Quantifying the performance of MEG source reconstruction using resting state data

Simon Little [a,*], James Bonaiuto [a,b], Sofie S. Meyer [c,d,e], Jose Lopez [f], Sven Bestmann [a,c,1], Gareth Barnes [c,1]

[a] Department of Clinical and Movement Neurosciences, UCL Institute of Neurology, Queen Square, London, UK
[b] Centre de Neuroscience Cognitive, CNRS UMR 5229-Université Claude Bernard Lyon I, 69675, Bron Cedex, France
[c] Wellcome Centre for Human Neuroimaging, UCL Institute of Neurology, 12 Queen Square, London, UK
[d] Institute of Cognitive Neuroscience, University College London, London, WC1N 3AR, UK
[e] Institute of Neurology, University College London, London, WC1N 1PJ, UK
[f] Electronic Engineering Department, Universidad de Antioquia, UdeA, Calle 70 No. 52-21, Medellín, Colombia

A B S T R A C T

In magnetoencephalography (MEG) research there are a variety of inversion methods to transform sensor data into estimates of brain activity. Each new inversion scheme is generally justified against a specific simulated or task scenario. The choice of this scenario will however have a large impact on how well the scheme performs. We describe a method with minimal selection bias to quantify algorithm performance using human resting state data. These recordings provide a generic, heterogeneous, and plentiful functional substrate against which to test different MEG recording and reconstruction approaches. We used a Hidden Markov model to spatio-temporally partition data into self-similar dynamic states. To test the anatomical precision that could be achieved, we then inverted these data onto libraries of systematically distorted subject-specific cortical meshes and compared the quality of the fit using cross validation and a Free energy metric. This revealed which inversion scheme was able to identify the least distorted (most accurate) anatomical models, and allowed us to quantify an upper bound on the mean anatomical distortion accordingly. We used two resting state datasets, one recorded with head-casts and one without. In the head-cast data, the Empirical Bayesian Beamformer (EBB) algorithm showed the best mean anatomical discrimination (3.7 mm) compared with Minimum Norm/LORETA (6.0 mm) and Multiple Sparse Priors (9.4 mm). This pattern was replicated in the second (conventional dataset) although with a marginally poorer (non-significant) prediction of the missing (cross-validated) data. Our findings suggest that the abundant resting state data now commonly available could be used to refine and validate MEG source reconstruction methods and/or recording paradigms.

1. Introduction

Magnetoencephalography (MEG) detects electromagnetic fields at sensors outside the head. The challenge for the researcher is to infer the neuronal current distribution responsible for the observed data, despite a much higher number of possible sources than sensors. The general approach is to restrict the number of potential solutions through a priori assumptions, including the temporal relationship between sources (i.e. source co-variance) and/or the anatomical manifold that gives rise to this function (e.g. the cortical mesh). These assumptions are continually being refined and debated (Baillet, 2015; Baillet et al., 2001; Lin et al.,

2006; Wipf and Nagarajan, 2009). Two recurring issues make it difficult for the community to come to a consensus on optimal source reconstruction methods - the first is the choice of test scenario, the second is the lack of ground truth.

Firstly, the choice of task, or simulation set-up used to compare source localisations will introduce a selection bias towards a specific temporal pattern predominant in certain cortical areas, which will suit some inversion assumptions but not others. Here we set out a framework which utilises diverse spatio-temporal patterns and minimizes selection bias by using a Hidden Markov model (Baker et al., 2014) to parcel endogenous resting-state data into collections of self-similar and quasi-stationary time

---

segments. Resting state data have been shown to arise from dynamic spatio-temporal network state fluctuations occurring on the scale of 100–200 ms (Baker et al., 2014; Koenig et al., 2002; Wackermann et al., 1993; Woolrich et al., 2013) that include the rehearsal of the transient dynamic patterns observed during task performance (O'Neill et al., 2017). These networks predominate in all M/EEG recordings (even those which are task based) and are key to healthy brain function (Barttfeld et al., 2015; Kaiser et al., 2015; Lewis et al., 2009; Li et al., 2012; Peterson et al., 2014; Philippi et al., 2015; Reineberg et al., 2015; Sheline and Raichle, 2013; Tessitore et al., 2012; Venkataraman et al., 2012; Wu et al., 2014; Wurina et al., 2012). We rely on these iterant dynamics, rehearsing multiple task scenarios, to provide a varied and unbiased spatio-temporal repertoire of the source reconstruction problems one might expect from any dataset on which to then test our inversion schemes.

The second problem then, having identified an appropriate and representative real dataset (as opposed to simulated data), is the lack of access to the ground truth with which to compare recording/inversion techniques. Here we leverage new analytic techniques to quantify the sensitivity of MEG source inversion schemes by progressively deforming the anatomical models (Lopez et al., 2013; López et al., 2017; Stevenson et al., 2014). Specifically, we quantify how distortions in the MRI-extracted cortical manifold (mesh) affect our ability to predict or model the underlying current distribution (using cross validation error and Free energy). The technique assumes that the MEG sensor level data are due to current flow normal to the cortical surface but makes no assumptions about how this current should be distributed. The rationale is that the best MEG inversion scheme will be the most sensitive to subtle distortions of the cortical anatomy (as we know that MEG data derives from grey matter structure). This spatial distortion metric then provides a principled basis for comparing different a priori inversion assumptions (i.e. different algorithms) and recording techniques. We are aiming for a generic method to provide a benchmark to refine inversion (or recording) methods based on human electrophysiological data from multiple labs.

The paper proceeds as follows: we first parcel resting state datasets into brief epochs using a hidden Markov model (Baker et al., 2014). The epochs for the four dominant networks were then amalgamated into four network-specific datasets for each subject and taken forwards for inversion. These datasets were then inverted onto a library of subject-specific distorted meshes, for which we had control over the spatial detail available in the forward model. For each of these meshes, and for each inversion scheme, we quantified the model fit using cross validation and Free energy metrics. As expected, we found that the greater the distortion from the true cortical mesh, the poorer the model fit. We then used this spatial quantification to compare different inversion schemes (implemented as different co-variance prior assumptions). For these data, we found that the beamformer-based priors (EBB) were the most sensitive to small deviations from the true anatomy. In addition to distinguishing between algorithms, here we also tested whether we could use the same methods to distinguish between datasets collected with and without a head-cast (Meyer et al., 2017a; Troebinger et al., 2014a; 2014b), where the accuracy of forward model is more precisely known, and those collected without and found marginal (but not significant) differences.

2. Methods

2.1. MRI

Subjects underwent two MRI scans using a Siemens Tim Trio 3 T system (Erlangen, Germany). For the head-cast scan, the acquisition time was 3 min 42 s, in addition to 45 s for the localizer sequence. The sequence implemented was a radiofrequency (RF) and gradient spoiled T1 weighted 3D fast low angle shot (FLASH) sequence with image resolution 1 mm^3 (1 mm slice thickness), field-of view set to 256, 256, and 192 mm along the phase (A–P), read (H–F), and partition (R–L; second 3D phase encoding direction) directions respectively. A single shot, high

readout bandwidth (425 Hz/pixel) and minimum echo time (2.25 ms) was used. This sequence was optimized to preserve head and scalp structure (as opposed to brain structure). Repetition time was set to 7.96 ms and excitation flip angle set to 12° to ensure sufficient SNR. A partial Fourier (factor 6/8) acquisition was used in each phase-encoded direction to accelerate acquisition. For the anatomical scan later used to construct the cortical model, multiple parameter maps (MPM) were acquired to optimise spatial resolution of the brain image (to 0.8 mm). The sequence comprised three multi-echo 3D FLASH (fast low angle shot) scans, one RF transmit field map and one static magnetic (B0) field map scan (Weiskopf et al., 2013).

