RELATING AUTISM OCCURRENCE TO FUNCTIONAL CONNECTIVITY IN RESTING-STATE MEG (2014 ADA REPORT)

JERZY WIECZOREK

Advanced Data Analysis (ADA) advisory committee:

- Advisor: Dr. Rob Kass, CMU, Department of Statistics
  kass@stat.cmu.edu
- Sponsor: Dr. Avniel Ghuman, UPMC, Brain Mapping Center
  ghumana@upmc.edu

1. Introduction

Many neuroscientists believe that the presence and severity of autism is related to impaired connectivity in the social brain. We use a novel combination of study design, brain imaging technology, and functional connectivity measures to examine differences in functional connectivity between subjects with Autism Spectrum Disorder (ASD) and Typically Developing (TD) subjects.

This exploratory report evaluates whether our novel study design has sufficient power to address the scientific questions of interest reliably. With the small available sample size, the large number of parameters generated by the all-to-all functional connectivity measure, and the use of noisy resting-state recordings rather than task-based recordings, we have low study power and our results are not statistically significant. We also attempt to avoid data fishing by only running a limited number of tests and by using appropriate multiple-comparisons corrections.

Nonetheless, our point estimates are compatible with the scientific hypothesis that functional connectivity is stronger in TD than ASD subjects, particularly in the social brain. If similar estimates and patterns held in a larger study, it would provide evidence in favor of this theory.

In Section 2, we explain the scientific terms and brain imaging technologies underlying our study. Our scientific and statistical hypotheses are stated in Section 3. Section 4 describes the data and the pre-processing we performed, while the statistical analyses of interest are presented in Section 5. Results are discussed in Section 6.

The appendix (Section A) summarizes data-access challenges that led us to change directions from our initial planned ADA topic.

2. Scientific background

Our study explores neurological aspects of Autism Spectrum Disorder (ASD), sometimes simply referred to as as Autism (Section 2.1).

Date: December 30, 2016.
Advanced Data Analysis project, 2014.

We have reason to believe this disorder is related to impairments in an aspect of brain activity known as functional connectivity (Section 2.2). Prior work suggests that functional connectivity is lower among ASD than TD subjects, especially in the spatial domain of the social brain (Section 2.3) and in the frequency domain of the alpha band oscillations (Section 2.4).

In order to study functional connectivity in our subjects, we measure their brain activity using magnetoencephalography (MEG) (Section 2.5). MEG is a non-invasive brain imaging technology with fine temporal resolution. We also use structural information about each subject's brain from magnetic resonance imaging (MRI) recordings (Section 2.6).

Of the many ways to define functional connectivity, we choose to use a metric developed by Ghuman et al. (2011) that is based on the phase-locking values of Lachaux et al. (1999) (Section 2.7).

2.1. **Autism Spectrum Disorder.** In revising its *Diagnostic and Statistical Manual of Mental Disorders* from the 4th to the 5th edition (DSM-5), the American Psychiatric Association grouped several disorders (including autistic disorder and Asperger's disorder) into a single diagnosis of Autism Spectrum Disorder (American Psychiatric Association, 2013a). "The essential features of autism spectrum disorder are persistent impairment in reciprocal social communication and social interaction [...] and restricted, repetitive patterns of behavior, interests, or activities [...] These symptoms are present from early childhood and limit or impair everyday functioning," and prevalence worldwide is estimated at around 1% of the population (American Psychiatric Association, 2013b, pp. 53-55).

Despite its frequency and possible severity, the causes and mechanisms of ASD are not yet well understood. Unlike some neurological disorders (such as Broca's aphasia, triggered by damage to Broca's area in the brain), ASD does not have a single, clear, localized root. Differences between ASD and TD subjects appear to occur throughout the brain.

The subjects in our study either have been diagnosed with autism or a related disorder, forming the ASD group, or have been diagnosed with no psychiatric disorder, forming the TD group.

2.2. **Functional connectivity.** "Structural connectivity" refers to physical, anatomical links between brain regions. "Functional connectivity," on the other hand, refers to related patterns of activity between brain regions: which regions tend to synchronize their activity, and under what conditions? Functionally-connected regions need not be structurally-connected and vice versa.

Many different measures of functional connectivity are in use, with different strengths and weaknesses depending on the researcher's goals. Most are essentially statistical measures of association between two time series, including correlation, coherence, Granger causality, etc. Our study uses a measure based on the phase-locking value of Lachaux et al. (1999), described in Section 2.7.

2.3. **The social brain.** The left side of Figure 7, from Gotts et al. (2012), highlights a set of brain regions including the temporo-parietal junction; the posterior superior temporal sulcus; and the fusiform gyrus, among others. This collection of brain areas is known to co-activate commonly across a range of social tasks and has been termed the "social brain." Prior studies have suggested that ASD appears

to be associated with impairments in the social brain, such as weaker functional connectivity (Gotts et al., 2012) or lower cortical thickness (Wallace et al., 2010).

2.4. **Alpha band oscillations.** When brain activity is recorded with a technology such as MEG, which measures the average activity of many neurons at once, synchronized spiking of groups of neurons can manifest as oscillations on the recorded time series. These oscillations may occur at different frequencies, and several frequency bands occur commonly enough to have been given names in the neuroscience literature. Oscillations in the 8-12 Hz frequency band (Purdon et al., 2013) are known as "alpha band" activity or "alpha waves." Alpha waves have been shown to be associated with eye opening and visual stimulation (Redlich et al., 1946) as well as with attention, sleep, and consciousness, for instance during anaesthesia (Purdon et al., 2013).

In previous exploratory data analyses, Dr. Ghuman has found that global differences between ASD and TD patients' functional connectivity are particularly large in the alpha band. For the present study, we compute functional connectivity using a wavelet-based measure (Section 2.7) centered at a single alpha-band frequency, 11 Hz, where these estimated differences are greatest, as shown in Figure 1.

This figure and the choice of 11 Hz are based on the same dataset used in this study. We acknowledge a possible danger of overestimating any effects in the data, since we have chosen to use the frequency at which those effects are maximized. However, as it turns out, the study lacks enough power to achieve statistical significance even after selecting for the largest effect size here. Furthermore, this effect itself is not statistically significant. For each patient, we compute a global average interhemispheric functional connectivity score, averaged over all locations in the brain. We perform a two-sided, two-sample, unequal variances t-test on these global averages, testing whether the difference between the ASD and TD lines at 11 Hz is nonzero in Figure 1. The result ($p = 0.08$, 95% CI (-0.002, 0.028), df = 33) is not statistically significant.

2.5. **MEG.** Magnetoencephalography (MEG) is a non-invasive neuroimaging tool with fine temporal resolution. MEG is used to study brain activity primarily in the cortex (the outer layer of neural tissue). Neurons deeper within the brain are not necessarily aligned in a common orientation, and they are farther from the MEG sensors, so their generated magnetic fields are too weak to measure.

