INVESTIGATING RELIABILITY AS A TOOL FOR PATIENT-SPECIFIC PIPELINE SELECTION

by

Sarah Christine McLeod

Submitted in partial fulfilment of the requirements
for the degree of Master of Applied Science

at

Dalhousie University
Halifax, Nova Scotia
July 2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Introduction: Functional neuroimaging plays a key role in pre-surgical planning and improving surgical outcomes. My team has developed ROCr (Receiver Operating Characteristic reliability) - a tool to assess intra-session reliability of functional neuroimaging data. **Objective:** Investigate ROCr as a tool for automated, patient-specific analysis. **Methods:** Ten subjects underwent three MEG (magnetoencephalography) imaging sessions of right MNS (median nerve stimulation). Twelve pipelines were used to produce activation maps for each dataset, and ROCr was used to assess the reliability of these maps. Various tests were done on the results to assess ROCr. **Results and Discussion:** Within subjects, the pipeline with the highest reliability varied across days, likely because the inter-session data variance was larger than assumed. Using the pipeline with highest reliability resulted in lower inter-session variability. Further analysis of the outcome measures gave new insight into how to improve use of ROCr in future studies and clinical applications.

LIST OF ABBREVIATIONS AND SYMBOLS USED

AFNI	Analysis of functional neuroimages
ANOVA	Analysis of variance
AUC	Area under the curve
BOLD	Blood oxygen level-dependent
B_z	Magnetic field
DC	Direct current
DCS	Direct cortical stimulation
ECD	Equivalent current dipole
ECG	Electrocardiogram
ED	Euclidean distance
EEG	Electroencephalography
EOG	Electrooculogram
e_z	Unit vector along the axis of the sensor
fMRI	Functional magnetic resonance imaging
FN	False negative
FP	False positive
FPR	False positive rate
FR	Reliable fraction
HP	High pass
HPF	High pass filter
HSD	Honestly significant difference
ICA	Independent component analysis
LCMV	Linearly constrained minimal variance
LP	Low pass
LPF	Low pass filter
MC	Motion correction
MEG	Magnetoencephalography

MNS	Median nerve stimulation
NPAIRS	Nonparametric prediction, activation, influence and reproducibility resampling
PSP	Post-synaptic potential
$\vec{Q}$	current dipole at rQ
$\vec{r}$	vector from origin to point where the magnetic field is computed
ROC	Receiver-operator characteristic
ROCr	Receiver-operator characteristic reliability
$\overrightarrow{r_Q}$	vector from origin to the current dipole
S1	Primary somatosensory cortex
S2	Secondary somatosensory cortex
SEF	Somatosensory evoked field
SSS	Signal space separation
t_1	Threshold 1
t_2	Threshold 2
TN	True negative
TP	True positive
TPR	True positive rate
tSSS	Temporal signal space separation
χ^2	Chi squared

ACKNOWLEDGEMENTS

I would like to thank Tim Bardouille and Steven Beyea for being excellent thesis supervisors, and for all the help and knowledge they've given me over the past two years. Steven was my entry point into Biomedical Engineering and the BIOTIC lab group, and I am especially grateful for his help with reviewing my work from the beginning to the end of my degree. I am also very grateful to Tim for his support and mentoring throughout the program, and helping me grow as a scientist.

I would also like to thank my committee members: Aaron Newman, Jeremy Brown, and Rob Adamson. Not only did they give excellent feedback on my work, I also had the opportunity to learn from each of them in a classroom setting. I would also like to thank everyone at BIOTIC, especially Alexander Rudiuk and Ron Bishop for their help with data analysis and collection.

I am very grateful to all my friends who have supported me over the past two years. Kristin and Rishima, having you two by my side for this journey means more than I could ever put into words. I hope we can continue to journey together. For fame and glory! Jena, you've been here for me through everything that has led to this point. Thank you for all the support through the good times and bad.

Finally, I would like to thank my family for their unconditional love and support. Mum and Dad, thank you for all that you've done for me, for always believing in me and putting me on the road to success. I wouldn't be where I am today without you.

CHAPTER 1 INTRODUCTION

1.1 Pre-Surgical Mapping

Pre-surgical functional brain mapping allows clinicians to identify so-called "areas of eloquence" in patients who are candidates for brain surgery. It is essential in these patients, with drug-resistant epilepsy or a brain tumour, to *identify* and *preserve* areas that are critical for eloquence in daily functions such as movement and speech, resulting in better surgical outcomes. While there are many similarities in the organization of the human brain between individuals, every brain is different. This is especially true in patients whose brains may have reorganized in the presence of a longstanding tumour or lesion (1). Pre-surgical mapping can be performed invasively or non-invasively.

1.1.1 Invasive Mapping

Invasive methods, such as the Wada test and direct cortical stimulation (DCS) are the clinical gold standard techniques. In the Wada test, a patient's language and memory functions are tested behaviourally following administration of an anaesthetic agent, which "shuts down" an entire hemisphere of the brain. The Wada test can only determine laterality of areas of eloquence, not their locations (2). In DCS, intracranial electrodes are surgically implanted, and these electrodes are activated to excite or inhibit brain activity. Areas of eloquence underlying implanted electrodes are inferred based on the behavioural response (e.g., muscle twitch or speech arrest) (3).

While very effective for pre-surgical functional mapping, the invasive nature of the Wada test and DCS pose challenges and risks such as infection and stroke. The Wada test is an invasive arterial procedure. A 2008 study by Loddenkemper et al. looking at 677 patients who underwent the Wada test found that 11% of patients were affected by complications. The most common complications were encephalopathy, seizures, and strokes (4). DCS requires either a craniotomy or burr holes to place electrode sheets or wires on the brain. A recent review by Enatsu et al. on DCS found that the complication rate ranged between 6% and 26%. Serious complications included transient cerebrospinal fluid leakage, infection, intracranial bleeding, and infarction (3). Both procedures require significant patient cooperation, and may not be usable by certain patient populations such as children and those with cognitive or neurological defects.

1.1.2 Non-Invasive Mapping

Non-invasive methods, such as functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG), can greatly reduce challenges and risks associated with invasive mapping techniques (5–8). Because they are non-invasive, all of the complications discussed above are eliminated. The procedures are far more patient-friendly and are available to more patient populations. Non-invasive methods are already being used clinically for pre-surgical mapping. A 2014 article by Pittau et al. discusses a variety of non-invasive techniques that can be used in pre-surgical planning for patients with epilepsy, including fMRI and MEG (9). These techniques can also be used for pre-surgical mapping before tumour resection (10). A significant advantage of non-invasive methods is that they can be done before the patient reaches the operating room. The surgical team can get the information they need well ahead of the surgery, giving them time to discuss the results and consider options as a team. In contrast, if an invasive method such as DCS is used, the team will not be aware of any abnormalities until the surgery is already in progress.

The challenge with non-invasive methods is proving that they are equivalent to the invasive clinical standard methods in terms of clinical efficacy. Importantly, these non-invasive methods must have high accuracy and low inter-session variability (11) to be seen as equivalent to invasive methods. Accuracy can be demonstrated by comparing the non-invasive results to clinical standard results. For example, the accuracy of MEG source localization can be determined by comparing the results to DCS results done during surgery. Inter-session variability can be demonstrated by repeating the procedure and seeing how similar the estimated locations are.

For functional localization applications, accuracy is a measure of how close the predicted values are to the actual location. Inter-session variability is a measure of how close the predicted values are to each other. Figure 1 demonstrates how accuracy and inter-session variability are defined mathematically. P indicates the known location (green dots), and Mindicates an estimated location (blue dots). M_{avg}indicates the average of the estimated locations (orange dot).

Accuracy is defined as the sum of the differences between each estimated location and the known location as shown in Equation (1):

$$Accuracy = \sum_{i} \left|\vec{M}_i - \vec{P}\right| \qquad \text{Equation (1)}$$

Inter-session variability is defined as the sum of the differences between each estimated location and the average estimated location as shown in Equation (2):

$$Inter - Session\ Variability = \sum_{i} \left|\vec{M}_i - \vec{M}_{avg}\right| \qquad \text{Equation (2)}$$

In scalar notation, inter-session variability can be defined as the mean of the Euclidean distances between each estimate and the average estimate. The Euclidean distance between two points (x_1, y_1, z_1) and (x_2, y_2, z_2) is defined as shown below in Equation (3):

$$Euclidean\ Distance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \qquad \text{Equation (3)}$$

Figure 1 a) demonstrates an example of poor accuracy and poor inter-session variability. The estimated locations are far from the average estimated location, as well as the known location. In Figure 1 b), the estimated points are very close to the average estimated location, but are quite far from the known location. This example has good inter-session variability, but the accuracy is poor.

