts of both sessions (LH_session1 ∩ RH_session1 ∩ BH_session1 ∩ LH_session2 ∩ RH_session2 ∩ BH_session2). Second, for each of these peak activations, a mask was created that included all voxels significantly activated at $p < 0.05$, family-wise error (FWE) corrected, within a sphere of 16 mm radius (M1 and V5) or 10 mm radius (PMC and SMA) centered on the respective group peak coordinates. The SMA masks were further constrained to one hemisphere (i.e., the mask for the left SMA was restricted to the left hemisphere only). To account for inter-subject variability in the location of the regions, coordinates of each ROI were then defined individually. First, individual ROI center coordinates were defined as the subject-specific maximum within the respective mask (Supplementary Table S1). Second, for each session and subject separately, time series were extracted from each ROI as the first eigenvariate of all voxels that survived $p < 0.001$, uncorrected, within a 4 mm sphere centered on the individual coordinates. Time series were mean-centered and signal variance due to head movements and task instructions was removed (i.e., adjusting the time series with regard to an F-contrast on the effects of interest, thus removing all variance there could be explained by effects of no interest).

Dynamic causal modeling

Dynamic causal modeling (DCM; Friston et al., 2003) is a Bayesian framework for the inversion and comparison of state-space models based on neuroimaging data. It is frequently used in fMRI research to assess effective connectivity within a network of interest, quantifying how experimental manipulations perturb both the neural activity in brain regions and the interactions among those regions. The original implementation described changes in the neural state using the following bilinear differential equation:

$$\frac{dz}{dt} = \left(A + \sum_{j=1}^{m} u_j B^j\right) z + Cu \quad (1)$$

where z represents a vector of neuronal population activities in the regions considered, A contains the endogenous connection strengths, B^j describes the effect of the j-th experimental perturbation on the connections among the network regions (modulatory connectivity), and C represents the strengths of driving inputs. Over the last decade, multiple extensions to the original framework have been suggested, allowing, for instance, the assessment of nonlinear (Stephan et al., 2008) and stochastic effects (Daunizeau et al., 2009; Friston et al., 2010; Friston et al., 2008). In this study, the original bilinear framework was utilized.

Following the approach by Grefkes et al. (2008), four DCMs were constructed, modeling various intra- and interhemispheric interactions of the motor network during uni- and bimanual hand movements. For all models, the pattern of endogenous connectivity and driving inputs (A- and C-matrix) were identical. Since bilateral V5 activity served as driving input for the motor network, exogenous influences (representing the cued response times) were set to excite activity within these regions. Both V5 areas then sent forward endogenous connections to ipsi- and contralateral PMC and SMA. Within the motor network (M1, PMC, SMA), all six ROIs were reciprocally connected to each other ROI, i.e., full endogenous connectivity (cf. Grefkes et al., 2008).

The effects of modulatory inputs on connectivity (B-matrix) differed across the four models and represented hypotheses of how uni- and bimanual hand movements affected connections within the motor network (Fig. 1).

Modulatory input structure differed in complexity, ranging from a sparsely modulated (model 1) to a complex model where all connections were subject to modulatory effects (model 4); additionally, symmetry between hemispheres was considered (asymmetric models 1–2, symmetric models 3–4). Random effects BMS (Stephan et al., 2009) was used to rank models according to their log model evidence

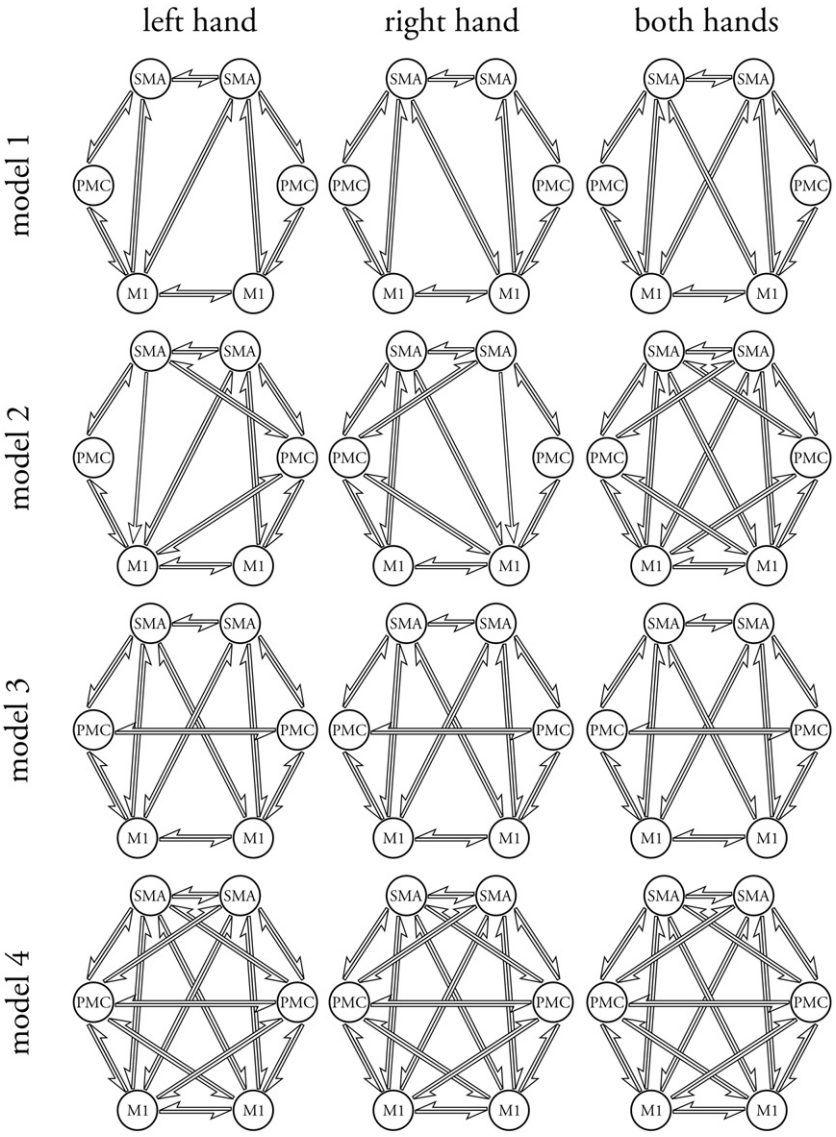

Fig. 1. Variations on the modulatory connectivity (B-matrix) resulting in four different models, implementing distinct hypotheses of how intra- and interhemispheric connections could be modulated by uni- or bimanual hand movements (cf. Grefkes et al., 2008). Models differed in their complexity, ranging from a sparse model (model 1) to a full model (model 4), and in their symmetry (asymmetric models 1–2, symmetric models 3–4).

(i.e., the likelihood of the data given a model). The log model evidence was approximated by the negative free energy, which provides a lower bound to the log model evidence and, under Gaussian assumptions about prior and posterior is given by

$$\begin{aligned} F = & -\frac{1}{2}e_y^T C_y^{-1} e_y - \frac{1}{2}log|C_y| - \frac{N}{2}log2\pi \\ & -\frac{1}{2}e_\theta^T C_\theta^{-1} e_\theta - \frac{1}{2}log|C_\theta| + \frac{1}{2}log|S_\theta| \\ & -\frac{1}{2}e_\lambda^T C_\lambda^{-1} e_\lambda - \frac{1}{2}log|C_\lambda| + \frac{1}{2}log|S_\lambda| \end{aligned} \quad (2)$$

where e_y is the error of the model's prediction, C_y is the error covariance matrix, N is the number of data points, e_θ is the deviation of the posterior from the prior, C_θ is the prior covariance matrix, and S_θ is the posterior covariance matrix (the equivalent notation applies to terms of the hyperparameters λ).