To record the miniscule magnetic fields produced by currents in the cortex, MEG uses a helment containing an array of powerful, sensitive, supercooled magnets in a shielded room. MEG measurements are recorded in Teslas, a unit of magnetic flux density, usually on the order of 10 femtoTeslas (fT). MEG readings can be taken on the order of every millisecond, while some other common brain imaging techniques, such as functional magnetic resonance imaging (fMRI), may take several seconds per reading. A tool like MEG is necessary in order to study millisecond-scale patterns in brain activity, on the same order of magnitude as the individual action potentials that form the basis of electrical activity in the brain.

On the other hand, the brain contains billions of neurons while MEG has only around 300 sensors, arranged in a helmet placed outside of the skull. The problem of inferring activity inside the brain from these 300 sensors is a "big p, small n" problem and an interesting statistical challenge (Section 4.2.3).

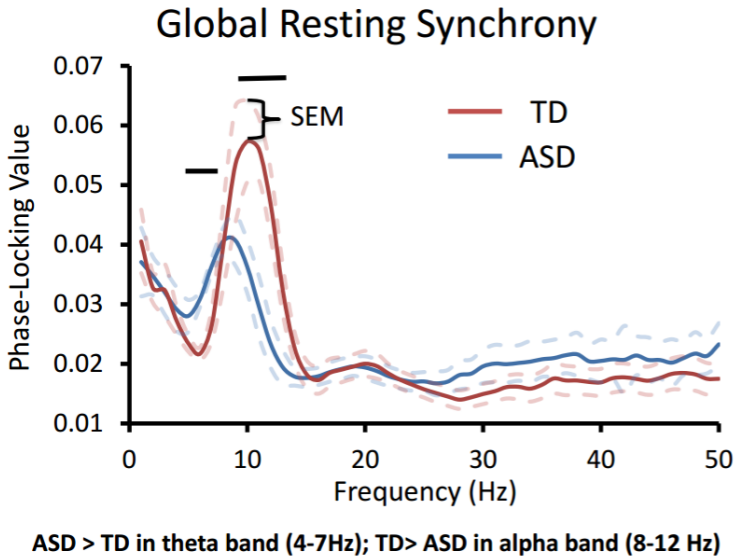

FIGURE 1. Average interhemispheric phase-locking values of ASD and TD patients over a range of frequencies from 1 to 50 Hz. Plot courtesy of Dr. Ghuman.

There are additional concerns related to data cleaning and pre-processing for removing the effects of magnetic fields outside the MEG room, as well as data artifacts caused by heartbeats, eye blinks, etc. In this study, we have simply applied standard data cleaning procedures (Section 4.2.1).

2.6. **MRI.** Magnetic resonance imaging (MRI) is an anatomical imaging technology used in neuroscience and other areas of medicine. Using powerful magnetic fields and radio waves, MRI can produce visual images (with fine spatial resolution) of 2D slices through a patient's body. In our study, MRI images of each subject's skull and brain were taken in order to model these structures, so that the sensor-space MEG recordings (localized in the MEG helmet) could be translated back into source-space estimates (localized on the subject's cortex).

2.7. **PLV/wavelet-based functional connectivity.** From Lachaux et al. (1999) we have the concept of phase-locking value (PLV), a measure of the functional connectivity at a given frequency between two locations in the brain. Our setting differs from theirs in that they compute PLV over trials at each time point, whereas we compute PLV over time. We model the sources of MEG-recorded brain activity as magnetic dipoles in the brain, with a single resting-state time series of activity at each dipole (location). For a given subject and a frequency of interest f, we can compute the PLV over time between two dipoles d_1 and d_2 as follows:

At each dipole d_j separately (for $j = 1, 2$), we take the (pre-processed and cleaned) time series of MEG measurements there; convolve it with a complex Morlet wavelet centered at frequency f; and normalize the amplitude, obtaining a time

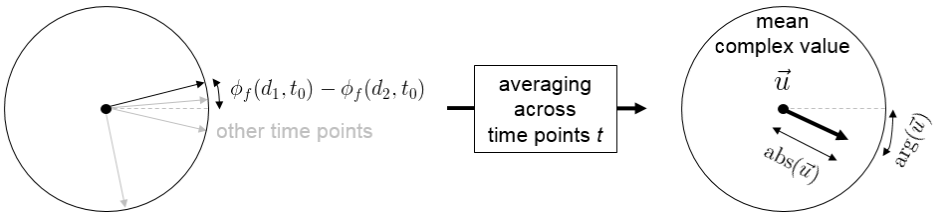

FIGURE 2. Illustration of the concept of phase-locking values (PLV). The PLV is $\text{abs}(\vec{u})$, the magnitude of the mean complex vector $\vec{u}$. Plot adapted from Lachaux et al. (1999).

series of phase values $\phi_f(d_j, t)$. Then the PLV is a transformation of phase differences, averaged over time:

$$\text{PLV}_f(d_1, d_2) = \frac{1}{T}\left|\sum_{t=1}^{T} \exp\left(i[\phi_f(d_1, t) - \phi_f(d_2, t)]\right)\right| \in [0, 1]$$

where T is the number of time points recorded. Figure 2 illustrates the interpretation of this metric. At each time point t, we represent the phase difference (between these two dipoles, at this time) as a complex unit vector. Next, we average these unit vectors over time and find the length of this average vector. If the phase differences are constant over time, then the unit vectors will all be identical, so their average will be a unit vector too and the PLV will be 1. (As the name "phase-locking value" suggests, this metric is highest when the dipoles stay at a constant phase difference, i.e. the phases are "locked.") On the other hand, if the phase differences vary wildly over time, then the unit vectors will point in all directions, their average over time will be a short vector near the origin, and the PLV will be close to 0.

As a measure of functional connectivity, we interpret that two dipoles with a high PLV are consistently firing in phase at that frequency. As Lachaux et al. (1999) argue, wavelet-based methods like PLV can be applied to nonstationary signals, which is not necessarily true of some other common functional connectivity measures such as simple correlation or coherence. Futhermore, correlation can give a misleading indication of weak connectivity between two dipoles which are in phase but lagged, so again PLV is a more suitable measure. Another benefit of PLV is that it separates the effects of phase from the effects of amplitude. Phase-locking is sufficient for concluding that two dipoles are functionally connected, so there is no need to conflate phase-locking and amplitude (as coherence does).

Ghuman et al. (2011) use adjusted PLVs to compute an all-to-all connectivity matrix between every pair of dipoles. Their adjustment is to subtract off the PLV computed on empty-room noise (i.e. a recording using the same MEG apparatus with no participant inside and projected onto a standard brain). They find that this successfully accounts for the spurious phase-locking that results from estimating each source dipole's data using a common set of MEG sensors. Every source estimate draws from every sensor, causing an artifact of false-positive phase-locking (especially strong between a dipole and its closest neighbors). By subtracting the

two all-to-all connectivity matrices (real-subject data minus empty-room data), they arrive at a matrix of adjusted PLVs which avoids this spurious phase-locking.