Accuracy is important in pre-surgical planning because resecting the wrong area of the brain can have a huge negative impact on the surgery outcome. However, accuracy can only be determined using invasive methods. Inter-session variability, however, is a quality of the data itself and can be calculated using non-invasive data. This makes it an attractive QA metric for pre-surgical planning.

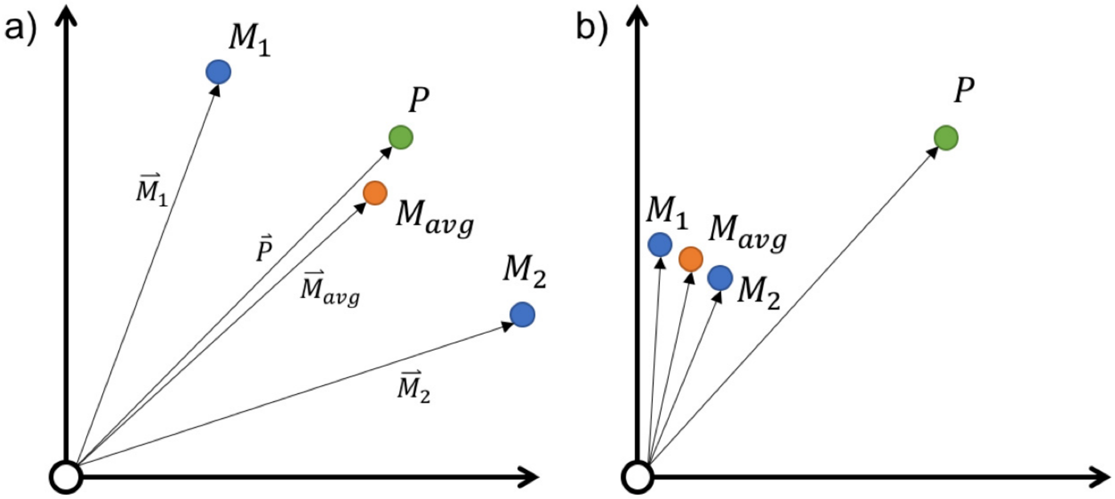


Figure 1: Accuracy and inter-session variability.

In each plot, the green dot (P) represents a known location. The blue dots (M) indicate estimated locations. The orange dots (M_{avg}) indicate the average of the estimated locations. (a) This example has poor inter-session variability because the distance between each estimate and the average estimate is large. (b) This example has good inter-session variability because each estimate is close to the average. Both examples have poor accuracy because the distance between each estimate and the known point is large.

1.2 Magnetoencephalography (MEG)

1.2.1 What is MEG?

MEG is a non-invasive, functional brain imaging technique that measures the magnetic fields generated by brain activity (12). Signals in the brain are transmitted by action potentials along a neuron's axon to synapse on other neurons, leading to modulations of the membrane potential. Action potentials are very transient, with a duration of only 1 ms. Post-synaptic potentials (PSPs) are modulations in the membrane potential that occur in the dendrites of a neuron following the arrival of an action potential at the synapse. Sufficient summation of PSPs will cause the neuron to generate an action potential, which is the basis of neural communication. The electric potential generated by a PSP resembles "a current dipole oriented along the dendrite", where a current dipole is defined as current flow over distance, as the distance approaches zero (12). Although these potentials have lower peak amplitudes than action potentials, they have a much longer duration on the order of tens of milliseconds. Because of this, PSPs have better temporal summation with neighbouring neurons than action potentials.

When neurons are aligned and fire synchronously, the combined PSPs generate a dipolar electric potential that can be detected at the scalp. It is this activity that is measured by EEG (electroencephalography). Electric potentials are distorted by the skull, limiting the spatial resolution of EEG, but magnetic fields are largely unaffected by different tissues. The electric potential generated by PSPs has a corresponding magnetic field, which can be detected by the MEG system; a set of special, cryogenically-cooled SQUID (superconducting quantum interference device) sensors connected to simple wire loops upon which a current flow is induced when a magnetic field change occurs. Field change can be measured with sub-millisecond temporal resolution. The sensors are placed in a "helmet" that the patient's head fits under.

Equation (4) below describes the magnetic field measured at a sensor that is normal to a spherical volume conductor.

$$B_z = \frac{\mu_0}{4\pi}\frac{\vec{Q}\times(\vec{r}-\overrightarrow{r_Q})\cdot\overrightarrow{e_z}}{\left|\vec{r}-\overrightarrow{r_Q}\right|^3} \qquad \text{Equation (4)}$$

Where:

$B_z = magnetic\ field$

$\mu_0 = permeability\ of\ free\ space$

$\vec{Q} = current\ dipole\ at\ r_Q$

$\vec{r} = vector\ from\ origin\ to\ the\ point\ where\ the\ field\ is\ computed$

$\overrightarrow{r_Q} = vector\ from\ origin\ to\ the\ the\ current\ dipole$

$\overrightarrow{e_z} = unit\ vector\ along\ the\ axis\ of\ the\ sensor$

This equation has two main implications for MEG, as discussed by Hämäläinen et al (12). First, MEG is most sensitive to neurons that are tangential to the skull, and least sensitive to axons perpendicular to the skull. As seen in the equation above, the magnetic field generated by a dipole is oriented perpendicular to the current flow. A radial current flow generates a tangential field (relative to the skull), which cannot be properly detected by the sensors near the scalp. Second, because the magnetic field strength falls off as a function of distance squared, MEG is (generally) more sensitive to neuronal activity near the skull and less sensitive to activity originating deep in the brain.

MEG is a much more direct measure of brain activity than fMRI, which measures the BOLD (blood oxygen level-dependent) signal. Via neurovascular coupling, the neural response results in an increase in the oxygen level near the area of the neural response. The BOLD response has excellent spatial resolution, but the physiology of this response limits the temporal resolution of fMRI to the order of seconds (13). MEG, on the other hand, has very high temporal resolution, capable of measuring over 1000 samples per second (12). The neural response to a stimulus can be detected directly using MEG as shown below in Figure 2. This means that MEG can record rapidly changing signals, and more importantly it can detect very small changes in the timing of brain activity.

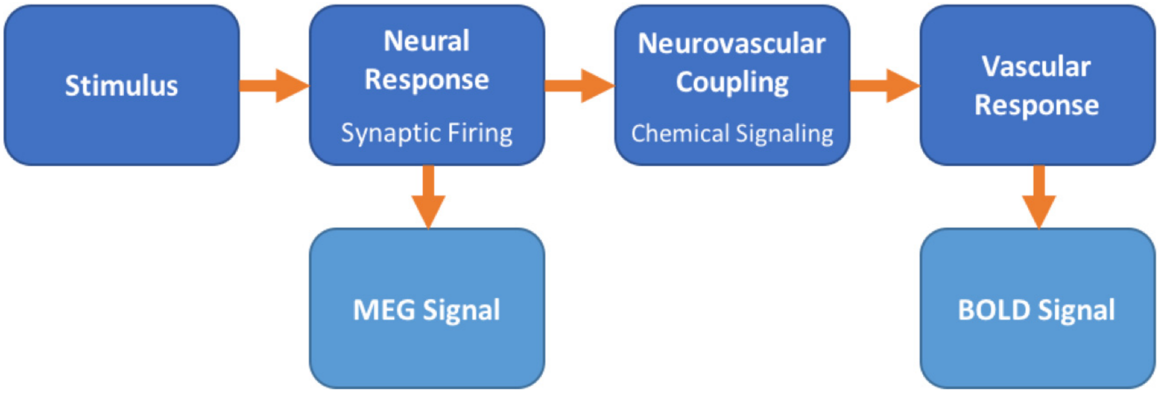


Figure 2: MEG Signal vs. BOLD Signal.
The stimulus is administered, resulting in a neural response. The neural response generates magnetic fields which can be detected using MEG. The neural response also generates a vascular response via neurovascular coupling. This coupling involves chemical signalling. The vascular response produces the BOLD signal, which is measured by fMRI. The extra physiological steps required to generate the signal that fMRI can measure reduces fMRI temporal resolution compared to MEG.