Two different DCM implementations were used for model inversion: classical DCM (cDCM) as implemented in SPM5, and DCM10 as implemented in SPM8 (version: 4290). Please note that experimental inputs were not mean-centered when using either DCM implementation to ensure comparability of parameter estimates across software versions. Choosing the same approach to mean-centering is essential for comparing different DCM implementations, since this affects the meaning of the parameter estimates. For example, the parameters of the endogenous connections correspond to the partial derivative of the neuronal state equation with respect to the neuronal states, when the inputs are zero. Depending on whether inputs are mean-centered or not, this changes the interpretation (i.e., coupling at the average input level vs. coupling when inputs are zero). Comparing results from both software versions (under the same mean-centering settings) allowed for an investigation of how changes to priors and technical refinements made in DCM10 affected the reliability of model selection results and parameter estimation.

Test-retest reliability of BOLD activity and DCM

Test-retest reliability of BOLD activity and DCM results were examined using the intra-class correlation coefficient (ICC). Specifically, the ICC(3,1) type (Shrout and Fleiss, 1979) was utilized, providing

an adequate relation of within-subject (σ^2_{within}) and between-subject variability ($\sigma^2_{between}$) in the context of fMRI:

$$ICC(3,1) = \frac{\sigma^2_{between} - \sigma^2_{within}}{\sigma^2_{between} + \sigma^2_{within}} \quad (3)$$

ICC values range from −1 to 1, and reliability is typically classified as "poor" for ICC < 0.4, as "fair" for 0.4 ≤ ICC < 0.6, as "good" for 0.6 ≤ ICC < 0.75, and as "excellent" for ICC ≥ 0.75 (Cicchetti, 2001). The reliability of BOLD activity was addressed by calculating the ICC(3,1) for each voxel in the brain and for the median contrast values, which were estimated from the subject-specific ROIs entering DCM analyses (i.e., M1, PMC, SMA, and V5 each in both hemispheres). This yielded information on the test-retest reliability of the BOLD signal relevant for later DCM analyses.

Furthermore, the test-retest reliability of DCM was addressed by computing the ICC(3,1) for the negative free energy (as an approximation to the log model evidence) and for the model parameter estimates after BMA (Penny et al., 2010), for both DCM implementations. Note that we used the reliability of the negative free energy as a measure of the reliability of random effects BMS. The test-retest reliability of the neuronal parameter estimates of the DCMs (i.e., endogenous connections, modulatory and driving inputs) was assessed by calculating an ICC for each parameter. Notably, as the ICC is particularly meaningful for parameters showing a substantial effect size, we additionally report reliability when restricting the analysis to parameter estimates that deviated strongly from the prior expectations (i.e., with a posterior probability larger than 0.95).

Results

BOLD activity during uni- and bimanual hand movements and its test-retest reliability

We found activation of a widespread cortical network when subjects performed visually synchronized uni- or bimanual hand movements. The activations included bilateral M1, lateral PMC, SMA as well as the motion-sensitive area V5 in the extrastriate cortex (Fig. 2A, Supplementary Fig. S1A-B and Supplementary Table S2). Across sessions, we found the test-retest reliability of BOLD activity in the network (i.e., voxels surviving a threshold of p < 0.05, FWE-corrected for the respective contrast of the first session) to be fair (median ICC; LH: 0.53, RH: 0.53, and BH: 0.54; Supplementary Fig. S1C), with high voxel-wise ICC values within the group-level ROIs selected for subsequent DCM analyses (Fig. 2B and Supplementary Table S3). Assessing the ICC of the median contrast value for each of these ROIs, we found the test-retest reliability to be fair ($ICC_{med} = 0.42$) for the right SMA (Table 1), as well as for the left SMA ($ICC_{med} = 0.51$). For all other ROIs, the test-retest reliability was considerably higher, ranging from good for the left PMC ($ICC_{med} = 0.65$) to excellent for the right V5 ($ICC_{med} = 0.87$). This suggests that BOLD activity within DCM-relevant ROIs was sufficiently stable across sessions to allow for a meaningful examination of test-retest reliability of DCM in further analyses.

Table 1
Test-retest reliability of the median BOLD activity (i.e., contrast value) within the ROIs entered to DCM analyses.

	ICC_{med}	95% CI	*p*
ROI test-retest reliability			
M1_L	0.71	0.49–0.85	1.20e-06
M1_R	0.75	0.56–0.87	1.30e-07
PMC_L	0.65	0.40–0.81	1.48e-05
PMC_R	0.75	0.55–0.87	1.66e-07
SMA_L	0.51	0.21–0.73	9.61e-04
SMA_R	0.42	0.10–0.67	6.30e-03
V5_L	0.76	0.57–0.87	9.51e-08
V5_R	0.87	0.75–0.93	1.80e-11

Bayesian model selection and test-retest reliability of BMS

Two subjects were excluded from DCM analyses, as they did not show consistent activation in all ROIs of the network. For some of the remaining subjects, a few DCMs under DCM10 "flat-lined", i.e., the Variational Laplace algorithm (Friston et al., 2007) converged almost immediately without notable changes in the parameter estimates. Since slight changes to the starting value (not to be confused with the priors) of the optimization scheme resulted in reasonable model fits, flat-lining was presumably due to local extrema close to the default starting value of the algorithm. Here, we addressed this issue in a pragmatic way: For flat-lined DCMs, we redefined the starting value of the algorithm by using the mean posterior parameter estimates across all models that had not flat-lined. Under this adjustment, initially flat-

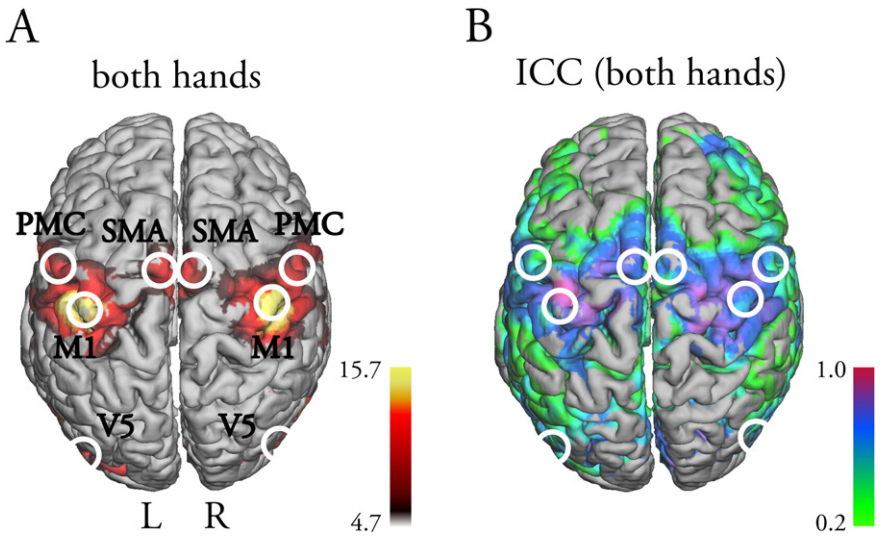

Fig. 2. BOLD activity during bimanual hand movements. (A) Activation pattern shows regions that were activated when subjects performed bimanual hand movements during the first session (L = left hemisphere, R = right hemisphere). White circles indicate regions of interest that were entered into subsequent DCM analyses. The activation pattern is thresholded at a voxel-level threshold of p < 0.05 (FWE-corrected). (B) Cortical reliability map of the voxel-wise ICC(3,1) values for the baseline contrast of bimanual hand movements. ICC maps were calculated using the ICC toolbox extension within SPM5 (Caceres et al., 2009). Results were rendered onto the surface of a standard anatomical template image.

Table 2
Results from random effects Bayesian model selection, as well as their reliability, for classical DCM (cDCM) and for DCM10. Model selection is based on the exceedance probabilities (i.e., ex. prob.). Reliability is given by the intra-class correlation coefficient (ICC) for the negative free energy, on which the random effects BMS procedure is based.