However, this adjustment over-corrects for pairs of dipoles that are spatially close. To step around this over-correction, our statistical analyses in Section 5 only look at interhemispheric PLVs (i.e. between two dipoles that are in separate hemispheres of the brain). Also, the resulting adjusted PLVs are no longer strictly in the range $[0, 1]$. We simply interpret any negative values as showing negligible functional connectivity compared to the level of noise in the data.

In the present study, we use the method of Ghuman et al. (2011) to construct an all-to-all functional connectivity matrix for all dipole pairs, for each subject, at every integer frequency from 1 to 50 Hz. Since we model the brain using 5124 dipoles, each of these is a 5124×5124 symmetric matrix of (adjusted) PLVs for each subject in the study. However, as described above, we end up reducing this to a smaller matrix containing only the interhemispheric PLVs; and we only use the 11 Hz matrix, in which the average difference between TD and ASD is most pronounced.

3. From scientific to statistical questions

Our scientific hypothesis is that functional connectivity in social brain regions is stronger among TD than ASD subjects. We translate this into two primary statistical questions:

- Globally, do we have evidence that connectivity patterns are significantly different for ASD vs. TD?
- Locally, where in the brain is connectivity significantly different for ASD vs. TD?

Under our scientific hypothesis, we hope that the answers will be "Yes" and "In the social brain," respectively.

We frame this study as a hypothesis testing problem. Using the study design and methods described here, do we have enough data to convincingly demonstrate an association between connectivity patterns and ASD vs. TD status, and hence to provide evidence about our scientific hypothesis?

Concretely, our present work differs from other studies addressing this scientific question in that we use:

- resting-state recordings, not task-based;
- MEG, not another brain-imaging technology such as fMRI;
- activity in the alpha band, not at all frequencies;
- PLV-based functional connectivity, not another measure such as correlation or coherence; and
- all-to-all connectivity, i.e. comparing each dipole to all others, not just evaluating the connectivity between a small number of prespecified seed regions.

The decision to use resting-state data introduces substantial noise, as opposed to task-based data (where subjects perform a certain task repeatedly and the recordings can be averaged over these repeated trials to reduce noise). Additionally, the choice of all-to-all connectivity leads to a large number of variables (big p, small n) and many possible multiple comparisons, making inference more challenging. In

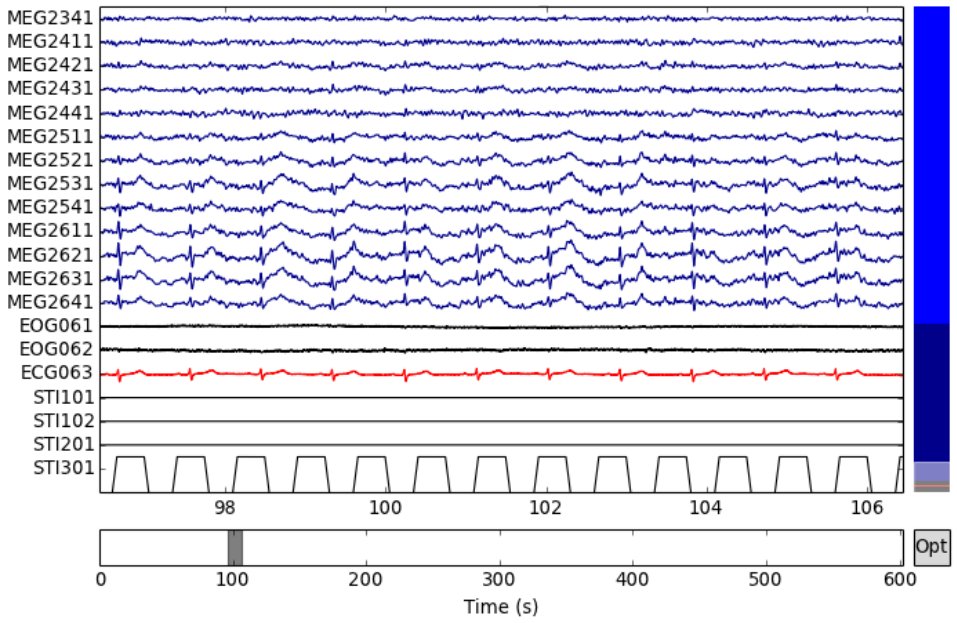

FIGURE 3. Time series plot of several MEG channels from one patient.

light of these decisions, it is unsurprising that our analyses ultimately found no statistically significant associations.

Section 4 details how we pre-process and summarize the data, in order to make it more tractable to address these questions via the analyses described in Section 5.

4. The data

4.1. **Data structure.** The data consists of resting-state MEG recordings from 17 ASD and 18 TD subjects, recorded at UPMC's Brain Mapping Center and partially pre-processed by members of Dr. Ghuman's lab. We also have structural MRI data for each subject, as well as empty-room MEG recordings taken for calibration shortly before or after each "real" recording.

For each subject, the data include 313 channels recorded over 10 minutes per patient. The 600 Hz recordings were downsampled to 150 Hz, for a total of 90,000 time points per recording. The 313 channels include 306 MEG channels; four stimulus channels (although there were no stimuli presented); two ocular channels; and one cardiac channel. The latter three channels allow correcting for the effects of eye movement and heartbeat.

In order to depict the data structure, Figure 3 depicts a section of one patient's MEG data. This figure shows a small section of the time series for a subset of the MEG channels, along with the eye-blink EOG channels, the cardiac ECG channel, and the reference stimulus STI channels. This figure illustrates how the cardiac activity is reflected in several MEG channels as well, so we must account for its effects as part of our data cleaning.

4.2. **Pre-processing workflow.** As detailed below, the data pre-processing had the following goals: remove noise in the raw data; combine the MEG data with the