A typical MEG study follows a basic structure. The subject sits or lies with their head in the scanner, where they perform a task, or perceive stimulus from a set, a number of times. The onset of task performance or stimulus defines an “event” of interest. The changing magnetic fields generated by the brain during the task are recorded continuously, including time prior to and following these events. An additional channel records the timing of the stimulus and/or response. Once the scan is complete, the data are analyzed offline to estimate the timing and/or location of the currents that generated the measured magnetic field deflections.

The collected data can be analyzed in a variety of ways. Most MEG studies, including the majority of clinical paradigms and the data used for my thesis, use an analysis “pipeline” (i.e., set of analysis algorithms and parameters) that includes event-related analysis and source localization. Event-related analysis isolates magnetic field changes that have a high temporal correlation to relevant events. This can be achieved, for example, by averaging the magnetic field time-course across multiple repetitions of the event to attenuate uncorrelated magnetic field deflections. Source localization interpolates the activity occurring at locations inside the brain using the signals measured at the sensor array. MEG data are registered to the patient’s anatomical MRI. As such, MEG sources can be visualized on the patient’s brain anatomy to assist in pre-surgical planning.

1.2.2 Clinical Application: Pre-Surgical Mapping of Primary Somatosensory Cortex

MEG has applications in research and clinical settings. As a key example for my work, the primary somatosensory cortex (S1) is an area commonly located in pre-surgical planning to ensure that the patient’s sense of touch is not disrupted by the surgery. Several studies, such as those by Castello et al. and Korvenoja et al. have validated MEG as a non-invasive method to locate S1 by comparing the source localization results to DCS localization done during surgery (14,15). These studies achieve localization via repetitive median nerve stimulation (MNS). MNS involves applying a percutaneous electrical stimulus to the wrist over the median nerve.

Median nerve stimulation activates primary afferent pathways at the wrist to produce a somatosensory evoked field (SEF) at the MEG sensor array. The cortical components of the SEF peak first in contralateral S1, followed by a weaker bilateral response in secondary somatosensory cortex (S2). The response in contralateral S2 is slightly stronger than the response in ipsilateral S2 (16). Source localization of the SEF response in contralateral S1 provides the pre-surgical map utilized by the surgical team. MNS "provides robust localization of the early SEF response in single subjects. In particular, the MNS SEF has been shown to contain three distinct peaks at 20, 35, and 60 ms" (17).

In clinical practice, the SEF is recorded using MEG, and the data is analyzed with a specific analysis pipeline to interpolate the location of S1. After collection and preprocessing, the data is separated into smaller data segments ("trials") aligned to the time of onset of each stimulus event, and an inter-trial average magnetic field map is generated, which is the SEF. Clinically, source localization is completed using the single equivalent current dipole method (11). Because the SEF is generated by a very focal source, the equivalent dipole method works well for source localization. Additional work by others, such as Cheyne et al. shows that the beamformer method also works well for localizing the primary somatosensory cortex (17). Additionally, Solomon et al. have demonstrated that this SEF paradigm and analysis pipeline gives very consistent results across multiple sessions, with across-session differences as low as 4-8mm (11).

1.2.3 Analysis Considerations and Options

All of the processes applied to MEG data following collection until the final output are referred to as an *analysis pipeline*. There are many parameters within the pipeline that can be modified, yielding slightly different final results as shown in Figure 3. This is an advantage and disadvantage of MEG source localization. The disadvantage with multiple results is that, short of using DCS to verify the accuracy of each result, it is difficult to know which result to use. The advantage of many processing options is that it allows for individual analysis. Every brain is unique, and a "one size fits all" approach to source localization may not be optimal. The challenge is determining which result will give the best surgical outcome for a given individual.

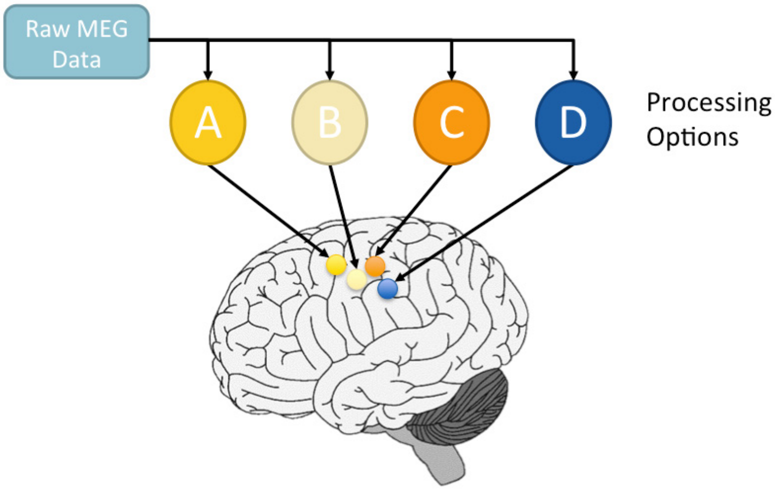


Figure 3: One dataset can produce many different results.

Circles A through D represent different sets of source localization processing options (pipelines) that can be applied to the raw data. Each pipeline gives a slightly different result. The challenge is determining which pipeline to trust.

Five important parameter categories in MEG data analysis pipelines are environmental noise reduction, head movement compensation, frequency filtering, artefact removal (commonly using independent component analysis; ICA) and source estimation techniques.

1.2.3.1 Environmental Noise Reduction

Magnetic shielding is a hardware solution to help prevent noise sources outside the scanning room from contaminating the data, but the shielding is not perfect. Even within the scanning room, noise can be generated by patient movement and metal objects. Temporal signal space separation (tSSS) is a software-based method of further reducing this environmental noise. tSSS is a temporal extension of signal-space separation (SSS), and uses temporal and spatial relations between signals to detect and remove sources of environmental noise (18).

1.2.3.2 Head Movement Compensation

Motion correction is a tool that can be used to improve data quality when head movement is present during the scan. If the subject's head moves with respect to the magnetic field sensors, the sources of brain activity will be in a different location in relation to the sensors, which can reduce source localization accuracy. If, during the scan, the head has translated more than 5mm, or rotated more than 3 degrees, the data may not be fit to use. The MaxFilter motion correction used in this study uses data from the HPI coils to determine head movement throughout the scan, and realigns the sensor data to the reference head position during the tSSS calculation (19).

1.2.3.3 Frequency Filtering

Frequency filtering suppresses unwanted frequencies from a signal. Despite best efforts to use hardware and software to reduce noise in recorded MEG data, some noise will always be present. One way to remove noise is by applying a frequency filter to the data. Neuromagnetic signals contain frequency components up to several hundred Hz. However, not all frequencies will be relevant for a given response of interest. Furthermore, magnetic signals will be present in the recording that are not generated by the brain; some at frequencies that are not relevant for the response of interest. If we know what frequency or frequencies a source of noise has, the data can be filtered to remove these frequencies. The problem is that sources of noise and signals of interest may have overlapping frequencies, so there is a trade-off between reducing noise and preserving the signal of interest.

The two types of filters I will use in my pipelines are high-pass (HP) filters and low-pass (LP) Butterworth filters. High-pass filters suppress frequencies below the cut-off frequency. This type of filter is useful for removing noise caused by sensor drift or removing DC offset from the data. Low-pass filters suppress frequencies above the cut-off frequency. This type of filter is useful for removing high-frequency noise sources such as power lines, communication devices, and muscle activity. Most brain activity of clinical interest occurs at frequencies below 70 Hz, so it is common in MEG data analysis to LP filter the data at 70 Hz or lower. Depending on the analysis being done, the cut-off frequency can be varied. Filter settings can be selected *a priori* based on expectations for signal and noise, or based on spectral analysis of the data. Filtering options include whether or not high- or low-pass filters are used, and the cut-off frequencies of these filters.

1.2.3.4 Artefact Removal

Independent component analysis (ICA) is a tool used to separate independent time series (also referred to as *sources*) in a dataset. A common example of an application for ICA is three microphones recording in a room where a three-instrument band is playing. Each microphone will pick up the sound from all three instruments, but with varying sensitivity. ICA is able to reconstruct three sources that improve the separation of the three instruments from the original recordings. At this point we can reject one or more of the sources (for example, the drums) and recombine the other two sources (guitar and vocals) to the mixed recordings. A key benefit of ICA is that it allows preservation of the sources of interest while rejecting sources of noise, instead of rejecting a whole segment of the data. For MEG data, important sources to separate could be brain activity and artefacts such as blinks. The output of ICA is a set of time-varying sources that are linearly combined to create the original dataset.