2.2. Head-cast construction

Scalp surfaces from the head-cast MRI data were extracted using SPM12 (http://www.fil.ion.ucl.ac.uk/spm/) by registering MRI images to a tissue probability map which classified voxels according to tissue makeup (e.g. skull, skin, grey matter etc.). The skin tissue probability map was transformed into a surface using the 'isosurface' function in MATLAB® and then into standard template library format with the outlines of three fiducial coils digitally placed at conventional sites (left/right pre-auricular and nasion). Next, a positive head model was printed using a Zcorp 3D printer (600 × 540 dots per inch resolution) and this model placed inside a replica dewar-helmet with liquid resin poured between the two, resulting in a flexible, subject specific, foam head-cast with fiducial indentations in MRI-defined locations (Meyer et al., 2017a).

2.3. MEG recording

Resting state data was acquired from 12 healthy subjects using head-casts (age: 26.6 ± 3.5 yrs (mean + sd)) and 12 other healthy subjects without head-casts (age: 25.2 ± 6.6 yrs). All subjects were right handed, had normal or corrected-to-normal vision, and had no history of neurological or psychiatric disease. Informed written consent was given by all subjects and recordings were carried out after obtaining ethical approval from the University College London ethics committee (ref. number 3090/001).

All subjects underwent a 10 min resting state scan with eyes kept open and instructed to fixate on a central cross on a screen, using a CTF 275 Omega MEG system. The head was localised using the three head-cast-embedded fiducials (head-cast subjects) or fiducials placed on the nasion and left and right pre-auricular points (non-head-cast subjects). Average range of absolute head movement within the 10 min resting state recording was 0.26 ± 0.06, 0.24 ± 0.05, 1.1 ± 0.54 mm (X,Y,Z directions; ± SEM) for head-cast and 3.2 ± 0.5, 3.0 ± 0.5, 3.3 ± 0.2 (X,Y,Z directions; ± SEM) for non-head-cast data. The data were sampled at a rate of 1200 Hz, imported into SPM12 and filtered (4th order butterworth bandpass filter: 1–90 Hz, 4th order butterworth bandstop filter 48–52 Hz) and downsampled to 250 Hz.

Traditional inverse problem solutions are based on the assumption that the data are stationary during the period of inversion. However, resting state data contains rapid dynamics that do not accord well with this assumption (Woolrich et al., 2013). Therefore, in order to improve stationarity within a given epoch, we parcellated the data into self-similar periods that capture the resting state network transitions (100–200 ms) using a Hidden Markov Model (HMM) that could identify the rapid formation and dissolution of recurring resting state networks (Baker et al., 2014). With this, a 'statepath' was estimated for each 10 min resting state block, which tracks the fine spatiotemporal dynamics and allocates each point in time to one of eight dominant network states (Baker et al., 2014). For this statepath determination, a copy of each subject's sensor level data was dimensionally reduced using principle component analysis (PCA) to derive 40 components of unit variance and mean (Woolrich et al., 2013). With these data, an 8 state Hidden Markov model (HMM; www.fmrib.ox.ac.uk/~woolrich/HMMtoolbox)

was then applied to derive the most probable state at all points in time (the statepath) (Fig. 1A) (Baker et al., 2014). The continuous 600s of data was then epoched into 200 ms blocks/epochs (the average scale of individual state transitions (Baker et al., 2014) and each 200 ms data epoch allocated to a new dataset according to the state that was most dominant during that time period. This resulted in 8 new datasets each with a collection of epochs that corresponded to a distinct network state as identified by the HMM. The four most dominant states for each individual subject (as measured by most time spent in that state) was taken forward for further analysis and spatial estimates averaged across those four inversions for each subject. As the HMM was performed on individual, rather than group concatenated data, state numbers did not directly correspond across subjects. This however permits superior partitioning within subjects since it allows the model to optimally fit states to the individual subjects, rather than fitting individual data to group states that are (or are not) common across all subjects. This resulted in 815 ± 56.9 data segments per partitioned dataset - equivalent to 163s of continuous data. Since the HMM selected periods of self-similarity within the resting state, partial correlation maps were examined for all subjects

to check that segregation into separate networks had occurred and this segregation did not relate to eye blinks, muscle artefacts or cardiac interference (Fig. 1B).

2.4. Subject-specific cortical mesh libraries

To extract the cortical pial mesh surface, we used Freesurfer software optimized for MPM scans by using PD and T1 volumes as inputs. A detailed description of these methods of cortical surface reconstruction and optimization can be found in earlier work (Carey et al., 2017; Lutti et al., 2014). Then, for each individual subject, the pial cortical mesh was taken and deformed using a 3D weighted Fourier analysis that effectively decomposes the original 3D mesh structure into spatial harmonic components (Chung et al., 2007). These are then sequentially combined to form a set of meshes of progressively increasing spatial detail (Fig. 2B) (Stevenson et al., 2014). These meshes are called the Weighted Fourier Series (WFS) and result in a library of meshes for each subject with different levels of spatial detail, from a completely smooth pair of ovoid surfaces (mesh 1) up to a mesh which is very similar to that extracted from the MRI of the subject's brain (mesh 50) (see Fig. 2B). The WFS can be expressed as follows:

$$F_\sigma^k[f](\omega) = \sum_{l=0}^{L} \sum_{m=-1}^{l} e^{-l(l+1)} f_{lm} S_{lm}(w)$$

where σ is the bandwidth of the smoothing kernel (set at 0.0001), L is the harmonic order of the surface, S_{lm} is the spherical harmonic of degree l and order m, and the Fourier coefficients are given by $<f_{lm}=f, S_{lm}>$, where f is determined by solving a system of linear equations (Chung et al., 2007). All meshes, including the true mesh, were downsampled by a factor of 10 in Freesurfer, to ~33,000 vertices per mesh to aid computational efficiency. This resulted in a mean vertex distance to nearest neighbour of 1.7 mm (mean within-subject 5 & 95 percentiles 0.8–2.6 mm).