		Model 1	Model 2	Model 3	Model 4
cDCM	ex. prob. (session 1)	0	0	0	1
	ex. prob. (session 2)	0	0	0	1
	ICC (F)	0.94	0.94	0.94	0.94
DCM10	Ex. prob. (session 1)	0.70	0.24	0.01	0.05
	Ex. prob. (session 2)	0.96	0	0.04	0
	ICC (F)	0.60	0.61	0.60	0.60

lining DCMs produced a reasonable fit, as indicated by SPM's routine for post-hoc diagnostics (i.e., the function *spm_dcm_fmri_check*), for each model in every subject (with all fits explaining more than 25% variance).

We then used random effects BMS (as implemented in DCM10) to compare our four alternative models, using the negative free energy as a lower bound approximation to the log model evidence. For cDCM, model 4 (i.e., the most complex model) was the most likely model in both sessions (exceedance probability: 1.00 in either session; see Table 2). For DCM10, however, model 1 (i.e., the model with sparsest modulation) was selected in both sessions as the most likely model (exceedance probability: 0.70 and 0.96 in sessions 1 and 2, respectively).

To address the test-retest reliability of BMS, we calculated an ICC for the negative free energy of each model. For cDCM, the negative free energy showed almost perfect reliability (ICC = 0.94; see Table 2), regardless of the specific model. By contrast, reliability was reduced when using DCM10, although still classified as "good" (with ICCs between 0.60–0.61; see Table 2).

Test-retest reliability of DCM connectivity parameters

Using Bayesian model averaging (BMA), individual connectivity parameter estimates were obtained by averaging over all models within the standard Occam's window. Concerning the stability of results, we found the model parameter estimates to be consistent across the different software versions (Supplementary Tables S4-S11): A statistical comparison of the group-level parameter estimates of both sessions between cDCM and DCM10 indicated that parameter estimates were significantly correlated across software versions. This was the case for all parameter types in DCM, i.e., endogenous connectivity parameters ($n = 76$): $r = 0.31$, $p < 0.01$; modulatory input parameters ($n = 180$): $r = 0.74$, $p < 0.001$; driving input parameters ($n = 12$): $r = 0.78$, $p < 0.01$. This indicates that, despite the differences in model selection results described in the previous section, the parameter estimates are fairly stable across software versions when taking into account model uncertainty. Similarly, our results qualitatively reproduced the main conclusions reported in Grefkes et al. (2008), regardless of the software version. For both cDCM and DCM10, unimanual hand movements positively modulated intra- and interhemispheric connections to the contralateral M1, as well as negatively modulated connections to or among motor regions of the ipsilateral hemisphere (Supplementary Tables S5-S6 and S9-S10). For bimanual hand movements, intra- and interhemispheric connections to M1 were positively modulated (Supplementary Tables S7 and S11). A detailed quantitative analysis of the reproducibility is, however, beyond the scope of this paper and will be reported separately (Frässle et al., in preparation).

In this paper, we focus on the test-retest reliability of the averaged parameter estimates (BMA) by means of the ICC, calculated separately for each endogenous, modulatory and driving input parameter. In a second step, we restricted our reliability analyses to those parameters with a substantial effect size (i.e., large deviations from their prior expectation; see Methods).

cDCM yielded highly consistent model parameter estimates across both sessions at the group level (see Fig. 3A for the modulatory influences of left hand movements as an example). ICCs were within a range from −0.17 (poor) to 0.78 (excellent), with a median ICC of 0.47 (Fig. 3B, black bars; see Supplementary Tables S4-S7 for a listing of the ICC of each parameter estimate). When analyzing each connectivity matrix separately, we found a median ICC of 0.50 for the endogenous connectivity estimates, 0.46 for the modulatory parameter estimates, and 0.24 for the driving input estimates. When restricting the analysis to the parameters that showed a substantial effect size, we found a slight elevation of test-retest reliability (median ICC: 0.48). This was due to an increase in the reliability of the modulatory parameter estimates (median ICC: 0.48), whereas the reliability of parameter estimates of endogenous connectivity and driving inputs remained unchanged.

For DCM10, the test-retest reliability of the model parameter estimates dropped considerably compared to cDCM. Most parameter estimates showed only poor reliability (median ICC: 0.13, range: −0.35 to 0.68; Supplementary Tables S8-S11). In particular, ICCs of parameter estimates of endogenous connectivity (median ICC: 0.22) and modulatory inputs (median ICC: 0.06) were low, whereas driving input parameter estimates were much more reliable (median ICC: 0.53). Restricting the analysis to parameters that deviated substantially from their prior mean, we found an increase in reliability (median ICC = 0.21; endogenous connectivity: 0.26; modulatory inputs: 0.11; driving inputs: 0.53), although it remained poor. Notably, despite the limited reliability of most parameter estimates, the overall group-level connectivity patterns were consistent across sessions, as mostly the same connections reached significance and hardly any sign error occurred (see Supplementary Tables S8-S11).

As suggested by one of our reviewers, we repeated the analyses on the test-retest reliability of BMS and model parameter estimation

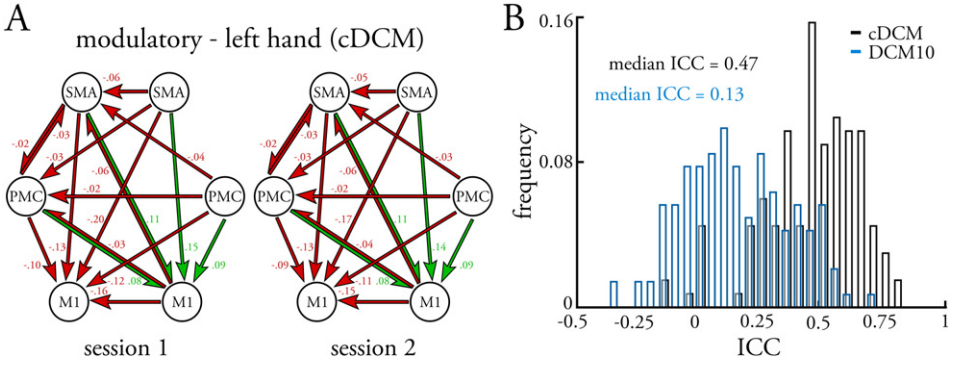

Fig. 3. Reliability of DCM model parameter estimates. (A) Group-level results for the modulation of connectivity (B-matrix) for left hand movements in (left) session 1 and (right) session 2 for cDCM. Connectivity patterns and the strength of the connections are almost identical across sessions. (B) Histogram of the ICC values of all model parameter estimates (i.e., A-, B- and C-matrix) as well as the median ICC for cDCM (black). Additionally, histogram and median of the ICCs are shown for model parameters obtained with DCM10 (blue).

using the most recent implementation of DCM (i.e., DCM12 as implemented in SPM12, version: 6225). Using DCM12, we found virtually the same results as for model inversion under DCM10 (negative free energy: ICCs between 0.56–0.59; parameter estimates (restricted to large effect sizes): median ICC = 0.17; range: −0.34 to 0.84; endogenous: 0.25; modulatory: 0.09; driving inputs: 0.79).

Potential sources of the differences in test-retest reliability

As mentioned above, in order to accommodate a variety of novel DCM variants in one unifying scheme, DCM10 introduced various changes in the implementation of DCM. In this section, we focus on some of the most salient changes and examine whether they explain the observed decrease in reliability from cDCM to DCM10.

Priors

In DCM10, priors were adjusted to accommodate the requirements of novel DCM variants (such as stochastic DCM) such that the prior variance of endogenous and modulatory parameters became much wider than in cDCM, whereas the prior variance of driving input parameters stayed identical (and thus became tighter in relative terms). As described above, we found that these changes in prior variance appeared to map onto the differences in test-retest reliability of the endogenous, modulatory and driving input parameter estimates. For cDCM, reliability was high for parameter estimates of endogenous connections and modulatory inputs, but poor for parameter estimates of driving inputs, whereas the opposite was true for DCM10. This suggested a possible role of the priors for the differences in reliability between cDCM and DCM10.