There are three main ways to reject components using ICA in MEG data analysis. The first is to set a rejection threshold, and any sources with signals over the threshold are rejected. This method works on the assumption that sources of noise such as blinks and other motion artefacts are much stronger than the brain activity being measured. The second method is to reject sources based on similarity to a known noise time-series. It is common for electrophysiological signals to contaminate MEG, so if a source is highly temporally correlated with a concurrently recorded electrooculogram (EOG) and electrocardiogram (ECG) signal, it can be rejected. The final method is manual investigation of the sources, rejecting those that appear to be noise. The MEG signal is then reconstructed by linearly combining sources that were not rejected. ICA can be done on raw or epoched data (20).

1.2.3.5 Data Epoching

The next step in analysis is epoching the data around the times when the MNS was delivered. This step involves separating the data into data segments ("trials") synchronized such that time zero is the onset of the stimulus in each trial. When the SEF is evoked via MNS, other brain activity occurs at the same time, and the sensors also pick up this activity. However, only the SEF is correlated with the stimulus onset. Epoching the data, towards inter-trial averaging, is an essential step in separating the signal of interest (correlated with the stimulus) from the other brain signals (not correlated with the stimulus). Inter-trial averaging is discussed in the following section. The epoching step has two parameters that can be modified: the length of the baseline (pre-stimulus interval) and how long after the stimulus to include. Each epoch is a 2-D matrix: channels by time. The dataset as a whole is a 3-D matrix: trials (epochs) by channels by time. Each epoched data segment is independent and can be separated temporally, which would not be the case for a similar study done using fMRI. This means that we can sample a subset of the epochs for comparison, or split the epochs into groups.

1.2.3.6 Data Averaging by Condition

Once the data has been epoched, the next step is averaging together the epochs across trials. Each channel's time course is averaged together across the epochs, resulting in a 2-D matrix of channels by time, producing single waveforms per channel for each condition. As discussed above, there are many unwanted signals present in the data that are not correlated with the stimulus onset. By averaging the data across trials, signals correlated to the stimulus (e.g., SEF) are preserved, but signals that are uncorrelated to the stimulus are suppressed.

1.2.3.7 Source Localization

Source localization is the process of estimating where in the brain activity is occurring based on evoked field data measured at the scalp. The first step in this process is calculating the forward solution. The forward solution represents the effects of the brain, skull and scalp on the magnetic fields generated by a current dipole of known location and orientation. It determines what the magnetic fields at the sensor array would look like as the result of known dipole sources on the cortex (20) using Equation (4) above. This equation is calculated at each sensor location for hypothetical current dipoles located at each location in the 3D brain space.

The next step is calculating the noise and data covariances. This step involves calculating the covariance between each sensor and every other sensor, over the pre-stimulus or post-stimulus window, giving an estimate of covariance of just the noise, and the noise plus signal of interest.

The final step is the source estimation. A common method is the equivalent dipole method, but Stevens et al. have shown that the beamformer method, which generates a volumetric activation map, can also be used for source localization using SEF data (21). For a given location in the brain, the beamformer method generates a data-driven weighting function to apply to the MEG data that spatially filters activity coming from other locations and maximizes the strength of a current dipole at the location of interest, based on the known forward solution and the measured covariance between sensors. This process is iterated across a grid covering the whole brain to produce an activation map (17). The activation map can be simplified to a single location by locating the coordinates of the strongest response (21).

1.3 Reliability

As discussed previously, non-invasive pre-surgical mapping techniques must have high accuracy and low inter-session variability to be used instead of invasive methods. Accuracy in pre-surgical mapping is challenging to assess without using clinical standards, such as DCS. Precision can be assessed without invasive methods, but there is not enough data in a single scan to properly determine precision.

Another quality assurance measure to consider is inter-session variability. If a second functional imaging scan gives almost identical results as the first scan, this indicates that this method has low inter-session variability. A limitation to this approach in a clinical setting is that there is often not enough time to do a second scan. An ideal method would be able to tell from a single scan whether or not the data is high quality, and if a second scan would likely give similar results.

Reliability may be such a method. We can look at reliability within a dataset by taking data from a single scan, splitting it in half, and the amount of overlap between the two half datasets indicates the level of reliability. In contrast to intra-session variability, which focuses on the variance in the estimated location, reliability assesses the whole dataset. Also, reliability can be determined using a single scan, which increases its appeal for clinical use.

My group has developed a framework called ROCr (Receiver-Operator Characteristic Reliability) that automatically assess reliability of functional neuroimaging data by quantitatively comparing the overlap of two half datasets. In order to understand how ROCr works and quantifies the overlap between two datasets, a review of classification parameters and the receiver-operator characteristic is required.

1.3.1 Assessing Reliability

1.3.1.1 Classification Parameters

A binary classifier predicts if a data point belongs in a category (positive) or not (negative) based on a cut-off value. For example, consider a group of students aged 10 to 19. It is simple to separate the teenagers from the young children using the cut-off value of 13 years old. If the child is 13 or older, they are classified as a teenager. However, in real data there is often no cut-off value that correctly separates the data points that belong in the category from those that do not. Often, the distributions overlap, and for a given data point the prediction does not always match the real classification. If we now try using age to separate the children who are in high school from the children who are not, the problem becomes more challenging. Depending on location, children start school at different ages, and the definition of which grades count as high school can vary as well. We could say that the cut-off is 14 years old, but this will incorrectly classify some students.

Consider two overlapping distributions of data as shown in Figure 4; one belonging to category one (distribution on the right) and one not belonging (distribution on the left). The vertical line in the figure indicates the cut-off x-value (or “threshold”) that separates data that is and isn’t in category one. The distribution of data belonging to category one are split into two populations by the classifier: true positives (TP) and false negatives (FN). A true positive occurs when the classifier correctly predicts the data point is a part of category one (e.g. a high school student classified as a high school student). A false negative, also known as type II error, occurs when the classifier incorrectly predicts the data point is not a part of category one (e.g. a high school student classified as not being a high school student). If the cut-off is too high, the chance of type II errors increases. The distribution of data not belonging to the category are also split into two populations by the classifier: true negatives (TN) and false positives (FP). A true negative occurs when the classifier correctly predicts the data point is not a part of category one (e.g. a student who is not in high school is classified as not being in high school). A false positive, also known as type I error, occurs when the classifier incorrectly predicts the data point is a part of category one (e.g. a student who is not in high school is classified as being in high school). If the cut-off is too low, the chance of type I errors increases.

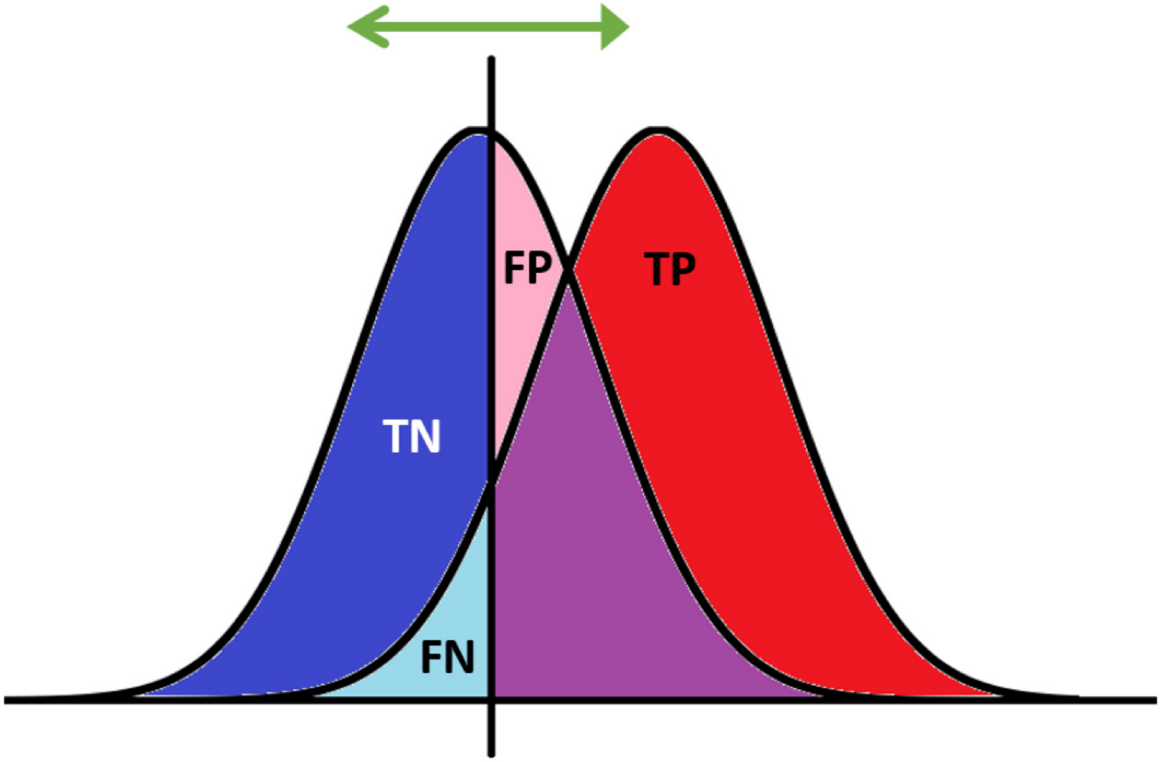


Figure 4: Overlapping Populations.