2.5. Source reconstruction

We used SPM12 which implements multiple different inversion schemes within a common empirical Bayesian pipeline, whereby all processing steps, besides the choice of prior source co-variance matrix are held constant (Friston et al., 2008; López et al., 2014). The choices of prior source covariance matrices used embody four popular inversion algorithms: The traditional unweighted Minimum norm (MMN) in which the source covariance is a diagonal (IID) matrix (Hämäläinen and Ilmoniemi, 1994). An (unweighted) LORETA-like solution (Pascual-Marqui, 2002) based on a surface (rather than volumetric) smoothing using Green's function (Harrison et al., 2007) (LOR). An Empirical Bayesian Beamformer (EBB) inversion (Belardinelli et al., 2012) in which a direct estimate of prior source co-variance is made based on the sensor-level data (and zero regularization). The main difference between the EBB scheme and the classic linearly constrained minimum variance (LCMV) (Van Veen et al., 1997) scheme is that the EBB attempts to fit the measured data (by optimizing model evidence). This has the advantage that EBB attempts to explain the measured data and can therefore provide a direct model evidence metric (which can then be compared to other linear inversion algorithms); but the disadvantage is that if there is external noise in the measured data the Bayesian scheme will still attempt to fit it (unless appropriately modelled) whereas the LCMV (spatially filtered) solution will be unaffected. The multiple sparse priors algorithm (MSP) (Friston et al., 2008) uses an optimized search to select mixtures of source covariance matrices each describing the activity of separate patches of cortex. For each of the different algorithms, the source level estimates were weighted against a diagonal sensor noise covariance matrix (effectively regularizing the solution) within the same empirical Bayes framework. For more details on each of these methods and the empirical Bayesian optimization please see (Belardinelli et al.,
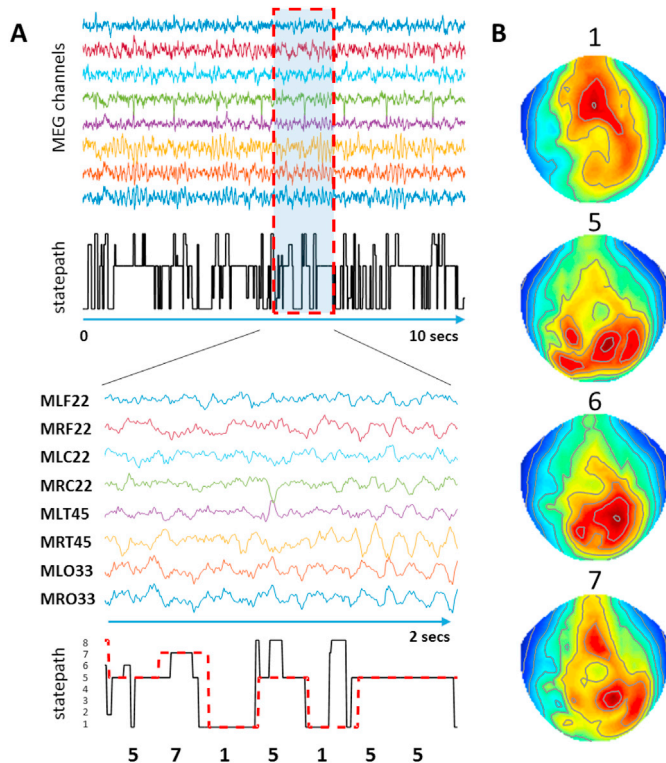

Fig. 1. MEG data and Hidden Markov Model network allocation. A. Top panels show a selection of MEG sensor time series for head-cast subject 1. Middle panel demonstrates the statepath as determined by the HMM, showing rapid transitions between different states. Timecourses recorded from a random subset of sensors is shown. Bottom panel shows a short (2s) period of data (MEG channel labels e.g. MLF22 shown), expanded with the corresponding section of the statepath (black line). Overlying the statepath trace (black) is the modal statepath for each 200 ms period (dash red line) and this modal statepath trace was used to sort each 200 ms data epoch into 1 of 8 states according to which state was most frequent during that period. These epochs of data for each of these states were aggregated together to form 8 new datasets. **B.** The statepath traces for the four most dominant states here (1, 5, 6, 7) are shown correlated against the time series amplitudes in all sensors to derive a map of activity (red, greater positive correlation, blue, less positive correlation) associated with each particular state. Data are shown in sensor space in order to check that differential network parcellation had indeed occurred according to the statepath detection method and was not artefactual (e.g. that the topography resembles that expected for an eye blink). In the example shown here – state 5 is the most common state – corresponding to the posterior alpha rhythm.
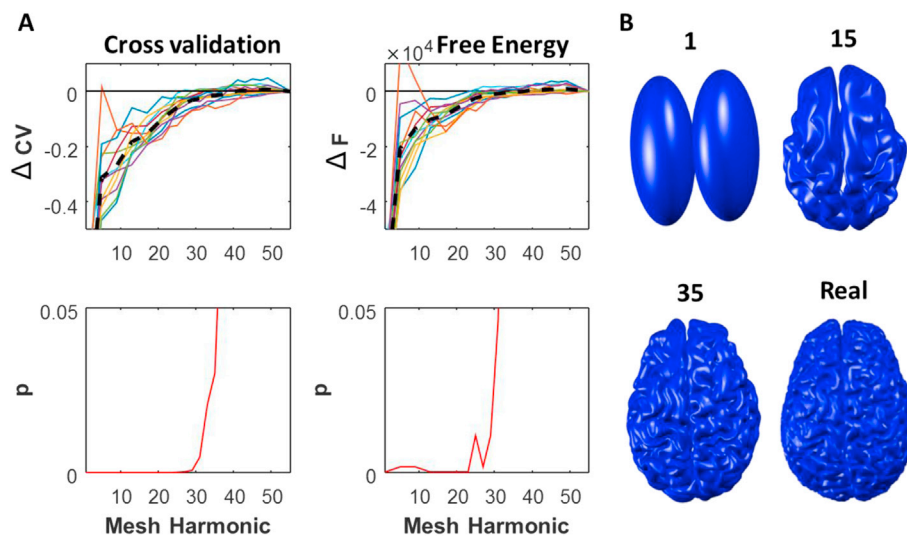
Fig. 2. Relative Cross Validation and Free energy results for a library of different meshes for head-cast resting data using EBB inversion. A. Increasing cross validation data explained (ΔCV) and relative Free energy ΔF with improving spatial resolution of harmonic meshes (from left to right). Top plots show individual subject ΔCV (left) and ΔF (right) values with superimposed mean value (dashed black line). Lower panels show the group level statistical significance using *t*-test (ΔCV) and Bayesian model comparison (ΔF) for cross validation and Free energy respectively. **B.** Example selection of meshes from subject 1 showing different levels of distortion, from smooth ovoid surfaces (WFS mesh 1) up to the real mesh.

2012; Friston et al., 2008; López et al., 2014).

The HMM parcellated datasets originally contained 272 channels and 50 time samples (3 channels were damaged and excluded from the data). As part of the cross validation procedure, we excluded 10% of these channels from the inversion stage, leaving 245 channels of data. This procedure left 27 unused channels out of the inversion, the data from which were then predicted based on the source estimates and compared to the real sensor data. For each of the ten cross-validation folds (each time a random 10% of channels were selected to be left out), the remaining (245 channels of) data were temporally decomposed into 16 orthogonal temporal modes and inverted onto each surface from the library of individualised and distorted cortical meshes (WFS) for each subject. Each inversion provided 10 cross-validation scores (the prediction of the hidden channels) which were then converted to a percentage of source data explained (more positive = better model fit) and averaged across the 10 folds and across the four state datasets inverted per subject.