To examine the effect of the priors more thoroughly, we transferred cDCM's prior expectations and prior variances to DCM10 and re-estimated all models under this setting. While the reliability of negative free energy estimates remained unaffected (ICC = 0.58, for the most likely model 4), the reliability of the model parameter estimates was considerably increased (median ICC: 0.51; range: 0.03 to 0.81). We found high reliability, not only for the parameter estimates of endogenous connections (median ICC: 0.56) and modulatory inputs (median ICC: 0.48), but also for driving inputs (median ICC: 0.66). Restricting the analysis to parameters with a substantial effect size, we found reliability to be further increased, with a median ICC of 0.56 (endogenous: 0.57; modulatory: 0.51; driving inputs: 0.66). In fact, reliability of the model parameter estimates in DCM10 (under the classical priors) was even significantly higher than for cDCM (Mann-Whitney U test: $Z = 2.63$, $p = 0.009$).

One possible explanation for the role of the priors for the drop in reliability across software versions is that the larger prior variance of endogenous connectivity and modulatory input parameters endowed the model with too much flexibility in fitting the data and thus led to overfitting. Such overfitting would naturally lead to poor generalizability and hence low reliability across sessions. This hypothesis can be tested by evaluating the log evidence – a principled measure for the balance between fit and complexity – under both types of priors (cDCM vs. DCM10). However, using random effects BMS at the family-level did not support our hypothesis of overfitting under DCM10 priors. On the contrary, we found that the model family with DCM10 priors was clearly superior (exceedance probability: 1.00 in both sessions).

In a next step, we investigated how the two different priors influenced the values of individual accuracy and complexity terms of the

Table 3
Individual terms of the negative free energy and their test-retest reliability for the models inverted using the original DCM10 priors as well as for the models using the cDCM priors. Individual terms are given by their mean and standard deviation, reliability is given by the ICC. Within each cell of the table, all four models are shown, in the order model 1 to model 4 from the upper to the lower entry.

Components of free energy (log evidence)	DCM10 (priors)			cDCM (priors)		
	Session 1	Session 2	ICC	Session 1	Session 2	ICC
$-\frac{1}{2}\left(e_y C_y^{-1} e_y + log\lvert C_y\rvert\right)$ *Accuracy (log likelihood)*	2006.6 ± 590.9	2094.6 ± 456.1	0.59	1853.8 ± 565.2	1926.3 ± 461.1	0.58
	2005.4 ± 600.9	2079.9 ± 451.1	0.61	1857.7 ± 566.9	1929.3 ± 458.9	0.58
	2004.6 ± 583.4	2094.8 ± 457.0	0.60	1860.4 ± 567.8	1933.4 ± 461.2	0.58
	2005.3 ± 584.4	2079.0 ± 457.2	0.60	1866.1 ± 567.8	1940.4 ± 459.7	0.58
$-\frac{1}{2} e_\theta^T C_\theta^{-1} e_\theta$ *Complexity (parameters)*	−21.1 ± 8.1	−21.4 ± 7.7	0.02	−119.9 ± 37.1	−119.1 ± 41.2	0.59
	−19.9 ± 7.1	−19.9 ± 6.7	0.08	−119.5 ± 35.8	−117.6 ± 34.8	0.58
	−20.9 ± 7.6	−21.0 ± 7.5	0.10	−119.5 ± 36.0	−119.1 ± 38.8	0.61
	−19.9 ± 6.8	−19.4 ± 6.2	0.12	−118.6 ± 35.5	−118.6 ± 37.7	0.58
$-\frac{1}{2} log\lvert C_\theta\rvert$ *Complexity (parameters)*	−816.5	−816.5	-	−625.6	−625.6	-
	−816.5	−816.5	-	−594.7	−594.7	-
	−816.5	−816.5	-	−603.6	−603.6	-
	−816.5	−816.5	-	−550.6	−550.6	-
$\frac{1}{2} log\lvert S_\theta\rvert$ *Conditional uncertainty (parameters)*	−356.4 ± 36.8	−366.5 ± 27.6	0.27	−424.3 ± 21.0	−425.9 ± 17.1	0.59
	−460.0 ± 40.3	−366.7 ± 32.3	0.41	−456.1 ± 20.9	−457.5 ± 16.7	0.59
	−460.2 ± 34.6	−373.0 ± 31.8	0.57	−447.4 ± 21.0	−448.9 ± 17.1	0.58
	−468.4 ± 33.4	−375.1 ± 30.8	0.46	−501.8 ± 21.1	−503.5 ± 17.1	0.59
$-\frac{1}{2} e_\lambda^T C_\lambda^{-1} e_\lambda$ *Complexity (hyper-parameters)*	−22.2 ± 7.6	−23.1 ± 6.0	0.62	−235.8 ± 45.5	−240.1 ± 18.4	0.56
	−22.2 ± 7.8	−22.9 ± 5.9	0.63	−243.0 ± 22.6	−240.0 ± 18.3	0.57
	−22.1 ± 7.6	−23.1 ± 6.1	0.62	−242.9 ± 22.6	−239.8 ± 18.4	0.57
	−22.1 ± 7.6	−22.9 ± 6.0	0.63	−242.6 ± 22.6	−239.5 ± 18.3	0.57
$-\frac{1}{2} log\lvert C_\lambda\rvert$ *Complexity (hyper-parameters)*	0	0	-	0	0	-
	0	0	-	0	0	-
	0	0	-	0	0	-
	0	0	-	0	0	-
$\frac{1}{2} log\lvert S_\lambda\rvert$ *Conditional uncertainty (hyper-parameters)*	−21.0 ± 3e-12	−21.0 ± 4e-12	0.52	−21.0 ± 3e-12	−21.0 ± 5e-12	0.56
	−21.0 ± 3e-12	−21.0 ± 4e-12	0.47	−21.0 ± 3e-12	−21.0 ± 5e-12	0.56
	−21.0 ± 3e-12	−21.0 ± 4e-12	0.53	−21.0 ± 3e-12	−21.0 ± 5e-12	0.55
	−21.0 ± 3e-12	−21.0 ± 4e-12	0.51	−21.0 ± 3e-12	−21.0 ± 5e-12	0.55

negative free energy from Eq. (2), as well as their reliability. Models inverted using DCM10 priors showed considerably higher accuracy (i.e., first row in Table 3) and lower complexity (i.e., sum of remaining terms in Table 3); as was also indicated by the BMS results described above. Remarkably, almost all terms showed comparable reliability between DCM10 and cDCM, the only exception being the (squared and precision-weighted) deviation of the posterior mean from the prior mean (second row in Table 3). This is consistent with the above findings that the reliability of posterior parameter estimates decreased far more markedly in DCM10 than the reliability of negative free energy estimates.

One possible reason why the wider priors in DCM10 impacted negatively on reliability is that they afforded less regularization than cDCM priors, thus increasing the possibility of local extrema in the objective function and hence more variable estimates across sessions. We return to this possibility below.

Hyperpriors

Although high reliability of the parameter estimates could be recovered when transferring the original cDCM prior distributions to DCM10, the initially observed problem with reliability in DCM10 is not necessarily attributable to the priors as such. In other words, the larger prior variance for endogenous connections and modulatory inputs in DCM10 might only provide a fundament for the influence of additional factors, which impact differently on model inversion in both software versions, such as the choice of hyperpriors, or use of highly correlated inputs.