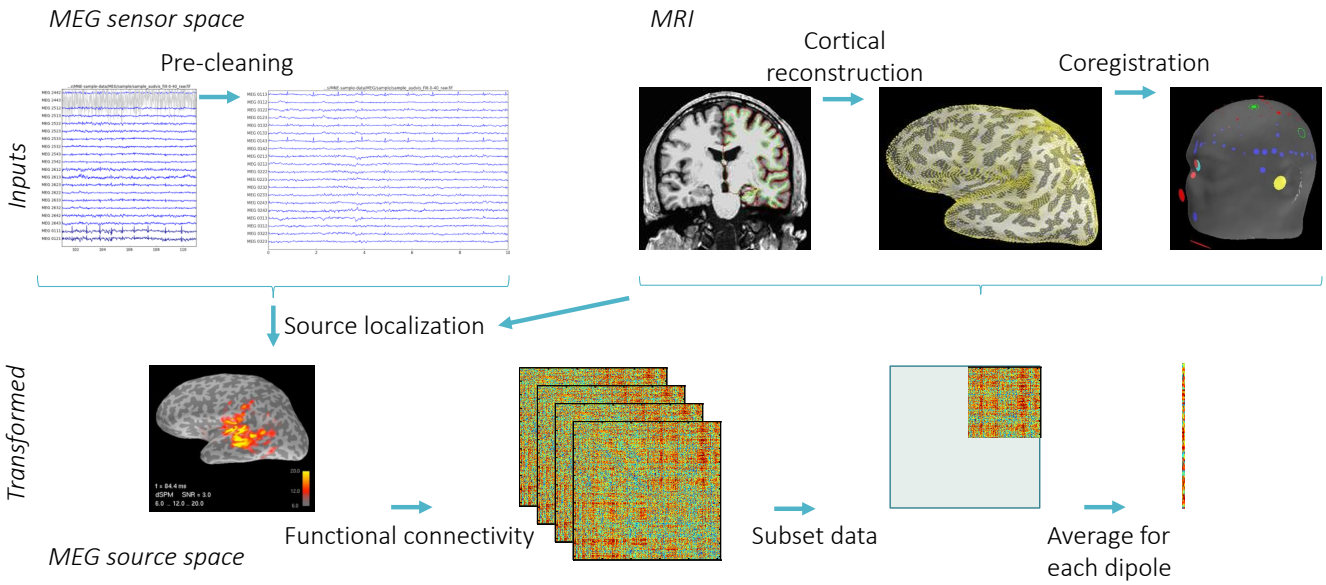

FIGURE 4. Pre-processing workflow.

MRI-derived spatial structure of the patient's brain; compute measures of functional connectivity between different locations; and simplify the resulting data down to a tractable level for statistical analysis.

Figure 4 summarizes the workflow, described in detail in Sections 4.2.1 through 4.2.5. In Figure 5, we overlay the data dimensions (per subject) at each step of the workflow to help illustrate how each step transforms the size of the dataset.

In practice, the first pre-processing stages were performed by members of Dr. Ghuman's lab. For the dataset actually used in this study, the author only performed the final stage, in Section 4.2.5, as well as all the statistical analyses in Section 5. However, the author did practice performing the earlier pre-processing stages on another dataset, as described in Appendix A.

4.2.1. *MEG sensor space.* The MEG data are in FIF file format, which can be read and analyzed with the open-source software MNE-C and MNE-Python (Gramfort et al., 2013).

We perform standard data cleaning steps to remove noise and artifacts. These steps consist of band-pass filtering, temporal signal space separation (TSSS), and signal-space projection (SSP).

Band-pass filtering for the range 1 to 50 Hz removes low-frequency drift and high-frequency noise, including the 60 Hz noise from power lines.

TSSS uses the known structure of the MEG sensor array and the basic physics of Maxwell's equations in order to decompose the MEG signal into two fields: arising inside vs. arising outside the helmet. By keeping only the signal from inside the helmet, we remove artifacts due to empty-room noise.

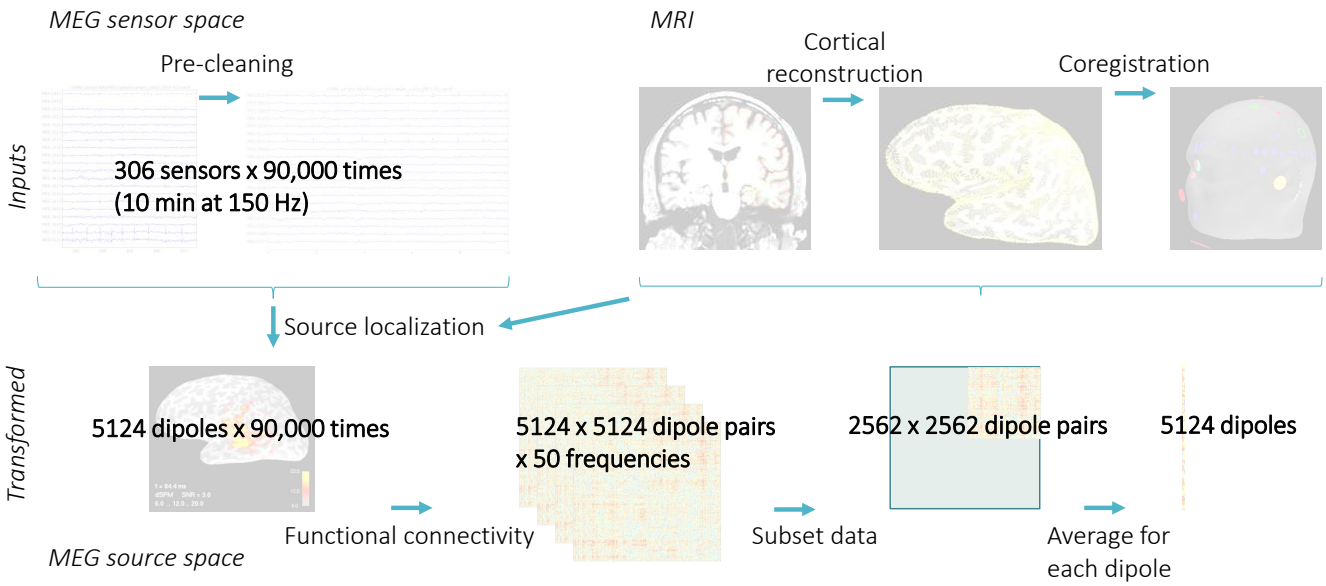

FIGURE 5. Pre-processing workflow, overlaid with dataset dimensions (per subject) at each step.

SSP removes ocular and cardiac artifacts by performing principal components analysis (PCA) on the EOG and ECG channels, computing a noise subspace which is then removed from the MEG data.

4.2.2. *MRI.* The MRI pre-processing converts each subject's 2D MRI images into a triangulated 3D mesh of their brain and skull structure. The patients' structural MRIs are analyzed with another open-source software tool, FreeSurfer$^{\text{TM}}$, in a step known as "cortical reconstruction." By measuring and modeling the shape of the patient's brain and skull, we can transform data from the sensor to the source space.

There is also a "coregistration" step, in which this 3D cortical reconstruction is aligned with the subject's head location and position inside the MEG helmet.

All together, using the reconstructed shape and location (relative to the MEG sensors) of the cortex and skull as anatomical constraints, we compute the gain matrix X described below in Section 4.2.3.

4.2.3. *Source localization.* This step transforms each of the 90,000 time points from the MEG sensor space of 306 sensors to the source space of 5124 dipoles.

The standard method for source localization is "minimum-norm estimation" (MNE), which can be seen as a set of L^2-norm regularized regressions or ridge regressions, computed separately at each moment in time. If we knew the current at a given point in the brain, we could use the MRI-derived skull shape and basic physics to evaluate how that current would be measured by each MEG sensor after passing through the skull. Instead, we have the inverse problem: estimating the distribution of currents in the brain, using the observed currents measured by the MEG sensors.