Both cross validation and Free energy provide metrics which can be used to directly compare different models of the same data, with different independent methods of preventing over-fitting (Bonaiuto et al., 2018; Henson et al., 2009). Whilst the Free energy provides a useful relative model fit metric for any given dataset, the absolute value is data dependent and therefore cannot be used to compare between datasets. In contrast, the cross validation gives a meaningful quantification of the total amount of data explained and can be used to compare across the different groups (e.g. head-cast versus non head-cast). After converting cross validation error to the percentage of total data explained by the model, both the CV % data explained and the Free energy increase with improving model fit.

Here we performed these inversions using our subject specific libraries of anatomically degraded meshes from the WFS with different levels of distortion and compared them to an inversion performed using the real mesh for each subject. We constrained the source estimates to lie on the cortical surface and normal to that surface (Phillips et al., 2002). In order to facilitate comparisons of how the anatomical distortion affected the inversions for different subjects with different baseline measures of model fit to their real brain meshes, we normalised the cross validation percentage data explained and Free energy metrics by subtracting the values for the real mesh to derive a relative measure (ΔCV and ΔF). As such, a worse fit gives a negative value of ΔCV (and ΔF) and we would predict that as the anatomical complexity of the mesh increases (mesh is less distorted) and approaches that of the real mesh – the quality of the model should improve and the ΔCV and ΔF values should approach zero. Across our group of subjects, we determined the level of distortion (harmonic) at which this first becomes statistically distinguishable from

the real mesh using a *t*-test of ΔCV values for each harmonic compared to zero. Similarly for ΔF, we use Bayesian model comparison to find out how much distortion is needed to make two meshes distinct. For example, using algorithm A we may find that inversions are not much worse (ΔCV and ΔF not significantly different from zero at the group level) using very distorted meshes all the way down to quite a smooth mesh harmonic (e.g. <10) i.e. the inversion algorithm is very insensitive to changes in the mesh shape. Alternatively, using algorithm B, we find that even with very subtle distortions of the mesh, that the inversion fit deteriorates substantially, so down at just harmonic 40 (which looks to the naked eye to be very similar to real mesh) the ΔCV and ΔF values are significantly below zero at the group level. This point (10 and 40 in this example), labelled the highest distinguishable harmonic (HDH), identifies the minimum amount of mesh distortion that can be reliably distinguished from the real mesh by inversion and can be converted into a conventional spatial metric (mms) by comparing vertex distances. The three dimensional Euclidean distance between every vertex from the HDH and the corresponding vertex on the true mesh was therefore calculated to give a distribution of 33,000 distances between the real mesh and the first mesh that can be distinguished from it statistically by inversion. We then took the upper, 95th percentile as an estimate of the distance between the two meshes. In addition to using the 95th percentile rather than the mean, this method is more conservative than that previously employed (Stevenson et al., 2014), as this directly matches corresponding vertices and therefore reduces the underestimation that could result if, following harmonic distortion, a vertex now lies closer to a non-corresponding other vertex. This distance therefore represents an upper bound on the spatial discriminability of both head-cast and non-head-cast resting state data.

2.6. Control analyses

In order to verify our findings, we performed a number of control analyses. Firstly, we used the same data but destroyed its correspondence with the MEG sensor locations by randomly shuffling the MEG channels labels and repeated the analysis above 10 times for each subject and averaged over the ΔCV for these different shuffled dataset inversions. It is worth noting that sensor shuffling also removes all the (real) spatial correlations between sensors, however since they are not independent, they are not perfectly exchangeable and thus we also performed a secondary control analysis. For this, we used the correct sensor labels but instead degraded our data by introducing different amounts of uncorrelated scaled white noise at the sensor level to change the signal-to-noise ratio of the sensor level data (5 dB to - 20 dB) (Troebinger et al., 2014a).

In both cases (sensor shuffling and noise addition), one would expect the ability to discriminate the true generative model from distorted ones to decrease.

3. Results

3.1. Anatomical cortical model

In order to determine the sensitivity of the inversion to the level of detail in the underlying anatomical mesh model we calculated the ΔCV and ΔF for each subject in the head-cast dataset across their subject-specific library of distorted meshes (Fig. 2). This showed increasing cross validation sensor data explained (ΔCV; reduced error) and increasing Free energy (ΔF) for all subjects. Statistical group level testing revealed that meshes lower (more deformed) than the 35 harmonic could be distinguished from the real mesh by ΔCV (HDH 35; $t_{11} = -2.49$, $p = 0.03$) and lower than 31 by ΔF (BMC, exceedance $p = 0.046$; Fig. 2). Having confirmed our results using the two metrics – we proceeded with the ΔCV metric which can be compared across subjects and recording methods.

3.2. Inversion algorithms/source covariance priors

The effect of source co-variance prior assumptions was then assessed by repeating the process for three other commonly implemented inversion algorithms (MMN, LOR, MSP) on our head-cast dataset. These showed lower anatomical mesh discriminability for all alternative algorithms with an HDH for MMN of 25 ($t_{11} = -1.80$ -, $p = 0.036$), an HDH of 25 ($t_{11} = -1.79$; $p = 0.039$) for LOR and an HDH of 17 for MSP ($t_{11} = -2.55$; $p = 0.027$) (Fig. 3A).

The mean distance between vertices on these meshes and corresponding vertices on the real (MRI-extracted) mesh was then calculated for each subject. These were then averaged to give a spatial measure of anatomical discriminability (Fig. 3A), under the different prior covariance assumptions as implemented in the different inversion algorithms. This ranged from 3.7 mm for EBB to 6.0 mm for MNN/LOR and 9.4 mm for MSP (Fig. 3B). Directly comparing the amount of data explained through cross validation across inversion conditions (using the real mesh) demonstrated a significant difference between EBB and the other 3 algorithms (EBB/MMN; paired t-test – $t_{11} = 7.6, p < 0.001$; EBB/LOR; paired t-test – $t_{11} = 7.6, p < 0.001$; EBB/MSP; paired t-test – $t_{11} = 14.4, p < 0.001$).

A 2-factor within-subject ANOVA of % cross validation data explained with factors – inversion type (EBB, MMN, MSP) and mesh smoothness (harmonic 1 and harmonic 50) showed a significant interaction between inversion type and mesh harmonic level ($F_2 = 29.9$, $p < 0.0001$). Post hoc examination showed that this was driven by a stronger effect of mesh distortion on the MSP inversion algorithm than EBB or MMN.

3.3. Head-cast versus conventional MEG

The EBB algorithm was therefore taken forwards for a comparison of head-cast versus non-head-cast datasets (Fig. 4). We found that the HDH was higher for the head-cast recorded dataset at 35 (3.7 mm; $t_{11} = -2.49$, $p = 0.03$) than for the non-head-cast related dataset at 29 (5.0 mm, $t_{11} = -2.70$, $p = 0.021$). Furthermore, the absolute (non-normalised) amount of data explained through cross validation (CV) was higher in the head-cast ($83.2 \pm 0.71\%$) than in the conventional recordings ($79.7 \pm 1.79\%$) although this difference was not significant (mixed ANOVA; main effect of hcMEG versus cMEG, $F = 3.06$; $p = 0.09$).