A hyperprior h describes prior expectations about measurement error e, e.g., $e \sim N(0, \exp(h)^{-1})$, $h \sim N(hE, hC)$; this essentially encodes prior beliefs about the signal-to-noise ratio (SNR) of the measured data. Hyperpriors are often overlooked aspects of model specification, but can have a profound effect on model inversion and comparison. Heuristically, the prior expectation of log precision (hE) corresponds to the expected SNR (under the simplifying assumption that the variance of the signal is 1). This means that by specifying the log precision, we can tell the model the (a priori) level of signal-to-noise that is typical of the fMRI time series at hand. Increasing the log precision means the inversion will strive to provide an accurate explanation and allow for larger deviations of posterior estimates from prior values, possibly at the price of overfitting. Generally speaking, increasing the log precision may reduce the model evidence but increase sensitivity to differences in log evidence among models. We therefore tested whether the drop in reliability could partially reflect a more "brittle" evaluation of model evidence, given the wider priors in DCM10. To that end, we repeated the analyses under DCM10 with three values of the expected log precision (hE = 2, 4 and 6). This corresponds roughly to SNRs of 7, 50, and 400, respectively. These may seem rather large values but the regional summaries used in DCM for fMRI are based upon regional "averages" (principal eigenvariates) that suppress noise in proportion to the volume of the region.

Re-running model inversion in DCM10 with different values of the expected log precision did not reveal any major effect of the hyperpriors on reliability, neither for BMS (Fig. 4A) nor for the averaged (BMA) model parameter estimates. Fig. 4B shows the results for taking into account only model parameters showing a substantial effect size. In summary, regardless of the chosen value for the expected log precision, reliability remained reduced in DCM10 as compared to cDCM.

Correlated inputs

So far, DCMs were defined as in Grefkes et al. (2008), using three fairly strongly correlated inputs (for left hand movements, right hand movements, and bimanual movements). Additionally, the same inputs were used as driving inputs to V5 and as modulatory influences on the coupling parameters. Such design choices can be expected to induce non-negligible conditional dependencies amongst model parameters; this, in turn, can lead to more variable parameter estimates across sessions (cf. Rowe et al., 2010). While there was no systematic difference, across models, between cDCM and DCM10 with regard to posterior dependencies (fourth row in Table 3), we examined whether a redefinition of inputs, leading to less correlation amongst parameters, would reduce differences between cDCM and DCM10.

In the redefined DCMs, instead of using one input for each hand-movement condition, one input implemented any visual stimulation, one input coded the difference between right and left hand movements, and a third input represented bimanual hand movements. For these re-parameterized DCMs, we again assessed the test-retest reliability of both the negative free energy and averaged (BMA) model parameter estimates in DCM10.

As expected, conditional dependencies amongst parameter estimates were reduced in the redefined models. Contrary to our expectation, however, reliability did not increase noticeably in DCM10, neither for the negative free energy (ICC = 0.59, for optimal model 1) nor for the model parameter estimates (median ICC = 0.12; range: −0.44 to 0.63; endogenous: 0.19; modulatory: 0.08; driving inputs: 0.55). Restricting the reliability analysis only to the parameters with large effect sizes, reliability was increased, although remaining poor (median ICC =

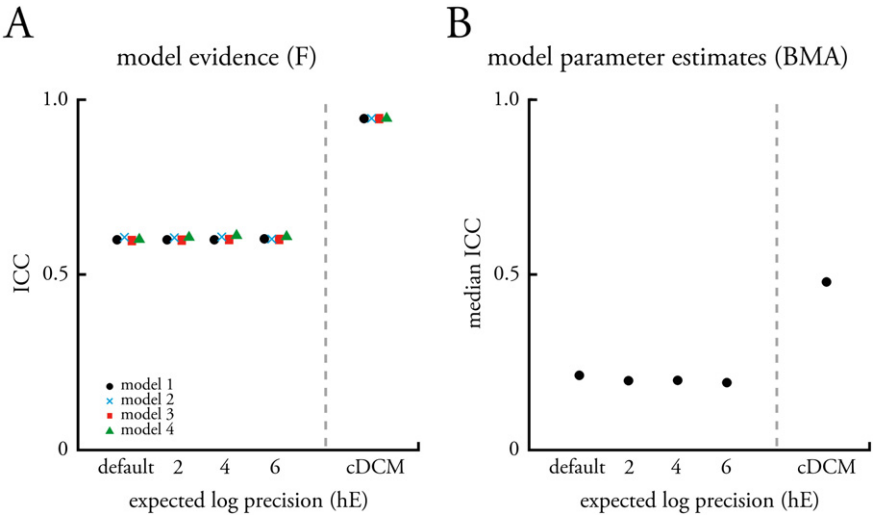

Fig. 4. Influence of the expected log precision on the test-retest reliability for (A) the negative free energy of each model (black dot = model 1, blue cross = model 2, red square = model 3, green triangle = model 4), and (B) the model parameter estimates (shown is the median ICC averaged over all parameters). Only parameters showing a substantial effect size were considered for the estimation of test-retest reliability. Reliabilities are shown for DCM10's default expected log precisions, for hE = 2, 4 and 6 in DCM10, as well as for the models inverted in cDCM.

0.15; endogenous: 0.19; modulatory: 0.09; driving inputs: 0.55). Finally, we varied the expected log precision (hE = 2, 4 and 6); as for the original input structure described above, this did not substantially affect reliability of model selection (Supplementary Fig. S2A) or model parameter estimation (Supplementary Fig. S2B).

Discussion

This study reports the results of a software comparison with respect to the test-retest reliability of DCM. Specifically, we examined the reliability of model selection and parameter estimation across two sessions of an established motor paradigm in healthy volunteers, using two different software versions, a classical (cDCM, SPM5) and a more recent (DCM10, SPM8) implementation of DCM.

Our results suggest that the reliability of DCM depends on the software version used for model inversion. For cDCM, reliability was excellent for model selection by BMS and satisfactory for model parameter estimates, with most of them showing fair or good reliability. This suggests that DCM can be a reliable tool for inferring effective connectivity from fMRI data, as previously shown by Schuyler et al. (2010) for auditory and visual tasks.

Using DCM10 (and similarly DCM12), we found a reduction in the reliability of both BMS and model parameter estimation. Whereas BMS results still exhibited good reliability, the majority of model parameter estimates had poor reliability. This is an important issue for investigating effective connectivity on a single-subject level (e.g., in clinical applications). We therefore tried to identify the reasons for the observed decrease of reliability in DCM10. Software changes in DCM10 were mostly motivated by the goal to integrate numerous DCM variants – such as nonlinear, two-state, and particularly stochastic DCM – under one implementation framework. Prominent modifications included numerical changes in the inversion scheme, simplifications of the hemodynamic model, reparameterization of self-connections, and adjustments of prior distributions. Specifically, prior variances were changed to allow for greater impact of connections, relative to driving inputs, on network dynamics; this was particularly motivated from the perspective of resting-state analyses under DCM.

Based on our observations on the relative reliability of endogenous, modulatory and input parameters across both software versions, we hypothesized that the change in prior distributions might be a key factor for explaining the observed changes in reliability across DCM versions. This was confirmed by transferring the priors of cDCM to DCM10: under these classical (tight shrinkage) priors, model inversion with DCM10 led to substantially increased reliability of the parameter estimates, even beyond the levels observed with cDCM.

Subsequent analyses tried to clarify why the choice of priors had such a marked influence on reliability. A straightforward explanation is that reliability might decrease due to overfitting, given that priors have higher variance in DCM10. However, this hypothesis was refuted by examining the log model evidence under both types of priors: models with DCM10 priors clearly outperformed models with cDCM priors.

We then examined the reliability of each individual term in the objective function of DCM, the negative free energy, which represents a bound approximation to the log evidence and contains components encoding model fit and complexity, respectively. We identified a single term whose reliability differed markedly between cDCM and DCM10. This was the (squared and precision-weighted) difference between posterior and prior mean. This finding is perfectly in line with the observation that reliability of BMS results decreased only moderately in DCM10, compared to cDCM, whereas the reliability of posterior mean estimates dropped sharply. By contrast, there was no evidence for a more general numerical problem in DCM10 (such as computing log determinants of covariance matrices) that could have affected multiple parts of the objective function.