We model sources in the brain as a large set of small magnetic dipoles, one dipole per voxel (volume pixel), with the voxels arranged in a triangulated mesh over the cortical surface. A separate time series of activity is estimated at each dipole. At each moment, the current at a dipole has both an orientation and a magnitude. In our model, each dipole's orientation remains constant (perpendicular to the cortical surface at that location), and only the magnitude changes over time.

We model the cortical surface as a commonly-used triangulation of 5124 vertices (5120 triangles), in which the source locations are about 6 mm apart on average. This triangulation is achieved by subdividing an icosahedron repeatedly. The chosen level of subdivision balances the need for a manageably-small number of dipoles without overly-large dipole sizes. It would not be scientifically meaningful to average together much larger clusters of neurons.

In other words, we partition the cortex into several thousand (oriented) dipoles and estimate the current strength in each voxel at each time point. At a given time point t, let β_t denote a vector of these current strengths, and let y_t be a vector of the MEG sensor measurements. Using the structural MRI, we create a "gain matrix" X deterministically to indicate how the currents at each dipole are transformed before being measured at the MEG sensors. We assume that $y_t \approx X\beta_t$, plus or minus noise in the measurements and variability in the currents themselves, which we can account for via estimated noise covariance matrices C_y and C_β, respectively. Since β_t is a much longer vector than y_t (5124 dipoles vs. 306 sensors), β_t is not estimable using ordinary linear regression. Instead, we estimate β_t via a regularized regression that shrinks elements of β_t towards 0, weighted appropriately by the estimated noise covariance matrices:

$$\hat{\beta}_t = \arg\min_{\beta} \left[(y_t - X\beta)^T C_y^{-1} (y_t - X\beta) + \beta^T C_\beta^{-1} \beta \right]$$

Alternative source localization methods might use L^1-norm regularization (as in "minimum-current estimation" or MCE), or may impose additional restrictions on covariance matrix estimation, encourage smoothness in the estimates over time, etc. However, we saw no reason not to use standard MNE in this study.

Finally, the source-localized estimates are projected once more, by morphing them from the subject's actual brain onto the corresponding anatomical regions of a "standard" brain. Morphing all subjects' data onto a common space allows us to make inter-subject comparisons at any given location on the cortex.

We acknowledge that this process is imperfect. Due to natural variability in the structure of human brains and the small dipole sizes used, a given dipole in one subject will not be perfectly comparable with "the corresponding dipole" in the standard brain. We try to compensate for this problem by using statistical analyses in Section 5 that make broad comparisons over sets of dipoles, rather than assigning importance to individual dipoles.

4.2.4. *Functional connectivity.* For each of the 5124×5124 pairs of dipoles, we compute the PLV/wavelet-based functional connectivity centered at each integer frequency between 1 and 50 Hz, as described in Section 2.7. In light of Figure 1, we later retain only the 11 Hz data for statistical analysis.

Furthermore, we subset the data to one off-diagonal quadrant of this symmetric connectivity matrix. This leaves us with connectivities for the 2562×2562 inter-hemispheric dipole pairs. We remove the intrahemispheric dipole pairs because of

the empty-room correction as described in Section 2.7. With no correction, nearby dipoles will naturally have similar time series and high PLVs, simply due to the process of estimating many dipoles using relatively few MEG sensors. The empty-room adjustment attenuates the estimated PLV between nearby dipoles, which corrects the autocorrelation at moderate and large distances, but overcorrects at small distances. Hence, by only looking at interhemispheric PLVs, we only compare distant pairs of dipoles and avoid being misled by this overcorrection.

4.2.5. *Averaging.* In the interest of simplicity, we reduce each subject's data further by averaging each row and each column of their subset (interhemispheric) connectivity matrix. This gives us a single connectivity score for each dipole. This score represents that dipole's average connectivity with all dipoles in the other hemisphere.

A low average score (near 0) suggests that this dipole does not show strong functional connectivity with the other hemisphere. A high average score (near 1) would indicate that this dipole does shows strong connectivity with the other hemisphere. A moderate score likely indicates one of two possibilities: either moderate connectivity with many other dipoles, or very strong connectivity with only a few dipoles.

5. Statistical analysis

At the end of this pre-processing, we have a dataset ready for statistical analysis. For each subject, we have a vector of 5124 average interhemispheric adjusted-PLV functional connectivity scores (one score per dipole); and we have a label for each subject (TD or ASD).

From here, we could take two possible approaches: prediction/classification or hypothesis testing. In the first approach, we could use the PLV scores as covariates for predicting TD vs. ASD status, e.g. using logistic regression or other classification algorithms. However, under this approach, we would still want to make statistical inferences regarding our scientific question of where in the brain functional connectivity differs between TD and ASD subjects. This would be challenging: usual significance testing of regression coefficients is difficult under the penalized regression or classification methods needed for this "big p, small n" scenario, and cross-validation would be too noisy with such a small number of participants.

Hence, we follow the second approach: we simply evaluate whether the study has enough power to distinguish the TD and ASD groups, either at individual dipoles or under global summary statistics. Section 5.1 summarizes our point estimates, while the following sections summarize our statistical inferences.

In Section 5.2, we explore the use of multiple-comparisons-corrected tests on individual dipoles. Furthermore, for any statistic we compute to compare the two groups, we can also randomly permute the subjects' labels (TD or ASD) and recompute the statistic repeatedly to simulate a permutation distribution. We conduct several such permutation tests: first in Section 5.3 using only the values of the PLV scores, and then in Section 5.4 also accounting for the spatial arrangement of the dipoles.

5.1. **Exploratory data analysis.** We compute a vector of two-sample, unequal-variances t-statistics for the difference in PLVs between groups (TD minus ASD). The t-scores are computed separately for each dipole.

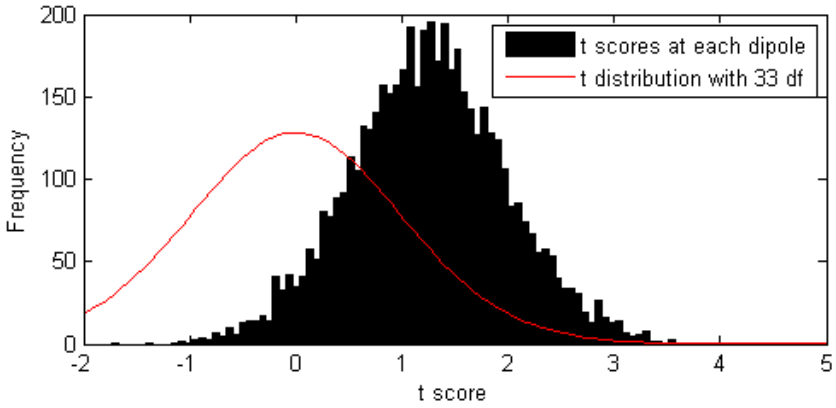

FIGURE 6. Histogram of t-scores (from each dipole) for the difference TD minus ASD, computed on averaged interhemispheric adjusted-PLV values. t(df=33) density overlaid for reference.