3.4. Control analyses

Finally, we checked our analyses by repeating our inversions (Head-cast dataset, EBB algorithm) but after degrading the consistent relationship between our sensor positions and sources by shuffling the sensor labels. As expected, this resulted in a breakdown of the previously shown relationship between cross validation and Free energy with mesh distortion. Notably we found that lower harmonics (smoother) meshes now showed higher ΔCV (Fig. 5A) (smoother surfaces superior when sensors shuffled). Thereafter we degraded our data (without sensor shuffling) by the addition of varying levels of Gaussian white noise (5 to -20 dB) and again repeated our analysis (Fig. 5B). This demonstrated that with increasing levels of noise (decreasing SNR), the curve describing the relationship between cross validation and harmonic mesh function flattened and the crossing point (HDH) reduced, indicating that the inversion was no longer able to statistically distinguish the more complex meshes from the real mesh, as would be expected if the data are primarily noise.

4. Discussion

Resting state data is dynamic, emulates task induced network changes, is predictive of behaviour and is simple to acquire in both control and patient populations (Baker et al., 2014; Larson-Prior et al., 2013; Niso et al., 2016; O'Neill et al., 2017; Philippi et al., 2015; Taylor et al., 2017). Here we show that it can provide a ready substrate for principled testing of MEG recording methods and inversion assumptions including anatomical forward modelling and functional (co-variance) priors.

We showed that in moving the cortical surface from a heavily distorted version to the true anatomy, a significant and monotonic improvement occurred in model fit. This improvement saturated for some inversion schemes before others, with the beamformer based
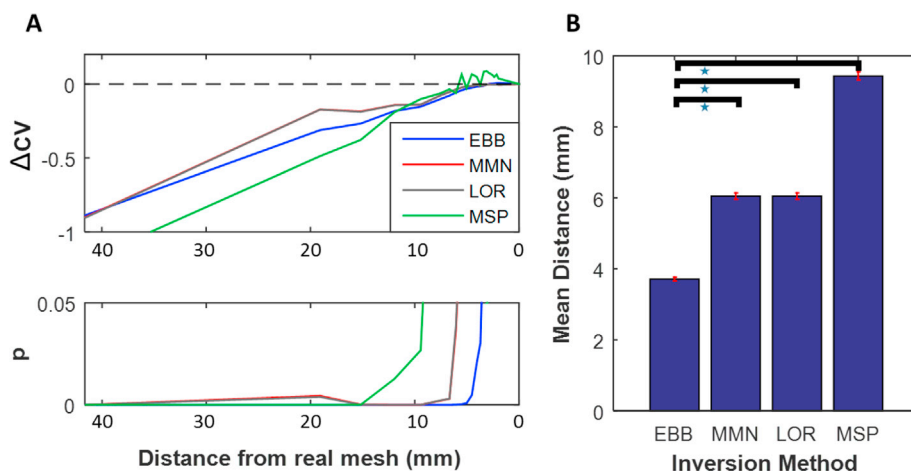

Fig. 3. Effect of different inversion schemes on head-cast resting state resolution. A. Relative cross validation data explained (ΔCV) for head-cast MEG dataset (averaged across 12 subjects), shown for 4 different inversion types (EBB, MMN, LOR & MSP) according to mean distance of the distorted mesh (x axis) from the real mesh. Lower panel shows the group level statistical significance by t-testing (note MMN (red) and LOR (grey) have very similar values and therefore lines are closely overlapping). **B.** Mean distance of vertices from highest significant mesh identified in the left hand panel and the real mesh. Error bars show SEM of this distance across subjects using their individualised meshes and group level HDH.
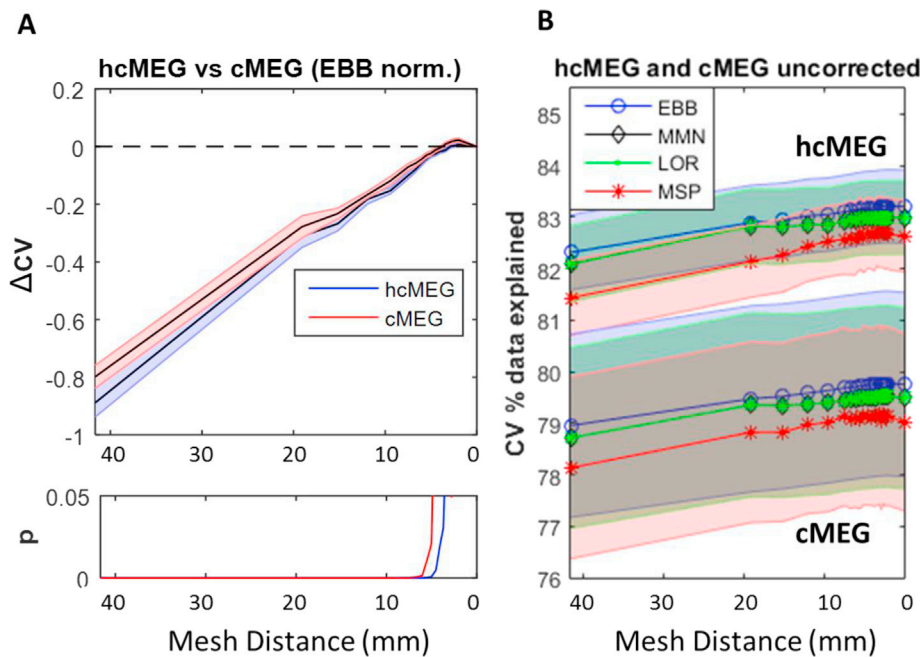
Fig. 4. Comparison of head-cast versus conventional MEG cross validation accuracy. A. Comparison of ΔCV for different recording methodologies with head-cast (hcMEG) data showing higher discrimination (3.7 mm) than conventional MEG (cMEG) data (5.0 mm). **B.** Absolute cross validation data explained is also higher in the head-cast versus the conventional MEG. Note that this holds for all inversion types and for all levels of mesh distortion, but was not significantly different ($F = 3.06$, $p = 0.09$).
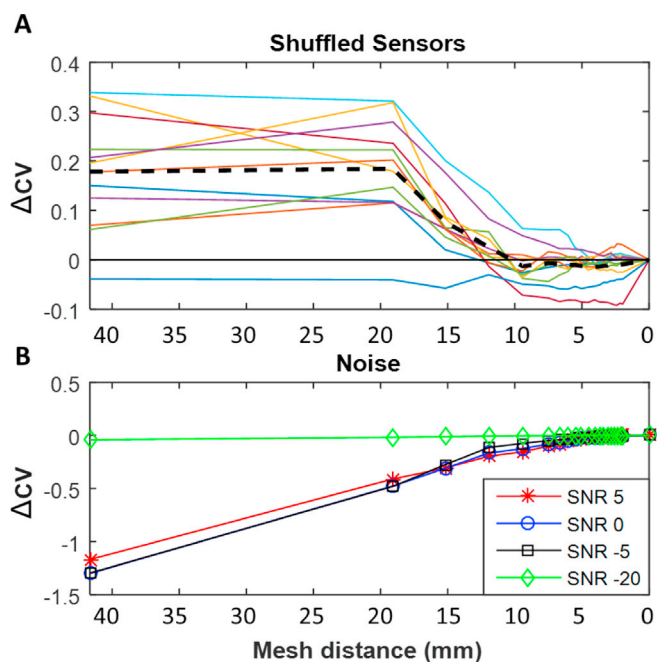