The distribution on the right is the population that belongs in the category. The distribution on the left is the population that does not belong in the category. The black vertical line is the cut-off value. Data points to the right of the line are classified as belonging to the category, and data points to the left are classified as not belonging to the category. This figure shows how some data points are incorrectly classified due to distribution overlap. The red and dark pink area indicates true positives (TP), and the pale blue area indicates false negatives (FN). These two populations add up to 1. The dark blue and pale blue area indicates true negatives (TN), and the light and dark pink area indicates false positives (FP). These two populations also add up to 1. Changing the threshold changes the populations.

1.3.1.2 ROC: Receiver-Operator Characteristics

The four populations discussed above (TP, FN, FP and TN) can be used to calculate several metrics describing the reliability of the data. Two such metrics are the true positive rate (TPR) and the false positive rate (FPR). The TPR is defined as the fraction of data belonging to the category (positive) that are correctly predicted by the classifier. TPR is formulated as follows:

$$TPR = \frac{TP}{TP + FN} \qquad \text{Equation (5)}$$

The FPR is defined as the fraction of data not belonging to the category (negative) that are incorrectly predicted by the classifier. FPR is formulated as follows:

$$FPR = \frac{FP}{FP + TN} \qquad \text{Equation (6)}$$

Two similar metrics are *sensitivity* and *specificity*. Sensitivity is defined as the proportion of positives that are correctly identified by the test:

$$Sensitivity = \frac{True\ Positives}{All\ Positives} = \frac{TP}{TP + FN} \qquad \text{Equation (7)}$$

Specificity is the proportion of negatives that are correctly identified by the test:

$$Specificity = \frac{True\ Negatives}{All\ Negatives} = \frac{TN}{FP + TN} \qquad \text{Equation (8)}$$

By comparing Equation (5) to Equation (7), we observe that the TPR is equivalent to sensitivity. By comparing Equation (6) to Equation (8), we observe that the FPR is equal to 1 - specificity:

$$\begin{aligned} &1 - specificity = FPR \\ &1 - \frac{TN}{FP + TN} = \frac{FP}{FP + TN} \\ &FP + TN - TN = FP \\ &FP = FP \end{aligned} \qquad \text{Equation (9)}$$

An ideal classifier would have a very high TPR and a very low FPR. Going back to the student example, a high TPR (or high specificity) indicates that the classifier is correctly identifying the majority of high school students as high school students. A low FPR (or high specificity) indicates that the classifier very rarely classifies a non-high school student as a high school student. It is possible for a classifier to have any combination of TPR and FPR.

The receiver operator characteristic (ROC) curve displays the relationship between TPR and FPR as the threshold is changed. The TPR and FPR are dependent on the threshold value used by the classifier. A lower threshold will increase the TPR, but this comes at the cost of an increased FPR. This results in an overestimation of the population belonging to the category. Sensitivity is increased, but specificity is lowered.

The ROC curve is obtained by calculating the TPR and FPR over a range of threshold values, then plotting the TPR values against the FPR values. The bottom left corner of the plot is at a very high threshold – so high that no data points are classified as belonging to the category. Returning to the example of children aged 10 to 19, this would be equivalent to setting the cut-off age for high school students at 20 years old. No young students would be incorrectly classified as high school students, but no high school students would be correctly classified either. The upper right corner of the plot is at a very low threshold – so low that all data points are classified as belonging to the category. This would be equivalent to setting the cut-off age for high school students at 10 years old. All the students would be classified as high school students, giving a high TPR and a high FPR. The range of thresholds in between give an indication of the classifier's reliability.

Consider Figure 5 below, which shows three different ROC curves. The ROC curve for threshold-based classification of two random distributions of data would be a straight line from the bottom left corner to the top right corner. Because the classifier works at random, the TPR and FPR increase at the same rate, resulting in a flat line. For curve (a), changing the threshold to increase the TPR increases the FPR by almost the same amount. Curve (b) represents a better classification than curve (a). The threshold can be changed to increase the TPR without an equal increase in the FPR. Curve (c) is close to perfect classification, which can achieve a TPR of one and a FPR of zero. There is a zone where changing the threshold increases the TPR with only a minimal increase in FPR. These different behaviours can be represented by calculating the area under the ROC curve (AUC).

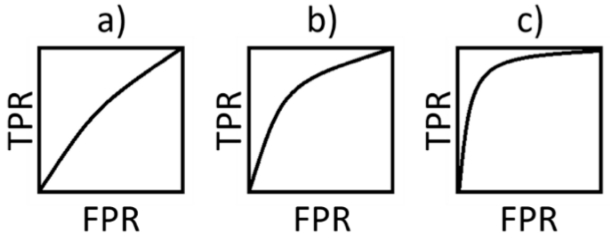


Figure 5: ROC curves.
(a) ROC curve with a low area under the curve. Changing the threshold to increase the TPR increases the FPR by almost the same amount. (b) ROC curve with a medium area under the curve, representing better classification than curve (a). (c) ROC curve with high area under the curve. This ROC curve is close to that of a perfect classifier. There is a zone where changing the threshold increases the TPR with only a minimal increase in FPR.

The area under the ROC curve (AUC) summarizes the reliability of the classifier. If there are no true positives at any false rate, the AUC will be 0. If there are no false negatives at any false positive rate, the AUC will be 1 (22). The AUC is generally greater than 0.5, because a classifier applied to random noise could achieve this classification accuracy. The AUC can be used to compare the reliability of two or more different classifiers. There is, however, a trade-off with increasing AUC. A classifier with a high AUC has a range of thresholds where a small increase in the TPR has a large increase in the FPR.

1.3.2 ROCr: Receiver-Operator Characteristic Reliability

1.3.2.1 ROC Applied to Functional Neuroimaging Data

The activation maps produced by functional imaging methods show the interpolated level of activity at each voxel or location. If the activity at a given voxel is above the chosen threshold, that voxel is considered to be active. If the activity is below the threshold, that voxel is considered to be inactive. However, due to noise from various sources, there is usually no threshold that perfectly separates the active voxels from inactive voxels. Additionally, there is no non-invasive way of knowing which voxels are actually active, so the TPR and FPR cannot be directly calculated. Despite these challenges, ROC analysis and similar overlap comparison methods have gained wide acceptance in medical imaging as a way to evaluate the reliability of the data (22–26).