Figure 6 shows the histogram of these t-scores. Nearly all t-scores are positive, so our point estimates do not tend to contradict the scientific hypothesis that connectivity is higher in TD subjects.

We can also show these t-scores arranged spatially by each dipole's location on the cortex. The right side of Figure 7 shows the spatial hotspots in our estimated t-statistics. The plot only shows dipoles with absolute t-scores above 2, plotted with 20 steps of spatial smoothing. All plotted dipoles are red or yellow, indicating that the TD score was greater than the ASD score for all such dipoles. (If any dipoles had had much higher t-scores for ASD than TD, they would have been blue, but no such effects occurred in the data.)

The left side of the figure (copied from Gotts et al., 2012) shows the social brain regions. The left figure shows a folded brain while the right side (our data) is plotted on an inflated brain. Although this may make comparison difficult for a nonexpert, the apparent hotspots in our data on the right do indeed match up with social brain areas on the left, particularly in the parietal and temporal cortex, posterior superior temporal sulcus, and fusiform gyrus.

5.2. **Separate tests with multiple comparisons corrections.** We have a statistical summary of the dataset: a vector of 5124 t-statistics for the difference (TD minus ASD) in average interhemispheric connectivity scores. If our scientific hypothesis holds and our study has sufficient power, at least some of these connectivity scores should be significantly different from zero.

The histogram of these t-scores in Figure 6 clearly has a positive sample mean, and many individual scores are above the usual t-score cutoff of approximately 2. However, we are implicitly carrying out 5124 distinct t-tests, so we need to correct for multiple comparisons.

A popular and useful multiple-comparisons approach is to set a target False Discovery Rate (FDR). However, the standard FDR approach assumes independence between the tests, which is not appropriate here. We know these t-statistics are

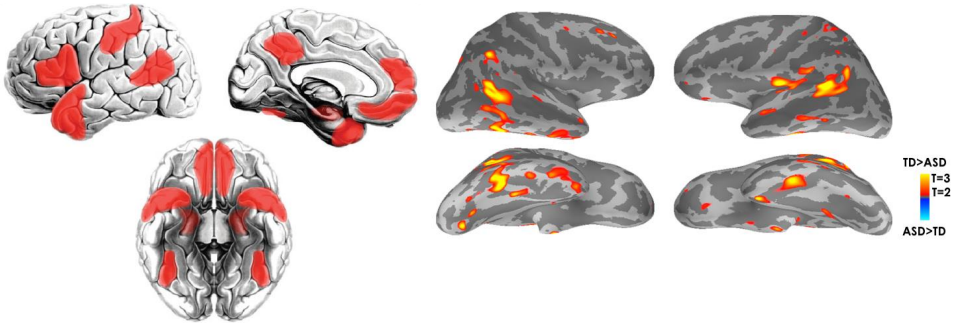

FIGURE 7. Left figure: social brain (copied from Figure 1 of Gotts et al., 2012). Right figure: spatial locations of dipoles with highest t-score estimates in our data.

not independent: neighboring dipoles are likely to have similar values of average interhemispheric connectivity, and hence to have positively correlated t-statistics.

Instead, we can perform a Bonferroni correction, which does not assume independence. For a usual family-wide error rate (FWER) of $\alpha_{FWER} = 0.05$, we compare each dipole's t-score to a Bonferroni-corrected critical value at $\alpha_{Bonf} = 0.05/5124 \approx 9.8 \times 10^{-6}$. However, the Bonferroni correction is known to be conservative, and it turns out to be underpowered here. For a two-sided test with degrees of freedom $n_1 + n_2 - 2 = 33$, the critical value is $t_{33,\ \alpha_{Bonf}/2} \approx 5.22$. None of our dipoles have absolute t-scores beyond this critical value (our largest was 4.46), so no dipoles are significant under this testing approach.

We see that, although many individual dipoles are unusual under the null t-distribution, none are sufficiently large to remain significant after multiple comparisons. Still, the histogram of all 5124 t-scores clearly does not match the null t-distribution. Consequently, we next seek significance through permutation tests on the set of 5124 t-statistics.

5.3. **Non-spatial permutation tests.** Our first permutation test is an informal visual judgment of whether the histogram of all 5124 t-scores, taken as a whole, appears unusual compared to histograms from the null permutation distribution.

The concept is a "visual hypothesis test," inspired by the `nullabor` R package (Wickham et al., 2011). We plot a "lineup" of several null histograms of t-scores from the permutation distribution. The real histogram of t-scores is also plotted, but in a random location among the null plots in this lineup. The use of 19 null plots and one real plot is roughly akin to testing with a significance level of $\alpha = 0.05$.

If the real histogram clearly stands out from the others, it would suggest that the null permutation distribution is not adequate for the real data and that there may be a real difference between the groups. However, if the real dataset does not dramatically differ from the null plots, then the null permutation distribution does appear to describe the real data adequately. Any difference between the groups is negligible compared to the noise in the data.

Figure 8 shows such a lineup of one real and 19 null histograms. We have overlaid N(0,1) distributions, simply to help guide the eye in comparing locations and scales of the histograms. Clearly none of these histograms actually match a N(0,1), presumably due to autocorrelation among neighboring dipoles.

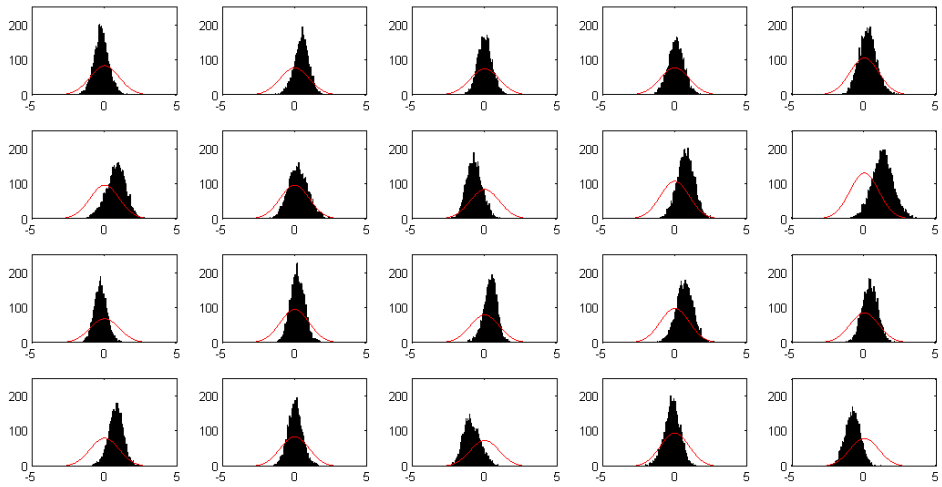

FIGURE 8. Visual hypothesis test in the spirit of Wickham et al. (2011). One histogram (identified in the text) plots the real data, while the others are drawn from the null permutation distribution. All histograms have N(0,1) density overlaid for reference.