Fig. 5. Effects of shuffling sensors (A) and replacing data with varying levels of Gaussian White noise (B). A. The relationship between the percentage of data explained through cross validation (ΔCV) of inversion (normalised to the real mesh) and increasing harmonic mesh after random shuffling of sensor labels. Note the decreasing cross validation fit with increasing mesh harmonic (increasing mesh detail) with the dashed line showing the average across all subjects. **B.** The relationship between (mean of all 12 subjects) ΔCV of inversion and increasing harmonic mesh after addition of Gaussian white noise across a range of noise additions (from SNR 5 to SNR -20 in decibels (dB)).

invasive human data. The HMM constrained the source estimates to select out self-similar and stationary epochs – under the assumption that this would also lead to a clearer picture of the underlying source distributions (essentially segmenting multiple overlapping source distributions into different temporal parcels and therefore simplifying the source reconstruction problem). We tested whether this step was justified by comparing the resolution estimates from the HMM parcelled data with arbitrary mixtures of parcels (supplementary Fig. S1). We found that the HMM approach demonstrated a significantly higher resolution in each subject compared to the non-HMM dataset ($t11 = 3.18$, $p < 0.009$). An alternative scheme, avoiding both resting state and the HMM, would be to define stationary periods of time around a common battery of tasks/stimuli to be used (but first agreed upon) universally; however, the concern would be that the ensuing cortical responses might still only occupy a small portion of the possible spatio-temporal solution space.

We then compared algorithms in two ways: by the model fit (or amount of data predicted), and by the sensitivity of each algorithm to the true anatomy. These two tests need not necessarily have been in accord. For example, had we used a bunny-shaped blancmange mould instead of a cortical surface, we would still have been able to rank the algorithms based on the amount of data predicted; but we would not have expected any monotonic improvement as features were added to the bunny. A related control analysis (Fig. 5) is that when we used the same data but with shuffled lead-fields (destroying the link between the sensors and the anatomy) the amount of data we are able to predict actually decreases as the cortical model approaches the truth. It is therefore striking that the models that benefitted most from the true cortical manifold were also those that predicted the most data. This adds anatomical validity, confirming that the data being described is indeed generated by pyramidal cell populations normal to the cortical surface. Furthermore however, it also allows us to quantify algorithm performance in millimetres without being restricted to or dependent on using a particular processing framework (e.g. SPM) or cost function (e.g Free energy or cross validation) (Stevenson et al., 2014).

Across anatomy (Fig. 2) and inversion assumptions (Fig. 3), the parametric (Free energy from empirical Bayes) and non-parametric (cross-validation) metrics of model fit were in accordance. This helps build confidence in the parametric Free energy metric which is considerably faster, makes use of all the available data, and has a direct probabilistic interpretation (i.e. how much more likely is one model than

algorithms (closely followed by Minimum norm) continuing to improve up until the cortical surface deviated by, on average, <4 mm from the ground-truth. Critically, rather than compare methods through simulation or a limited task set (with ground truth from another modality), we have presented a method to optimise MEG forward and inverse models whilst minimizing selection bias and based on a plentiful supply of non-

another?). The Bayesian formalism is however predicated on comparing how different models explain the same data; the use of cross-validation, which provides an absolute quantitative measure of data predicted, also allowed us to compare between different datasets (head-cast versus non-head-cast). We confirm here earlier reports that within-session head movements are reduced for head-cast versus non-head-cast MEG (Bonaiuto et al., 2017; Liuzzi et al., 2017; Meyer et al., 2017a; 2017b). We were surprised that the head-cast and non-head-cast data were statistically indistinguishable in terms of spatial resolution (mixed ANOVA; main effect of hcMEG versus cMEG, F = 3.06; p = 0.09). However, we believe that the comparison was particularly conservative in that the non-headcast group showed very little average head movement (3 mm), comparable to the average mesh vertex spacing (1.7 mm). Another limitation could be that the models we are using do not sufficiently capture the physics or physiology of the generators of the measured magnetic fields; and that the resolution is constrained by the models and not the recording. This could include for example, un-modelled noise sources such as the heartbeat, eye-blinks and other sources of noise. Although the HMM states we used were visually inspected to avoid common artefacts such as eye-blinks, it is possible that some of the modelling deficiencies come from failure to explain data that does not arise from the cortex (e.g. heart-beats). Finally, the head-cast and non-head-cast cohorts did not overlap and a more sensitive analysis would have been to perform a within-subject comparison.

Here we used individualised meshes with current flow constrained to be normal to the cortical surface. These normal constraints, based on the MRI surface extraction, will themselves have some noise (which will be attenuated by smoothing in the lower harmonic orders) and could be one of the limiting factors in our spatial resolution estimates. One further consideration was that our estimates of resolution might have been confounded by absolute levels of signal power. We found no evidence for this in any frequency band (supplementary Fig. S2). We did, however, find that our subject group was divided into subjects showing a dominant theta/delta peak and those showing a peak in the alpha range (supplementary Fig. S2). We found that consistent with previous (MEG within-subject) findings (Stevenson et al., 2014), we achieved a greater sensitivity to spatial distortion in the subjects with a peak in a higher frequency range (p < 0.002). This is consistent with invasive observations that higher frequency oscillations tend to have smaller spatial coherence domains (Leopold et al., 2003). We should note however that the theta cycle is close to our chosen window length (200 ms) and this could be another reason for the lower spatial resolution estimates in those subjects dominated by lower frequencies.

We found that the Multiple Sparse Priors algorithm had the least dependence on the true anatomy and also explained the least data. We should note however that the MSP-based analyses implemented here were generic and constructed from a limited set of 512 patches (or priors) placed at approximately evenly spaced vertices. The MSP algorithm, although perhaps the most elegant and comprehensive method we tested, is also computationally disadvantaged by the need to search over a large space of possible patch/prior combinations and the inherent pitfalls of local extrema in this optimization. A more robust way to implement this algorithm would have been to select the best model from many random patch choices (Troebinger et al., 2014b). Moreover, the relationship between sparsity assumed by the algorithm, and resting state networks used here for analysis, is beyond the scope of this study. We should note that the implementations of all the algorithms we tested were generic and not individually optimized. For example, had we used an optimized MNE framework (Gramfort et al., 2014) then no doubt its relative sensitivity would have improved. Here we focus on developing the benchmark and showing its application, rather than an exhaustive comparison of algorithms.

This study was analysed using broadband (1–90 Hz) resting state data. Therefore, whether a similar level of spatial discriminability at the mm scale can be demonstrated when more selective data is used (e.g. frequency filtered or spatially restricted) remains to be shown. For example, future work might test different frequency bands (eg <30 Hz, >30 Hz) against different anatomy for example infra/supra granular cortical surfaces (Arnal and Giraud, 2012; Bastos et al., 2012; Bonaiuto et al., 2018, 2017).