There are different ways to apply ROC analysis to functional neuroimaging data, but my research uses a technique called ROC reliability (ROCr) reported by Stevens et al. (22), based on work by Le and Hu (27). Stevens et al.'s technique uses two datasets known as the test and retest sets. The second dataset is acquired in the same way as the first (same task and recording parameters). By this logic, either dataset can be called the test set. The test set is thresholded at a conservative level, and any voxels classified as active are considered to be the real activation. The retest set is overlaid on the test set as shown in Figure 6, and the overlap between these two sets is used to calculate the test-retest positive rate and the test-retest negative rate. In this test-retest application, the TPR is defined as the ratio of the detected true positives to the "real" activation (as determined from the test dataset):

$$TPR = \frac{Overlap\ Area}{Test\ Set\ Area} = \frac{TP}{TP + FN} \qquad \text{Equation (10)}$$

The FPR is defined as the ratio of false positives to the "real" negatives (as determined from the test dataset):

$$FPR = \frac{Non - overlap\ Retest\ Set\ Area}{Area\ not\ covered\ by\ Test\ Set\ Area} = \frac{FP}{FP + TN} \qquad \text{Equation (11)}$$

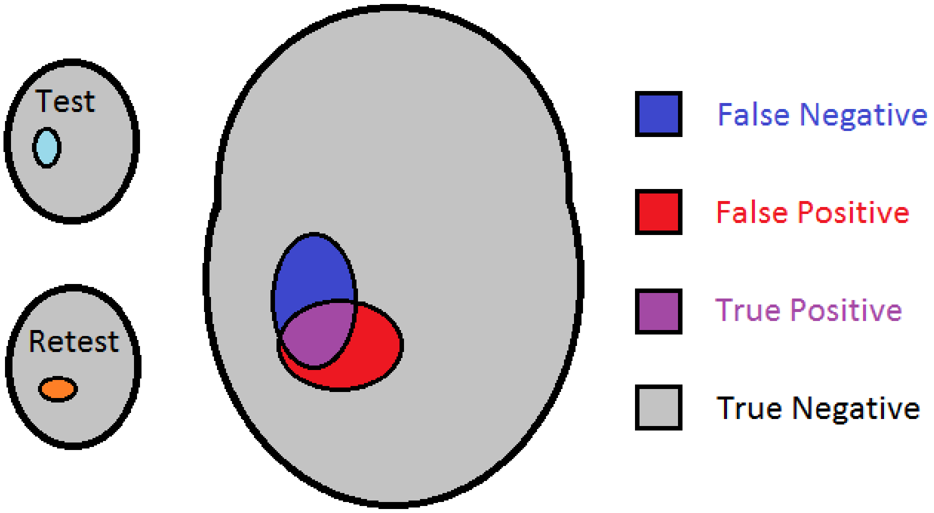


Figure 6: Overlap of test and retest datasets.

The area where the test and retest datasets overlap (pink) is considered to be true positives. Where neither set overlaps (grey) is considered a true negative. The blue area is false negatives. They belong to the test set and are considered as active, but the retest dataset did not cover this area. The red area is false positives. The retest dataset covered this area, but the test dataset did not.

To generate the full range of TPRs and FPRs, the threshold used with the retest set is varied as shown in Figure 7. Although these rates are not real TPR and FPR, they can be used in the same way to create an ROC curve that is indicative of the test-retest reliability of the two datasets. An ROC curve with a large AUC indicates that the test and retest sets overlap consistently across thresholds, indicating high test-retest reliability. In contrast, a small AUC indicates that the test and retest sets do not overlap consistently across thresholds, indicating low test-retest reliability.

The dataset columns in Figure 7 demonstrate how test-retest overlap relates to the AUC. In the first column, the area of overlap starts off large at low thresholds, and decreases as the threshold increases. The corresponding ROC plot has a low AUC. In the third column, the area of overlap is small but consistent as the threshold changes. The corresponding ROC plot has a high AUC. The second column is a compromise between the two; both in area of overlap consistency and AUC.

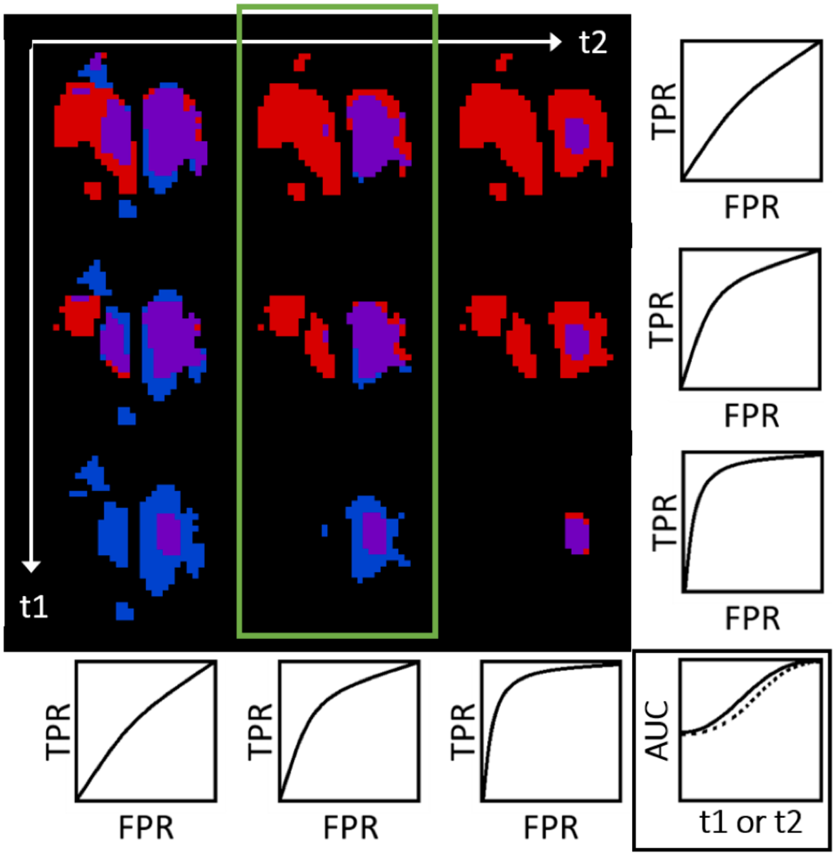


Figure 7: Creating ROC curves with neuroimaging data.

The combination of the blue and purple areas indicate the test set, and the combination of the red and purple indicate the retest set. The red areas indicates false positives, and the blue areas indicate false negatives. Purple areas indicate true positives, and black areas indicate true negatives. In the outlined column, the test set threshold (t_2) is held constant while the retest threshold (t_1) is increased (top to bottom). Each threshold change produces a new TPR and FPR, which are used to create the ROC curve shown below the column. This process is repeated for different test set thresholds to create the set of ROC curves at the bottom of the figure. The AUC is calculated for each test set threshold, and plotted in the AUC plot (bottom right).

1.3.2.3 AUC vs. Threshold Plot

The technique described above (hold the test set threshold constant and vary the retest set threshold) produces a single ROC curve with an associated area under the curve (AUC). If the procedure is repeated many times with new test set thresholds, a set of AUCs is produced as a function of test set threshold as shown in Figure 7. Plotting the AUCs against the test set thresholds gives the AUC curve shown in the bottom right-hand corner of Figure 7, and in greater detail in Figure 8.

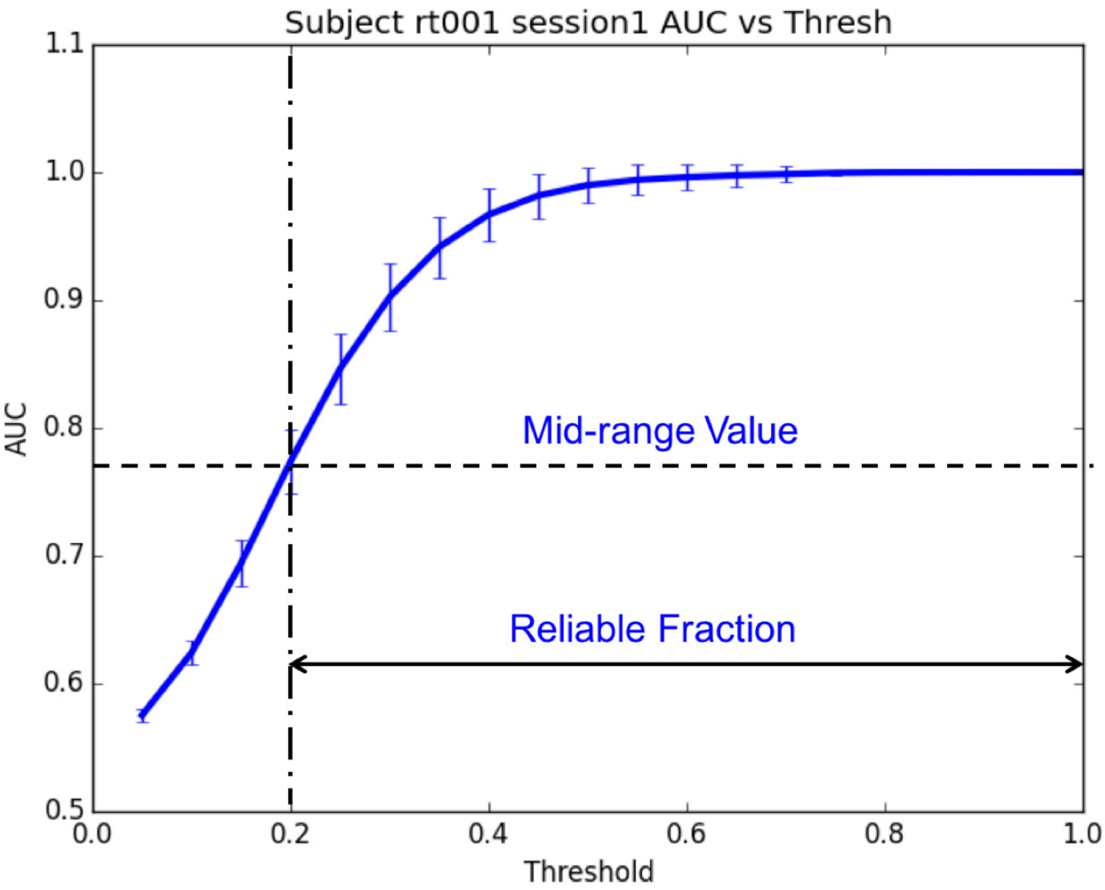


Figure 8: AUC vs. Threshold plot.