We also examined the role of additional factors, which might have interacted with the choice of priors, such as the specification of hyperpriors (i.e., assumptions about SNR) and correlations in inputs encoding experimental conditions. These analyses, however, did not support the possibility that these factors could represent additional reasons for the drop in reliability under DCM10.

We do not wish to imply that our results question previous group-level results obtained with DCM10. This is for several reasons. First, we examined an unusually challenging case here, with complex models (with up to >100 parameters) fitted to relatively few data points (i.e., 380 scans per subject and session). Most previous empirical applications of DCM deal with a much more graceful ratio of data points to parameter numbers. Second, almost all DCM studies to date report group-level inferences based on summary statistics approaches (e.g., null hypothesis tests applied to maximum a posteriori parameter estimates); this approach depends on between-subject variability. This means that studies reporting significant effects at the between-subject level have probably discovered large effect sizes, because the between-subject random effects are effectively revealed by less informative priors (rendering inference more conservative). Third, our current analyses demonstrate that despite the low test-retest reliability of model parameter estimation under DCM10, parameter estimates at the group-level were consistent across DCM versions. This was demonstrated by statistical analyses which showed significantly correlated parameter estimates across software versions. This correspondence was particularly significant for the estimates of modulatory inputs (B-matrix); these are the estimates of main interest in most DCM studies as they encode context-dependent changes in connectivity.

Two previous studies have investigated the test-retest reliability of DCM for fMRI in other contexts. Using classical DCM implemented in SPM5, Schuyler et al. (2010) reported fair to excellent reliability of model parameter estimates during auditory and visual tasks, but using consecutive within-subject sessions, smaller models and not addressing the reliability of model selection. Also using cDCM in SPM5 for model inversion, Rowe et al. (2010) reported high reliability of BMS for an action selection paradigm, but poor correlations of model parameter estimates across sessions. Several reasons for this finding were suggested by Rowe et al. (2010), including high posterior covariances among model parameters, model complexity and long inter-session intervals. In contradistinction to Rowe et al. (2010), our findings show that both model selection and model parameter estimation can be reliable, even when separating sessions by one month and when using complex models with non-trivial posterior covariances. An alternative explanation for the results by Rowe et al. (2010) is the relatively low number of data points (156 scans per session). This might have rendered model inversion brittle and thus reduced the reliability of model parameter estimates. Having said this, low number of data points might be a critical issue in the present study as well. Although twice as many volumes (i.e., 380 scans) were acquired, we also used a model with twice as many regions (eight as opposed to four) and more than twice the number of parameters.

In fact, brittleness and problems of model inversion can be a potential issue for fitting any nonlinear model, particularly when using a small number of data points relative to the number of model parameters. In analyses under parametric (normal) assumptions, a sufficiently high ratio of data points and model parameters is required to ensure that the posterior distribution is well approximated by a Gaussian; in DCM, this is an important prerequisite for the stability of the Variational Laplace (VL) scheme (cf. Daunizeau et al., 2011).

One of our reviewers asked for diagnostics of when DCM results could be “trusted”. Generally, when fitting nonlinear models (not just DCMs), simple binary criteria or clear-cut thresholds that indicate the absence or presence of problems with inversion or identifiability rarely exist. Any such thresholds are essentially arbitrary, as in other domains of statistics (cf. significance thresholds in frequentist statistics), and cannot replace a thorough understanding of both the model and the inversion scheme. For example, the former can enable a reparameterization to avoid identifiability issues (see above and Brodersen et al., 2008 for an example of fitting behavioral data); the

latter helps one to detect pathological updates in the sequence of objective function values returned by an optimization algorithm, such as the VL scheme in DCM.

Having said this, there are some general heuristics which apply to the inversion of any generative model, not just DCM, and can help detecting problems. For example, a common rule of thumb is that a model requires ten data points for each free parameter (Penny and Roberts, 1999). Furthermore, it is informative to inspect the posterior covariance matrix (which DCM provides by default). High posterior variances are related to conventional indices of sensitivity analyses (Deneux and Faugeras, 2006) and signal interdependencies among parameters. This is most easily interpreted when transforming posterior covariances into posterior correlations (cf. Stephan et al., 2007b); the finite range of the latter facilitates detecting potential identifiability problems of specific parameters. High posterior correlations may reflect a redundant parameterization or badly behaved regimes of the objective function (e.g., ridges or ravines) that pose a problem for many model inversion schemes.

Importantly, however, potential problems with identifiability are taken into account automatically by Bayesian model comparison. This is because the posterior covariance among parameters increases model complexity (cf. Daunizeau et al., 2011); in other words, in model comparisons, models with identifiability problems will be judged as inferior and will contribute little to Bayesian model averages. Finally, when problems are suspected, it is possible (albeit laborious) to examine the identifiability of parameters in simulations using realistic signal-to-noise ratios (which can be estimated, for example, from a conventional GLM). This approach has been used by previous empirical DCM studies (see, for example, Stephan et al., 2007a), and DCM offers tools to run such simulations (e.g., see the function *spm_dcm_create*).

In summary, our results demonstrate that the choice of priors can have substantial influence on the test-retest reliability of DCM results. Notably, the influence of the priors could not be attributed to overfitting and did not affect the reliability of any other objective function component than the posterior mean estimates. The most likely explanation for this finding is that the large prior variance of the endogenous and modulatory parameters in DCM10 exerted less regularization of the objective function than the tighter priors in cDCM, thus leading to more local extrema in the objective function (or greater likelihood of bifurcations induced by parameter updates; cf. Daunizeau et al., 2011). Low reliability might therefore be attributable to the Variational Laplace algorithm being trapped in different local extrema in each session. This hypothesis is supported by our observations that some DCMs converged almost immediately ("flat-lined") when inverted under DCM10, while producing sensible results as soon as the optimization algorithm's starting position was slightly varied. To test this hypothesis systematically, we will use global optimization schemes in a future study, such as Markov Chain Monte Carlo (MCMC) and Gaussian Process Optimization (GP) schemes, which are currently under development for DCM for fMRI.

Local extrema in the objective function might also provide an explanation for the observed influence of the prior distributions on the selection of the winning model. Under the DCM10 priors, the sparsest model (model 1) was most likely, whereas the full model (model 4) was selected under cDCM priors. Our analyses of the individual terms of the objective function of DCM illustrate that model 1 had the highest accuracy and lowest complexity of all models under DCM10 priors. For cDCM priors, model 1 still had the lowest complexity but at the same time provided only poor accuracy. In contrast, the winning model (model 4) provided the largest complexity, yet this was outweighed by the benefit in accuracy. As for the differences in parameter estimates, these differences in model selection might result from a higher incidence of local extrema under the wide DCM10 priors, compared to the tighter regularization provided by cDCM priors; the latter may have enabled cDCM to exploit the explanatory power of additional parameters without getting stuck during optimization.

Three important issues remain to be emphasized. First, to ensure a clear focus, the present report was restricted to an assessment of test-retest reliability. By contrast, a separate quantitative analysis concerning the reproducibility of parameter estimates (with regard to the previous group results by Grefkes et al., 2008) will be reported in a separate paper (Frässle et al., in preparation). In brief, we found that our group results reproduce those by Grefkes et al. (2008) well, regardless of the DCM implementation used for model inversion.

Second, we have focused exclusively on bilinear deterministic DCM for fMRI and did not consider other DCM variants (e.g., nlDCM or sDCM). These variants differ from bilinear DCM and their use is not easily motivated for the simple task and motor network in this paper. Furthermore, we could not extend our comparison to these variants because they were not implemented in SPM5. Thorough analyses of the test-retest reliability of nlDCM and sDCM thus remain subject to future research. Similarly, our findings are – strictly speaking – only applicable to the hand-movement paradigm used in the present study. Repeating the test-retest procedures we have introduced, with different experimental paradigms, may further elucidate the impact of prior assumptions on test-retest reliability across software versions.