Overall, we have here demonstrated and validated a robust and unbiased method for the comparison of inversion and MEG recording methods that can in future be used to evaluate and optimise future techniques.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2018.07.030.

References

Arnal, L.H., Giraud, A.-L., 2012. Cortical oscillations and sensory predictions. Trends Cognit. Sci. 16, 390–398. https://doi.org/10.1016/j.tics.2012.05.003.

Baillet, S., 2015. Forward and inverse problems of MEG/EEG. In: Encyclopedia of Computational Neuroscience. Springer, New York.

Baillet, S., Mosher, J.C., Leahy, R.M., 2001. Electromagnetic brain mapping. IEEE Signal Process. Mag. 18, 14–30. https://doi.org/10.1109/79.962275.

Baker, A.P., Brookes, M.J., Rezek, I. a, Smith, S.M., Behrens, T., Smith, P.J.P., Woolrich, M., 2014. Fast transient networks in spontaneous human brain activity. Elife 2014, 1–18. https://doi.org/10.7554/eLife.01867.

Barttfeld, P., Uhrig, L., Sitt, J.D., Sigman, M., Jarraya, B., Dehaene, S., 2015. Signature of consciousness in the dynamics of resting-state brain activity. Proc. Natl. Acad. Sci. U. S. A 112, 887–892. https://doi.org/10.1073/pnas.1418031112.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., Friston, K.K.J., 2012. Canonical microcircuits for predictive coding. Neuron.

Belardinelli, P., Ortiz, E., Barnes, G.R., Noppeney, U., Preissl, H., 2012. Source reconstruction accuracy of MEG and EEG Bayesian inversion approaches. PLoS One 7, e51985. https://doi.org/10.1371/journal.pone.0051985.

Bonaiuto, J., Meyer, S.S., Little, S., Rossiter, H., Callaghan, M.F., Dick, F., Barnes, G.R., Bestmann, S., 2017. Laminar-specific Cortical Dynamics in Human Visual and Sensorimotor Cortices. https://doi.org/10.1101/226274 bioRxiv 226274.

Bonaiuto, J., Rossiter, H.E., Meyer, S.S., Adams, N., Little, S., Callaghan, M.F., Dick, F., Bestmann, S., Barnes, G.R., 2018. Non-invasive laminar inference with MEG: comparison of methods and source inversion algorithms. Neuroimage 167. https://doi.org/10.1016/j.neuroimage.2017.11.068.

Carey, D., Caprini, F., Allen, M., Lutti, A., Weiskopf, N., Rees, G., Callaghan, M.F., Dick, F., 2017. Quantitative MRI provides markers of intra-, inter-regional, and age-related differences in young adult cortical microstructure. Neuroimage. https://doi.org/10.1016/J.NEUROIMAGE.2017.11.066.

Chung, M.K., Dalton, K.M., Shen, L., Evans, A.C., Davidson, R.J., 2007. Weighted Fourier series representation and its application to quantifying the amount of gray matter. IEEE Trans. Med. Imag. 26, 566–581. https://doi.org/10.1109/TMI.2007.892519.

Friston, K.K.J., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., Henson, R., Flandin, G., Mattout, J., 2008. Multiple sparse priors for the M/EEG inverse problem. Neuroimage 39, 1104–1120. https://doi.org/10.1016/j.neuroimage.2007.09.048.

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Parkkonen, L., Hämäläinen, M.S., 2014. MNE software for processing MEG and EEG data. Neuroimage 86, 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027.

Hämäläinen, M.S., Ilmoniemi, R.J., 1994. Interpreting magnetic fields of the brain: minimum norm estimates. Med. Biol. Eng. Comput. 32, 35–42. https://doi.org/10.1007/BF02512476.

Harrison, L.M., Penny, W., Ashburner, J., Trujillo-Barreto, N., Friston, K.J., 2007. Diffusion-based spatial priors for imaging. Neuroimage 38, 677–695. https://doi.org/10.1016/J.NEUROIMAGE.2007.07.032.

Henson, R.N., Mattout, J., Phillips, C., Friston, K.K.J., 2009. Selecting forward models for MEG source-reconstruction using model-evidence. Neuroimage 46, 168–176. https://doi.org/10.1016/j.neuroimage.2009.01.062.

Kaiser, R.H., Andrews-Hanna, J.R., Wager, T.D., Pizzagalli, D.A., RJ, D., J, J., 2015. Large-scale network dysfunction in major depressive disorder. JAMA Psychiatry 72, 603. https://doi.org/10.1001/jamapsychiatry.2015.0071.

Koenig, T., Prichep, L., Lehmann, D., Sosa, P.V., Braeker, E., Kleinlogel, H., Isenhart, R., John, E.R., 2002. Millisecond by millisecond, year by year: normative EEG microstates and developmental stages. Neuroimage 16, 41–48. https://doi.org/10.1006/nimg.2002.1070.

Larson-Prior, L.J., Oostenveld, R., Della Penna, S., Michalareas, G., Prior, F., Babajani-Feremi, A., Schoffelen, J.-M., Marzetti, L., de Pasquale, F., Di Pompeo, F., Stout, J., Woolrich, M., Luo, Q., Bucholz, R., Fries, P., Pizzella, V., Romani, G.L., Corbetta, M., Snyder, A.Z., 2013. Adding dynamics to the human connectome project with MEG. Neuroimage 80, 190–201. https://doi.org/10.1016/J.NEUROIMAGE.2013.05.056.

Leopold, D.A., Murayama, Y., Logothetis, N.K., 2003. Very slow activity fluctuations in monkey visual cortex: implications for functional brain imaging. Cerebr. Cortex 13, 422–433.

Lewis, C.M., Baldassarre, A., Committeri, G., Romani, G.L., Corbetta, M., 2009. Learning sculpts the spontaneous activity of the resting human brain. Proc. Natl. Acad. Sci. U. S. A 106, 17558–17563. https://doi.org/10.1073/pnas.0902455106.

Li, P., Li, S., Dong, Z., Luo, J., Han, H., Xiong, H., Guo, Z., Li, Z., 2012. Altered resting state functional connectivity patterns of the anterior prefrontal cortex in obsessive-compulsive disorder. Neuroreport 23, 681–686. https://doi.org/10.1097/WNR.0b013e328355a5fe.

Lin, F.-H., Witzel, T., Ahlfors, S.P., Stufflebeam, S.M., Belliveau, J.W., Hämäläinen, M.S., 2006. Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. Neuroimage 31, 160–171. https://doi.org/10.1016/J.NEUROIMAGE.2005.11.054.

Liuzzi, L., Gascoyne, L.E., Tewarie, P.K., Barratt, E.L., Boto, E., Brookes, M.J., 2017. Optimising experimental design for MEG resting state functional connectivity measurement. Neuroimage 155, 565–576. https://doi.org/10.1016/j.neuroimage.2016.11.064.

López, J.D., Litvak, V., Espinosa, J.J., Friston, K.K.J., Barnes, G.R., 2014. Algorithmic procedures for Bayesian MEG/EEG source reconstruction in SPM. Neuroimage 84, 476–487. https://doi.org/10.1016/j.neuroimage.2013.09.002.