We find that this informal visual test is inconclusive. The real dataset's histogram is probably the "most unusual," in that it is most dissimilar to the reference N(0,1) curve, but its difference from the null histograms is not dramatic. (The real data are in the right-most column, second row down.)

For a more objective evaluation, we run permutation tests for several summary statistics computed on the real and null histograms. Without certainty about what statistic would give the most powerful test, we try several plausible statistics:

- How many t-scores are "large"?
  (proportion of absolute t-values that are above 2)
- How "massive" is the set of the largest t-values?
  (sum of top 10% of absolute t-values)
- How "different" is the histogram from a reference distribution?
  (Kolmogorov-Smirnov test statistic (maximum absolute difference in CDFs) between the t-scores' empirical CDF and a N(0,1) CDF)

The K-S test-statistic-based permutation test gave a p-value of 0.055, while the other two tests' p-values were above 0.1. Although we could continue trying other summary statistics, this would run the risk of fishing for significance. Consequently, we cannot conclude that this set of dipoles is significantly unusual under the permutation distribution. We lack sufficient power to distinguish TD and ASD by the set of average interhemispheric PLVs alone.

5.4. **Spatial clustering permutation tests.** However, the histogram permutation tests above neglect to use one important piece of information:which dipoles are spatial neighbors. We can use this information in a spatial clustering permutation test, in the spirit of Maris & Oostenveld (2007) and Xu et al. (2011).

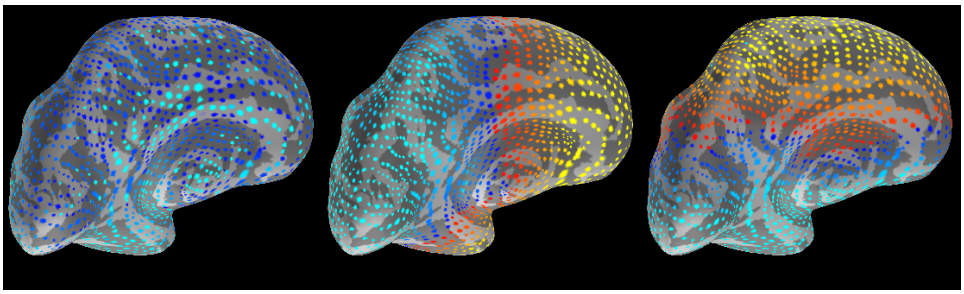

FIGURE 9. From left to right: The values of x, y, and z coordinates in our matrix of dipole locations, plotted on the left brain hemisphere. The color gradient in each subfigure runs from light blue through dark blue to red to yellow. The y (back to front) and z (low to high) coordinates are ordered correctly, but the x (left to right) coordinates are scrambled.

A spatial clustering test draws from the same permutation distribution as before, but then also accounts for spatial adjacency between dipoles. A common type of test statistic finds all the spatial clusters (spatially-contiguous sets of dipoles whose t-values are "large enough") and computes the size or mass of the largest such cluster. If the largest spatial cluster in the real data is larger than most such clusters in permutation data, we can reject the permutation null distribution as an implausible description of our real data and conclude that we have evidence of a real difference between groups.

For our test statistic, we chose to consider only dipoles whose absolute t-scores are above 2. For these dipoles, we find every connected spatial cluster, count the number of dipoles in that cluster, and record the size of the largest cluster.

Although we have written functioning MATLAB code for conducting the spatial clustering test, unfortunately, we could not obtain an accurate adjacency matrix (indicating which dipoles are each other's neighbors) from the MNE software or find its representation in the stored data. We attempted to compute adjacencies directly from a matrix of (x, y, z) coordinates provided by Dr. Ghuman's lab, but we discovered that the x-coordinates are scrambled, as shown in Figure 9. When we plot the y or z coordinate values (middle and right subplots) as colors on the brain surface, they form a smooth gradient in the correct direction. However, the x coordinates (left subplot) are scrambled relative to the order in which they are plotted. With no dipole ID to use for matching, we cannot align the coordinates with the data in order to create an adjacency matrix for spatial clustering.

6. Discussion

As we see in the statistical analyses of Section 5, the t-scores for functional connectivity differences between ASD and TD are not statistically different from noise. Neither multiple-comparisons-corrected tests on each dipole, nor global tests on the vector of t-scores, nor a global test on the spatial clustering of t-scores can distinguish our results from chance. Therefore, we must answer our first question of Section 3 in the negative: our study lacks power to conclude that functional connectivity levels are significantly different in ASD and TD.

As a consequence, it is also unsafe to use the spatial map of estimated t-scores to infer where t-scores might be significantly different if only we had a larger sample size. Hence we are unable to answer our second question of Section 3: we cannot locate where functional connectivity is significantly different for ASD vs. TD.

We could certainly have explored more modern methods for estimating an empirical null distribution (Efron, 2010, Chapter 6) or tried using other test statistic variants for our spatial and non-spatial clustering permutation tests. However, we fear that any more testing at this point would just be significance fishing.

Rather than seek out more powerful tests, it may be more productive to consider what sample size we would have needed to attain significance. Roughly speaking, we can imagine a dataset with more subjects but the same within-group sample means and sample variances of average PLVs at each dipole. For every dipole j, we would have the same values of $\bar{x}_{j(TD)}, \bar{x}_{j(ASD)}, s^2_{j(TD)}, s^2_{j(ASD)}$, but computed from a larger n. In this case, as n increases, the standard error estimate decreases and each dipole's t-score

$$t_j = \left(\bar{x}_{j(TD)} - \bar{x}_{j(ASD)}\right) / \sqrt{s^2_{j(TD)}/n + s^2_{j(ASD)}/n}$$

becomes more extreme. The t-distribution's degrees of freedom also rise and the critical value becomes smaller. We calculate that at $n = 23$ subjects per group, or 46 total subjects, the largest absolute t-score in our dataset would reach 5.09 and surpass the new Bonferroni-adjusted critical value of $t_{44,\ 0.05/(2 \cdot 5124)} \approx 5.00$.

This is actually quite heartening: a moderate increase of 5 or 6 subjects per group may be enough to achieve multiple-comparisons-corrected statistical significance, at least at some dipoles. Power for permutation tests is harder to study, but the K-S permutation test of Section 5.3 already had a p-value near 0.05 and may also have attained significance with a slightly larger sample.

We could likely achieve more decisive results with an even larger, but still manageable, increase in sample size. In a comparison of frequentist and Bayesian hypothesis testing, Johnson (2013) illustrates how doubling the sample size can be more than enough to power a simple study for significance threshold $\alpha = 0.005$ instead of the traditional $\alpha = 0.05$. The lower frequentist threshold corresponds to substantially more-convincing evidence on standard Bayes-factor scales.