Third, one might argue that identifying the choice of prior variance as key influence on reliability is somewhat non-surprising, in the sense that, all other factors being equal, reliability can be expected to increase with decreasing prior variance. Indeed, in the limit of a delta function prior, perfect reliability is guaranteed. Having said this, cDCM priors do allow for non-trivial deviations of posterior from prior mean and have proven practical suitability for fitting empirical fMRI data in many studies. Furthermore, our results using the ICC of posterior means that deviate from the prior means cannot be explained by a non-specific effect of increasing shrinkage priors. Therefore, our present findings suggest that the choice of priors should represent a future focus of DCM developments. An attractive approach in this regard is Empirical Bayes (EB), which constitutes a powerful Bayesian inference scheme for hierarchical models (Efron and Morris, 1973; Kass and Steffey, 1989) and has been introduced to other neuroimaging analyses in the past (e.g., posterior probability maps; Friston and Penny, 2003). For standard DCM group-analyses, EB would estimate prior distributions from the data across subjects, under the hierarchical structure of a multi-subject random or mixed-effects model. A prototype of such a scheme for DCM analyses is presently being developed in our group. Our expectation here is that the empirical priors from the second (between-subject) level will be sufficiently informative to shrink subject-specific estimates and therefore improve reliability in the same way that we have shown when shrinking the priors in the DCM10 analyses of single subjects. We will explore the utility of this approach, and its benefits for reliability, in forthcoming studies.

Acknowledgements

This work has been founded by the Research Foundation of the University of Marburg (PhD scholarship) and the German Academic Exchange Service (to S.F.), the René and Susanne Braginsky Foundation (to K.E.S), the German Research Foundation (DFG; KR3803/2-1, KR3803/7-1) and the START-programme of the RWTH Aachen University (to S.K. and F.M.P.) and the Else Kröner-Fresenius Stiftung (grant 2012_A219, to A.J.).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2015.05.040.

References

Almeida, J.R., Versace, A., Mechelli, A., Hassel, S., Quevedo, K., Kupfer, D.J., Phillips, M.L., 2009. Abnormal amygdala-prefrontal effective connectivity to happy faces differentiates bipolar from major depression. Biol. Psychiatry 66, 451–459.