The x axis is the test set threshold, and the y axis is the AUC. As the curve approaches the top right corner (test set threshold increases), there is an increase in false negatives. As the curve approaches the bottom left corner (test set threshold decreases), there is an increase in false positives. The mid-range value is the median AUC value. The reliable fraction (F_R) is the fraction of threshold values for which the AUC is greater than the mid-range value. The faster the AUC curve crosses the mid-range value, the larger the reliable fraction and the more reliable the data is.

Metrics and data reliability information can be obtained from the AUC curve. As the test set threshold is increased, there is an increase in false positives. A higher test set threshold reduces the area covered by the test set compared to the retest set, so there is less area available for the retest set to overlap. As the test set threshold decreases, there is an increase in false negatives. A lower test set threshold increases the area covered by the test set compared to the retest set, so there is more area not covered by the retest set.

According to the ROCr approach, the AUC curve can be used to find a threshold that strikes a balance between type I and II errors. The mid-range value can be calculated from this curve by averaging together the maximum and minimum AUC values, and the linear rate can be found by determining the slope of a line that would connect the initial and final AUC values (21). Given the shape of the AUC curve shown in Figure 8, a good threshold may be the value at which the actual rate of change of AUC with respect to test set threshold matches the linear rate. At values higher than this test set threshold, there are diminishing returns on reliability and an increased loss in sensitivity.

Finally, the reliable fraction (F_R) is a value between 0 and 1, showing the percentage of threshold values for which the AUC curve is above the mid-range value. A high reliable fraction indicates that reliability increases quickly with image threshold, and remains high for a large range of thresholds (21). The faster the curve passes the mid-range value, the larger the F_R and the more image threshold values for which the data is considered reliable.

1.3.2.4 Data Splitting Approaches with ROCr

The test and retest sets can be obtained in two different scans, as done by Stevens et al. (21,22) . In this case, the reliable fraction provides a measure of inter-session reliability. This approach increases the amount of time required to obtain the required data, and any different conditions between the two scans will affect the results. An alternative is to take the original dataset and split it randomly into two sets, then treat them as the test and retest sets, as done in a later study by Stevens et al.

(28). In this case, the reliable fraction provides a measure of intra-session reliability. Using only one dataset eliminates the risk of conditions changing between scans. In MEG, data can be easily split once it has been split into trials; half of the trials make up the test dataset and the other half make up the retest dataset. Trial selection can be randomized for each split, giving many estimates of reliability from a single dataset. Once this distribution of reliability estimates has been obtained, statistics can be used to determine metrics such as the average reliability and standard deviation.

1.3.2.5 ROCr Framework in Practice

This procedure of taking the data and automatically obtaining the F_R (as a measure of data quality) and a threshold is termed the ROCr framework. This framework was originally used with fMRI data. A 2013 paper by Stevens et al. presented a novel method of using ROCr to select the optimal threshold for fMRI activation maps. Using this method, they were able to achieve localization results that better matched DCS results (22).

There are still some challenges to be overcome to promote the use of MEG in clinical settings. It is challenging to perform source mapping at the individual level that is reproducible and robust. Another challenge is that the ECD method doesn't work well for distributed cortical activity, and requires operator input for multi-source activity. This introduces an opportunity for operator bias, gives poor inter-rater reliability and limits automated approaches. Volumetric source models can accurately model distributed or multiple sources, overcoming a key limitation of the ECD method, but these models do not have well-established methods of quality assurance, such as GoF used with ECD.

To overcome these challenges, Stevens et al. presented a practical use of the ROCr framework to assess the quality of MEG functional maps in a recent 2016 paper (28). Specifically, they proposed the use of the ROCr reliable fraction for quality assurance and data-driven thresholding of volumetric source maps. ROCr offers two key benefits. The first benefit is that it increases confidence in the localization results by providing a quantitative measure of source map reliability. The second key benefit is that ROCr uses a data-driven approach to identify optimal threshold levels. This allows a push-button approach to thresholding, decreasing variability between raters.

Stevens et al. set out to demonstrate that ROCr analysis could be used as a quality assurance and automated thresholding tool. The first step in this process was to validate their solution using several sets of simulated MEG data. The simulated datasets were created by calculating the forward solution from a current dipole with a known location, orientation and time course, and superimposing this data on resting state MEG data. The time course of the dipole was varied between datasets to produce ten simulated datasets. Once a simulated dataset was created to mimic a real dataset, ECD modeling and beamformer mapping were done on the data.

Each simulated dataset was analyzed in the following way, starting by epoching the dataset into trials synchronised to the onset of each simulated trial. For ECD modeling the epochs were averaged, and for each time point in the evoked response, an ECD was modelled that maximised the GoF. Next, the GoF time course was used to find the latency of the GoF peak nearest to P35m component of the SEF. This procedure was repeated using CV instead of GoF, as a measure of data quality. For beamformer mapping, the dataset was epoched and divided into split-halves, then the beamformer source estimation was calculated for each split-half. This step was done eight times per dataset, and the resulting 16 SEF maps were used to create an average SEF map.

Quality analysis on the beamformer activation maps was done using the ROCr framework. Each of the eight split-half pairs was used as ROCr inputs to calculate eight time series of reliable fractions (F_R). The mean and variance of the eight F_R time series were calculated, and compared to the ECD GoF and CV time series. The average F_R time series was used to find the latency of F_R closest to P35m, as was done with the ECD GOF and CV time series. The latencies of these three quality assurance methods were compared to validate ROCr.

In addition to calculating F_R, the ROCr framework was also used to find an “optimal” threshold for each split-half pair, balancing high reliability and high sensitivity. The ROCr optimal thresholds were averaged together, and applied to the averaged beamformer map. The thresholded beamformer map was then used to extract the peak location and latency. Because the location and timing of the underlying dipole in the simulated data was known, the peak location and latency calculated using ROCr, GoF and CV could be compared to the actual values.

The results of the simulated data analysis showed that maximizing F_R was just as effective at providing accurate localization of the simulated source as maximizing GoF or minimizing CV. Validating ROCr on simulated data was an important step, because the localization results could be compared against a known exact location and time, which is not the case with real datasets.

The next stage in validating ROCr was to repeat the steps discussed above, but with real data. Data collection and pre-processing was done in a manner similar to the MNS paradigm discussed above.   Eighteen volunteers participated in the study, and the same processed data sets were used for ECD modeling and beamformer mapping using the procedures discussed above. ROCr thresholding was done to the beamformer activity maps to obtain localization results. ECD quality assurance was done using GoF and CV, and beamformer quality assurance was done using F_R. With the real datasets, the true locations of S1 were not known, so the beamformer localization results were compared with the ECD localization results.

The results were once again positive. F_R was shown to have a correlation to other conventional quality assurance metrics such as GoF and CV. The results also indicated that better localization results are more likely to be obtained from higher quality data, regardless of the method used.

Stevens et al. met their goal of demonstrating that ROCr analysis could be used as a quality assurance and automated thresholding tool. This has great significance for the future of clinical MEG for pre-surgical mapping. ROCr solves the issue of providing a quality assurance method for volumetric localization methods, which are superior to the ECD method for mapping distributed source activity. ROCr also allows volumetric images to be produced and analysed without the need for expert intervention. The data-driven threshold selection that is possible with ROCr has potential to standardize volumetric image thresholding, further reducing operator bias. These advantages bring MEG closer to being a technology that can be used by non-experts to enhance pre-surgical mapping in patients.