Lopez, J.D., Troebinger, L., Penny, W., Espinosa, J.J., Barnes, G.R., 2013. Cortical surface reconstruction based on MEG data and spherical harmonics. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 6449–6452. https://doi.org/10.1109/EMBC.2013.6611031.

López, J.D., Valencia, F., Flandin, G., Penny, W., Barnes, G.R., 2017. Reconstructing anatomy from electro-physiological data. Neuroimage 163, 480–486. https://doi.org/10.1016/j.neuroimage.2017.06.049.

Lutti, A., Dick, F., Sereno, M.I., Weiskopf, N., 2014. Using high-resolution quantitative mapping of R1 as an index of cortical myelination. Neuroimage 93, 176–188. https://doi.org/10.1016/j.neuroimage.2013.06.005.

Meyer, S.S., Bonaiuto, J., Lim, M., Rossiter, H., Waters, S., Bradbury, D., Bestmann, S., Brookes, M., Callaghan, M.F., Weiskopf, N., Barnes, G.R., 2017a. Flexible head-casts for high spatial precision MEG. J. Neurosci. Meth. 276, 38–45. https://doi.org/10.1016/j.jneumeth.2016.11.009.

Meyer, S.S., Rossiter, H., Brookes, M., Woolrich, M., Bestmann, S., Barnes, G.R., 2017b. Using generative models to make probabilistic statements about hippocampal engagement in MEG. Neuroimage 149, 468–482. https://doi.org/10.1016/j.neuroimage.2017.01.029.

Niso, G., Rogers, C., Moreau, J.T., Chen, L.-Y., Madjar, C., Das, S., Bock, E., Tadel, F., Evans, A.C., Jolicoeur, P., Baillet, S., 2016. OMEGA: the open MEG archive. Neuroimage 124, 1182–1187. https://doi.org/10.1016/J.NEUROIMAGE.2015.04.028.

O'Neill, G.C., Tewarie, P.K., Colclough, G.L., Gascoyne, L.E., Hunt, B.A.E., Morris, P.G., Woolrich, M.W., Brookes, M.J., 2017. Measurement of dynamic task related functional networks using MEG. Neuroimage 146, 667–678. https://doi.org/10.1016/j.neuroimage.2016.08.061.

Pascual-Marqui, R.D., 2002. Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. Methods Find. Exp. Clin. Pharmacol. 24 (Suppl. D), 5–12.

Peterson, A., Thome, J., Frewen, P., Lanius, R.A., 2014. Resting-state neuroimaging studies: a new way of identifying differences and similarities among the anxiety disorders? Can. J. Psychiatr. 59, 294–300. https://doi.org/10.1177/070674371405900602.

Philippi, C.L., Pujara, M.S., Motzkin, J.C., Newman, J., Kiehl, K.A., Koenigs, M., 2015. Altered resting-state functional connectivity in cortical networks in psychopathy. J. Neurosci. 35.

Phillips, C., Rugg, M.D., Friston, K.K.J., 2002. Systematic regularization of linear inverse solutions of the EEG source localization problem. Neuroimage 17, 287–301. https://doi.org/10.1006/NIMG.2002.1175.

Reineberg, A.E., Andrews-Hanna, J.R., Depue, B.E., Friedman, N.P., Banich, M.T., 2015. Resting-state networks predict individual differences in common and specific aspects of executive function. Neuroimage 104, 69–78. https://doi.org/10.1016/j.neuroimage.2014.09.045.

Sheline, Y.I., Raichle, M.E., 2013. Resting state functional connectivity in preclinical Alzheimer's disease. Biol. Psychiatr. 74, 340–347. https://doi.org/10.1016/j.biopsych.2012.11.028.

Stevenson, C., Brookes, M., López, J.D., Troebinger, L., Mattout, J., Penny, W., Morris, P., Hillebrand, A., Henson, R., Barnes, G.R., 2014. Does function fit structure? A ground truth for non-invasive neuroimaging. Neuroimage 94, 89–95. https://doi.org/10.1016/j.neuroimage.2014.02.033.

Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Cam-CAN, Henson, R.N., 2017. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. Neuroimage 144, 262–269. https://doi.org/10.1016/j.neuroimage.2015.09.018.

Tessitore, A., Amboni, M., Esposito, F., Russo, A., Picillo, M., Marcuccio, L., Pellecchia, M.T., Vitale, C., Cirillo, M., Tedeschi, G., Barone, P., 2012. Resting-state brain connectivity in patients with Parkinson's disease and freezing of gait. Park. Relat. Disord. 18, 781–787. https://doi.org/10.1016/j.parkreldis.2012.03.018.

Troebinger, L., López, J.D., Lutti, A., Bestmann, S., Barnes, G., 2014a. Discrimination of cortical laminae using MEG. Neuroimage 102, 885–893. https://doi.org/10.1016/j.neuroimage.2014.07.015.

Troebinger, L., López, J.D., Lutti, A., Bradbury, D., Bestmann, S., Barnes, G.R., 2014b. High precision anatomy for MEG. Neuroimage 86, 583–591. https://doi.org/10.1016/j.neuroimage.2013.07.065.

Van Veen, B.D., Van Drongelen, W., Yuchtman, M., Suzuki, A., 1997. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans. Biomed. Eng. 44, 867–880. https://doi.org/10.1109/10.623056.

Venkataraman, A., Whitford, T.J., Westin, C.-F., Golland, P., Kubicki, M., 2012. Whole brain resting state functional connectivity abnormalities in schizophrenia. Schizophr. Res. 139, 7–12. https://doi.org/10.1016/j.schres.2012.04.021.

Wackermann, J., Lehmann, D., Michel, C.M., Strik, W.K., 1993. Adaptive segmentation of spontaneous EEG map series into spatially defined microstates. Int. J. Psychophysiol. 14, 269–283. https://doi.org/10.1016/0167-8760(93)90041-M.

Weiskopf, N., Suckling, J., Williams, G., Correia, M.,M.M., Inkster, B., Tait, R., Ooi, C., Bullmore, T.,E.T., Lutti, A., 2013. Quantitative multi-parameter mapping of R1, PD*, MT, and R2* at 3T: a multi-center validation. Front. Neurosci. 7, 1–11. https://doi.org/10.3389/fnins.2013.00095.

Wipf, D., Nagarajan, S., 2009. A unified Bayesian framework for MEG/EEG source imaging. Neuroimage 44, 947–966. https://doi.org/10.1016/j.neuroimage.2008.02.059.

Woolrich, M., Baker, A.P., Luckhoo, H., Mohseni, H., Barnes, G.R., Brookes, M., Rezek, L., 2013. Dynamic state allocation for MEG source reconstruction. Neuroimage 77, 77–92. https://doi.org/10.1016/j.neuroimage.2013.03.036.

Wu, J., Srinivasan, R., Kaur, A., Cramer, S.C., 2014. Resting-state cortical connectivity predicts motor skill acquisition. Neuroimage 91, 84–90. https://doi.org/10.1016/j.neuroimage.2014.01.026.

Wurina, Zang, Y.-F., Zhao, S.-G., 2012. Resting-state fMRI studies in epilepsy. Neurosci. Bull. 28, 449–455. https://doi.org/10.1007/s12264-012-1255-1.