Aron, A.R., Gluck, M.A., Poldrack, R.A., 2006. Long-term test-retest reliability of functional MRI in a classification learning task. NeuroImage 29, 1000–1006.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26, 839–851.
Birn, R.M., Cornejo, M.D., Molloy, E.K., Patriat, R., Meier, T.B., Kirk, G.R., Nair, V.A., Meyerand, M.E., Prabhakaran, V., 2014. The influence of physiological noise correction on test-retet reliability of resting-state functional connectivity. Brain Connect. 4, 511–522.
Brandt, D.J., Sommer, J., Krach, S., Bedenbender, J., Kircher, T., Paulus, F.M., Jansen, A., 2013. Test-retest reliability of fMRI brain activity during memory encoding. Front. Psychol. 4, 163.
Braun, U., Plichta, M.M., Esslinger, C., Sauer, C., Haddad, L., Grimm, O., Mier, D., Mohnke, S., Heinz, A., Erk, S., Walter, H., Seiferth, N., Kirsch, P., Meyer-Lindenberg, A., 2012. Test-retest reliability of resting-state connectivity network characteristic using fMRI and graph theoretical measures. NeuroImage 59, 1404–1412.
Brodersen, K., Penny, W., Harrison, L., Daunizeau, J., Ruff, C., Duzel, E., Friston, K., Stephan, K., 2008. Integrated Bayesian models of learning and decision making for saccadic eye movements. Neural Netw. 21, 1247–1260.
Brodersen, K., Schofield, T., Leff, A., Ong, C., Lomakina, E., Buhmann, J., Stephan, K., 2011. Generative embedding for model-based classification of fMRI data. PLoS Comput. Biol. 7.
Brodersen, K.H., Deserno, L., Schlagenhauf, F., Lin, Z., Penny, W.D., Buhmann, J.M., Stephan, K.E., 2014. Dissecting psychiatric spectrum disorders by generative embedding. NeuroImage Clin. 4, 98–111.
Caceres, A., Hall, D., Zelaya, F., Williams, S., Mehta, M., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. NeuroImage 45, 758–768.
Cicchetti, D., 2001. The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. J. Clin. Exp. Neuropsychol. 23, 695–700.
Daunizeau, J., Friston, K., Kiebel, S., 2009. Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. Physica D Nonlinear Phenomena 238, 2089–2118.
Daunizeau, J., David, O., Stephan, K.E., 2011. Dynamic causal modelling: A critical review of the biophysical and statistical foundations. NeuroImage 58, 312–322.
David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., Depaulis, A., 2008. Identifying neural drivers with functional MRI: an electrophysiological validation. PLoS Biol. 6, 2683–2697.
Deneux, T., Faugeras, O., 2006. Using nonlinear models in fMRI data analysis: Model selection and activation detection. NeuroImage 32, 1669–1689.
Deserno, K., Sterzer, P., Wüstenberg, T., Heinz, A., Schlagenhauf, F., 2012. Reduced prefrontal-parietal effective connectivity and working memory deficits in schizophrenia. J. Neurosci. 32, 12–20.
Di, X., Biswal, B.B., 2014. Modulatory interactions between the default mode network and task positive networks in resting-state. PeerJ 2, e367.
Dima, D., Roiser, J.P., Dietrich, D.E., Bonnemann, C., Lanfermann, H., Emrich, H.M., Dillo, W., 2009. Understanding why patients with schizophrenia do not perceive the hollow-mask illusion using dynamic causal modelling. NeuroImage 46, 1180–1186.
Dima, D., Dietrich, D.E., Dillo, W., Emrich, H.M., 2010. Impaired top-down processes in schizophrenia: a DCM study of ERPs. NeuroImage 52, 824–832.
Dima, D., Jogia, J., Frangou, S., 2014. Dynamic causal modeling of load-dependent modulation of effective connectivity within the verbal working memory network. Hum. Brain Mapp. 35, 3025–3035.
Efron, B., Morris, C., 1973. Stein's estimation rule and its competitors – Empirical Bayes Approach. J. Am. Stat. Assoc. 68, 117–130.
Eickhoff, S., Stephan, K., Mohlberg, H., Grefkes, C., Fink, G., Amunts, K., Zilles, K., 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. NeuroImage 25, 1325–1335.
Fairhall, S.L., Ishai, A., 2007. Effective connectivity within the distributed cortical network for face perception. Cereb. Cortex 17, 2400–2406.
Fliessbach, K., Rohe, T., Linder, N., Trautner, P., Elger, C., Weber, B., 2010. Retest reliability of reward-related BOLD signals. NeuroImage 50, 1168–1176.
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., 2008. Test-retest and between-site reliability in a multicenter fMRI study. Hum. Brain Mapp. 29, 958–972.
Friston, K.J., Penny, W., 2003. Posterior probability maps and SPMs. NeuroImage 19, 1240–1249.
Friston, K., Holmes, A., Poline, J., Grasby, P., Williams, S., Frackowiak, R., Turner, R., 1995. Analysis of fMRI time-series revisited. NeuroImage 2, 45–53.
Friston, K., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273–1302.
Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. NeuroImage 34, 220–234.
Friston, K., Trujillo-Barreto, N., Daunizeau, J., 2008. DEM: A variational treatment of dynamic systems. NeuroImage 41, 849–885.
Friston, K., Stephan, K., Li, B., Daunizeau, J., 2010. Generalised filtering. Math. Probl. Eng.
Friston, K., Kahan, J., Biswal, B., Razi, A., 2014. A DCM for resting state fMRI. NeuroImage 94, 396–407.
Goulden, N., Khusnulina, A., Davis, N.J., Bracewell, R.M., Bokde, A.L., McNulty, J.P., Mullins, P.G., 2014. The salience network is responsible for switching between the default mode network and the central execution network: replication from DCM. NeuroImage 99, 180–190.
Grefkes, C., Eickhoff, S., Nowak, D., Dafotakis, M., Fink, G., 2008. Dynamic intra- and interhemispheric interactions during unilateral and bilateral hand movements assessed with fMRI and DCM. NeuroImage 41, 1382–1394.
Grefkes, C., Nowak, D., Wang, L., Dafotakis, M., Eickhoff, S., Fink, G., 2010. Modulating cortical connectivity in stroke patients by rTMS assessed with fMRI and dynamic causal modeling. NeuroImage 50, 233–242.
Grèzes, J., Wicker, B., Berthoz, S., de Gelder, B., 2009. A failure to grasp the affective meaning of actions in autism spectrum disorder subjects. Neuropsychologia 47, 1816–1825.
Grol, M., Majdandzic, J., Stephan, K., Verhagen, L., Dijkerman, H., Bekkering, H., Verstraten, F., Toni, I., 2007. Parieto-frontal connectivity during visually guided grasping. J. Neurosci. 27, 11877–11887.
Kass, R., Steffey, D., 1989. Aproximate Bayesian inference in conditionally indepedent hierarchical models (parametric empirical Bayes models). J. Am. Stat. Assoc. 84, 717–726.
Kellermann, T., Regenbogen, C., De Vos, M., Mößnang, C., Finkelmeyer, A., Habel, U., 2012. Effective connectivity of the human cerebellum during visual attention. J. Neurosci. 32, 11453–11460.
Lee, L., Friston, K., Horwitz, B., 2006. Large-scale neural models and dynamic causal modelling. NeuroImage 30, 1243–1254.
Li, J., Liu, J., Liang, J., Zhang, H., Zhao, J., Rieth, C.A., Huber, D.E., Li, W., Shi, G., Ai, L., Tian, J., Lee, K., 2010. Effective connectivities of cortical regions for top-down face processing: a dynamic causal modeling study. Brain Res. 1340, 40–51.
Loubinoux, I., Carel, C., Alary, F., Boulanouar, K., Viallard, G., Manelfe, C., Rascol, O., Celsis, P., Chollet, F., 2001. Within-session and between-session reproducibility of cerebral sensorimotor activation: a test–retest effect evidenced with functional magnetic resonance imaging. J. Cereb. Blood Flow Metab. 21, 592–607.
Marreiros, A., Kiebel, S., Friston, K., 2008. Dynamic causal modelling for fMRI: A two-state model. NeuroImage 39, 269–278.
Nguyen, V.T., Breakspear, M., Cunnington, R., 2014. Fusing concurrent EEG-fMRI with dynamic causal modeling: Application to effective connectivity during face perception. NeuroImage 102, 60–70.
Oldfield, R.C., 1971. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9, 97–113.
Penny, W., Roberts, S., 1999. Bayesian neural network for classification: how useful is the evidence framework? Neural Netw. 12, 877–892.
Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Modelling functional integration: a comparison of structural equation and dynamic causal models. NeuroImage 23 (Suppl. 1), S264–S274.
Penny, W., Stephan, K., Daunizeau, J., Rosa, M., Friston, K., Schofield, T., Leff, A., 2010. Comparing families of dynamic causal models. PLoS Comput. Biol. 6.
Plichta, M.M., Schwarz, A.J., Grimm, O., Morgen, K., Mier, D., Haddad, L., Gerdes, A.B., Sauer, C., Tost, H., Esslinger, C., Colman, P., Wilson, F., Kirsch, P., Meyer-Lindenberg, A., 2012. Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. NeuroImage 60, 1746–1758.
Radulescu, E., Minati, L., Ganeshan, B., Harrison, N.A., Gray, M.A., Beacher, F.D., Chatwin, C., Young, R.C., Critchley, H.D., 2013. Abnormalities in fronto-striatal connectivity within language networks relate to differences in grey-matter heterogeneity in Asperger syndrome. NeuroImage Clin. 2, 716–726.
Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J., Kahn, R.S., Ramsey, N.F., 2007. Test-retest reliability of fMRI activation during prosaccades and antisaccades. NeuroImage 36, 532–542.
Reyt, S., Picq, C., Sinniger, V., Clarencon, D., Bonaz, B., David, O., 2010. Dynamic causal modelling and physiological confounds: A functional MRI study of vagus nerve stimulation. NeuroImage 52, 1456–1464.
Rizzolatti, G., Luppino, G., 2001. The cortical motor system. Neuron 31, 889–901.
Roiser, J.P., Wigton, R., Kilner, J.M., Mendez, M.A., Hon, N., Friston, K.J., Joyce, E.M., 2013. Dysconnectivity in the frontoparietal attention network in schizophrenia. Front. Psychol. 4, 176.
Rowe, J., Hughes, L., Barker, R., Owen, A., 2010. Dynamic causal modelling of effective connectivity from fMRI: Are results reproducible and sensitive to Parkinson's disease and its treatment? NeuroImage 52, 1015–1026.
Schlösser, R.G., Wagner, G., Koch, K., Dahnke, R., Reichenbach, J.R., Sauer, H., 2008. Fronto-cingulate effective connectivity in major depression: a study with fMRI and dynamic causal modeling. NeuroImage 43, 645–655.
Schuyler, B., Ollinger, J., Oakes, T., Johnstone, T., Davidson, R., 2010. Dynamic causal modeling applied to fMRI data shows high reliability. NeuroImage 49, 603–611.
Shrout, P., Fleiss, J., 1979. Intraclass correlations - Uses in assessing rater reliability. Psychol. Bull. 86, 420–428.
Siman-Tov, T., Mendelsohn, A., Schonberg, T., Avidan, G., Podlipsky, I., Pessoa, L., Gadoth, N., Ungerleider, L., Hendler, T., 2007. Bihemispheric leftward bias in a Visuospatial attention-related network. J. Neurosci. 27, 11271–11278.
Smith, S., 2012. The future of fMRI connectivity. NeuroImage 62, 1257–1266.
Stephan, K.E., Marshall, J., Penny, W., Friston, K., Fink, G., 2007a. Interhemispheric integration of visual processing during task-driven lateralization. J. Neurosci. 27, 3512–3522.
Stephan, K.E., Weiskopf, N., Drysdale, P.M., Robinson, P.A., Friston, K.J., 2007b. Comparing hemodynamic models with DCM. NeuroImage 38, 387–401.
Stephan, K.E., Kasper, L., Harrison, L., Daunizeau, J., den Ouden, H., Breakspear, M., Friston, K., 2008. Nonlinear dynamic causal models for fMRI. NeuroImage 42, 649–662.
Stephan, K.E., Penny, W., Daunizeau, J., Moran, R., Friston, K., 2009. Bayesian model selection for group studies. NeuroImage 46, 1004–1017.
Summerfield, C., Koechlin, E., 2008. A neural representation of prior information during perceptual inference. Neuron 59, 336–347.
Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., Hirsch, J., 2006. Predictive codes for forthcoming perception in the frontal cortex. Science 314, 1311–1314.
Valdes-Sosa, P.A., Roebroeck, A., Daunizeau, J., Friston, K., 2011. Effective connectivity: influence, causality and biophysical modeling. NeuroImage 58, 339–361.
Worsley, K., Friston, K., 1995. Analysis of fMRI time-series revisited – again. NeuroImage 2, 173–181.
Zeki, S., Watson, J., Lueck, C., Friston, K., Kennard, C., Frackowiak, R., 1991. A direct demonstration of functional specialization in human visual-cortex. J. Neurosci. 11, 641–649.