If the apparent patterns from our small sample did indeed hold in a large sample, then we would have evidence in favor of both our hypotheses. Our t-score estimates are mostly large and positive, so we would conclude that TD has higher average connectivity than ASD. Also, the apparent spatial clusters of highest t-scores are largely in social-brain areas, as the scientific hypothesis predicts.

Alternatively, instead of increasing the sample size, we can consider changing the study design and methodology. Although we did not have sufficient power when using all-to-all connectivity, we may have more success by averaging over larger areas. By grouping dipoles *a priori* in larger brain regions (known anatomical and functional structures of the brain) and computing a single average t-score in each region, we would have fewer comparisons for which to correct. We would also reduce the noise that comes from morphing each subject's dipoles to a common brain at such a fine spatial scale. Alternatively, by grouping dipoles *a priori* into two groups (social-brain regions vs. all other brain regions) and testing just these two groups of dipoles, we would likely be able to draw statistically significant conclusions with the existing sample size. Finally, using trial-based recordings instead of resting-state

data would allow us to average over repeated trials and reduce the noise in our time series. On the other hand, this would also limit the generalizability of the study, in that we could only draw conclusions about TD and ASD differences on such trials, not on general resting-state activity.

In conclusion: Although our tests failed to reach statistical significance, it does appear that all-to-all functional connectivity in resting-state MEG data is a promising approach. It should not require too many more subjects before the sample size is sufficiently large to draw at least some limited conclusions. Alternatively, our neuroscientific questions could be addressed productively through less ambitious methods, by reducing either the noise levels or the number of comparisons.

7. References

American Psychiatric Association (2013a) "Autism spectrum disorder fact sheet," Washington, DC.

American Psychiatric Association (2013b) *Diagnostic and statistical manual of mental disorders* (5th ed.), Washington, DC.

Antony, A.R., Alexopoulos, A.V., Gonzalez-Martinez, J.A., et al. (2013) "Functional Connectivity Estimated from intracranial EEG Predicts Surgical Outcome in Intractable Temporal Lobe Epilepsy." *PLoS ONE*, vol. 8, no. 10.

Efron, B. (2010) *Large Scale Inference*, Cambridge Uni. Press, Cambridge, UK.

Ghuman, A., McDaniel, J., & Martin, A. (2011) "A Wavelet-based Method for Measuring the Oscillatory Dynamics of Resting-State Functional Connectivity." *NeuroImage*, vol. 56, no. 1.

Gotts, S.J., Simmons, W.K., Milbury, L.A., et al. (2012) "Fractionation of social brain circuits in autism spectrum disorders." *Brain*, vol. 135.

Gramfort, A., Luessi, M., Larson, E., et al. (2013) "MEG and EEG data analysis with MNE-Python." *Frontiers in Neuroscience*, vol. 27, no. 267.

Johnson, V. (2013) "Revised standards for statistical evidence." *Proc Natl Acad Sci USA*, vol. 110, no. 48.

Lachaux, J-P., Rodriguez, E., Martinerie, J., & Varela, F.J. (1999) "Measuring Phase Synchrony in Brain Signals." *Human Brain Mapping*, vol. 8.

Maris, E. & Oostenveld, R. (2007) "Nonparametric statistical testing of EEG- and MEG-data." *Journal of Neuroscience Methods*, vol. 164.

Purdon, P.L., Pierce, E.T., Mukamel, E.A., et al. (2013) "Electroencephalogram signatures of loss and recovery of consciousness from propofol." *Proc Natl Acad Sci USA*, vol. 110, no. 12.

Redlich, F.C., Callahan, A., & Mendelson, R.H. (1946) "Electroencephalographic changes after eye opening and visual stimulation." *Yale J Biol Med*, vol. 18, no. 5.

Wallace, G.L., Dankner, N., Kenworthy, L., Giedd, J.N., & Martin, A. (2010) "Age-related temporal and parietal cortical thinning in autism spectrum disorders." *Brain*, vol. 133.

Wickham, H., Cook, D., Roy Chowdhury, N., & Hofmann, H. (2011) "An introduction to the `nullabor` package."
http://cran.r-project.org/web/packages/nullabor/vignettes/nullabor.html

Xu, Y., Sudre, G.P., Wang, W., Weber, D.J., & Kass, R.E. (2011) "Characterizing global statistical significance of spatiotemporal hot spots in magnetoencephalography/electroencephalography source space via excursion algorithms." *Statistics in Medicine*, vol. 30.

Appendix A. ADA project chronology and changes in scope

In January 2014, the original plan for this ADA project was to study epilepsy, not autism. Patients whose epilepsy does not respond to medication are candidates for surgery to remove the seizure focus. However, some proportion of surgical patients will have undesirable post-surgery side effects and/or will experience no reduction in seizures. There is currently no conclusive way to predict which patients will or will not benefit from surgery. We intended to evaluate the use of several resting-state MEG-based measures of functional connectivity to predict surgical outcomes, using records from 20-30 temporal lobe epilepsy (TLE) patients at UPMC.

However, we were unable to obtain the complete data required for the intented epilepsy project. Each major component of the data (MEG, MRI, and neuropsychological testing records) was managed by a different data owner and was de-identified for privacy protection. We were unable to get all of each subject's components as well as the unique subject IDs needed to link each subject's records correctly.

We spent the first portion of the year working with a small ($n = 12$ subjects) MEG and MRI dataset, without the patients' neuropsychological records or surgical outcomes. During this time we learned to perform data pre-processing and work with MEG data using Freesurfer and MNE-Python; became comfortable with Python programming; and studied the science behind epilepsy and MEG.

In May and June we received more MEG records as well as a neuropsychological outcomes dataset (pre- and post-surgery testing of each patient's memory, verbal skills, etc. to allow us to evaluate the surgery's side effects), but without the IDs needed for linking them together. In the following months we performed marginal exploratory data analyses on these outcomes; implemented the correlation-based functional connectivity measure of Antony et al. (2013); and practiced working with the wavelet-based measure of Ghuman et al. (2011). Over the summer we also observed an MEG experiment and met with a clinician to learn how MEG imaging is read and used in practice. Nonetheless, we could not continue the analysis without knowing which MEG records linked to which outcomes.

By October, with no further progress towards linking the records from different data owners, we agreed to switch projects. The new autism project was chosen largely for convenience due to the late date in the ADA cycle. Like the original epilepsy project, it still involved the analysis of functional connectivity measures from resting-state MEG data, which meant that most of our learning and work throughout the year was still relevant. However, the autism dataset had already been combined and pre-processed. This saved considerable time and effort and made it possible to complete a reasonable analysis by the December deadline. On the other hand (without having the raw, unprocessed version of the data), we were limited in the kind of analyses we could do. Without being able to split the MEG timeseries into exploration and validation datasets, and with such a small sample size, we had to be especially careful to rein in our exploratory analyses to avoid data fishing. This led to the restrained scope of our final analysis.

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA
E-mail address: jerzy@stat.cmu.edu
URL: http://www.stat.cmu.edu/~jwieczor/