1.3.2.6 Using Reliable Fraction to Guide Analysis Pipeline Choices

The reliable fraction has demonstrated potential as an indicator of data quality, in terms of reliability of localization for the SEF generated by MNS. Given that, the ROCr framework may also be useful to rank the reliability of different analysis pipelines for a given individual's dataset.

In theory, a clinical application of the reliable fraction would involve the following steps. First, MNS data would be collected using standard MEG MNS acquisition protocols. Next, the collected data would be split into the test and retest sets. Each set would be analysed using a set of pre-determined analysis pipelines, giving test and retest activation maps for each pipeline. The thresholds of the test and retest activation maps would then be varied to produce AUC vs. test threshold plots, from which reliable fractions could be calculated. The pipeline with the highest reliable fraction would be selected for the patient's data. Finally, the patient's original dataset would be analysed using the ROCr selected pipeline to determine the location of S1. A key clinical benefit of using ROCr to select an analysis pipeline is that it eliminates the influence of the experimenter on the final activation maps, reducing the need for expert intervention. This would also ensure that the analysis pipeline applied for the patient's data provides the most reliable activation map.

1.3.3 NPAIRS: An Alternative Method to Assess Reliability

ROCr is not the only method to assess reliability of functional neuroimaging data. Another prominent method is NPAIRS (Nonparametric Prediction, Activation, Influence, and Reproducibility resampling) (29). Like ROCr, the first steps in the NPAIRS method are splitting the data from a scan in half, and analyzing each half separately to get two activation maps.

From the two activation maps, NPAIRS estimates two metrics of the data: reproducibility and prediction. Reproducibility “quantifies the global similarity” of activation maps, and is quantified using the Pearson’s correlation coefficient. Reproducibility ranges from 0 (worst) to 1 (best). Prediction “measures how consistently the analysis results of split 1 can predict the class, or brain state, of individual scans in split 2 and vice-versa, via Bayes’ posterior probability” (30). In other words, split 1 is used as a training set for a classifier which is then used to classify scans in split 2. Prediction ranges from 0 to 1, with 0.5 corresponding to a random classifier.

An ideal dataset would have high reproducibility and high prediction (R, P = 1), but it is difficult to maximize one metric without reducing another. To overcome this difficulty, “model performance is defined as Euclidean distance … from perfect prediction and reproducibility. Better pipeline performance is given by smaller [Euclidean distance]” (30,31).

1.3.3.1 Similarity to ROCr

ROCr and NPAIRS share some common elements, but also have their differences. Both methods split the raw data in half, generating two activation maps, but the comparison of these two maps is a key difference. NPAIRS does not use activation map thresholding, which is a key part of the ROCr framework, and ROCr does not consider prediction. However, the trade-off between repeatability and predictability in NPAIRS is comparable to the trade-off between TPR and FPR in ROCr. Churchill et al. have investigated NPAIRS as a quality assurance measure for fMRI data (31), similar to work done by Stevens et al. (22), and as a tool to select pipeline parameters for analyzing an individual’s fMRI data (30). It is not known, however, if there is value in using ROCr for pipeline selection.

1.4 Objectives and Hypotheses

The objective of my thesis is to investigate ROCr as a tool for automated pipeline selection. To achieve this objective, I analyzed previously collected MEG data from a median nerve stimulation task repeated on three different days. My lab has already shown good inter-session reliability (low inter-session variability) in localization of S1 based on the SEF in these data sets, using a modified version of the clinical standard pipeline. There are a number of parameters that can be adjusted in the analysis pipeline that localizes S1 from MEG data. Given the robust nature of the SEF response, however, we expect that the most reliable approach to analysing the data for a given subject will not change from day to day.

For each participant's data at each session, I will modify five analysis parameters (environmental noise reduction, motion correction, high-pass filtering, low-pass filtering and ICA) to generate twelve analysis pipelines that can localize the SEF. The outcome measure for this analysis will be the reliable fraction as a function of analysis pipeline parameters, participant and session. I will also be able to determine the coordinates that define the localization of S1 for each of these data points.

I have two hypotheses regarding these data. **Hypothesis A:** Given the robust nature of the SEF to MNS, ROCr analysis will indicate the highest reliable fraction for the same pipeline for a subject's data on each of the three days. The pipeline with the highest reliability may vary across subjects due to inter-subject variances, but within subjects the pipeline should be the same. **Hypothesis B:** Across all subjects and sessions, the variability of the localization results from analysis pipelines with highest reliable fraction will be the same or better than the variability of localization from the clinical standard pipeline.

Given that functional mapping is often completed on a region of interest, rather than the whole brain, I will also investigate the impact of selecting a smaller subset of brain areas on the outcome measures described above. Most simply, this can be achieved by splitting the maps into two hemispheres, given the contralateral dominance of early SEF responses. **Hypothesis C:** Across all subjects and sessions, the reliable fraction should improve if only the hemisphere with the strongest predicted response is considered in the analysis.

CHAPTER 2 METHODS

2.1 Participants

The study was approved by the Research Ethics board at the IWK Health Centre, and participants provided informed written consent. The original participants were eighteen healthy right-handed volunteers (10 females; age 19-29, mean 24 years). Of the original volunteers, data from ten participants were used for this study. Two subjects could not be used because they did not complete all three scans. Another subject was eliminated from previous studies with these data due to poor data quality. The remaining dropped subjects were not used because the ICA algorithm could not converge for these subjects when their data was analyzed with Pipeline 1.

2.2 Data Acquisition

Each participant completed three sessions, each on consecutive days at approximately the same time. In each session, somatosensory evoked fields were generated to allow localization of the primary somatosensory cortex. To do this, the left and right median nerves were stimulated percutaneously (DS7A Constant Current Simulator, Digitimer, England). The stimulation pulse is administered from the current simulator to the subject via a small module with two hydrated electrodes positioned along the median nerve at each wrist. The output voltage and current of each stimulator, and the stimulator position, were adjusted such that the participant reported a sensory response in the innervated thenar muscle, and a slight visible twitch of the thumb was observed. Once the ideal output was established, 80 to 100 stimuli were applied to each wrist with an inter-stimulus interval between 1 and 2 seconds, while the participant watched a short movie with no sound. The movie was included to control for participant arousal. The sequence of stimuli was quasi-random, with no more than four consecutive stimuli on a side. To allow event-related analysis, the timing and order of the stimuli were recorded continuously along with the MEG data. Stimulus timing and order was controlled by the Presentation software (Neurobehavioral Systems Inc., Berkeley, CA, USA).

The MEG data was recorded on a 306 channel MEG system (Elekta Neuromag Oy, FL). Electrooculography (EOG) was used to track horizontal and vertical movement using four electrodes: one electrode at the outer corner of each eye (horizontal) and electrodes inferior and superior to the left eye (vertical). Head position was tracked continuously throughout the data collection, via four indicator coils. One coil was placed on each mastoid, and two coils were placed on the forehead. All data were acquired continuously at a sampling rate of 1500 Hz and a bandwidth of 0.1-500 Hz, and recorded to a file for off-line analysis (11). Head digitization was performed using the Isotrak system (Polhemus Inc., Colchester, USA) to provide a model of the head shape for source localization. The nasion, left/right preauricular points, and HPI coils were digitized, and a 150-200 point digitization of the scalp surface was done (11,28).

2.3 MEG Data Analysis

2.3.1 Analysis Overview

My work focuses primarily on intra-session reliability as indicated by the reliable fraction. For the remainder of this thesis, "reliability" refers to intra-session reliability unless indicated otherwise. Figure 9 below provides an overview of the analysis done in this project, and the desired outcome measures. Outcomes related to Hypothesis A are shown in green, and outcomes related to Hypothesis B are shown in orange. Each of the 10 subjects had three datasets. The first dataset was processed with each of the twelve pipelines shown below in Table 1 (including the standard pipeline), producing activation maps. The activation maps were analyzed with the ROCr framework to obtain a measure of reliability for each pipeline, and the pipelines were ranked based on reliability. The activation maps were also used to localize S1 for each pipeline by locating the highest-intensity voxel. For each pipeline we get a reliable fraction and an estimated S1 location. For each subject, the second and third datasets were processed in the same manner as the first dataset.

For each subject, the pipeline rankings were compared across the three datasets to test Hypothesis A. If Hypothesis A is true for a given subject, the top-ranked pipeline should be the same on each of the subject's three scans, and the rankings should be similar across scans.   To test Hypothesis B, the Euclidean distance of S1 location from the highest-ranking pipeline was compared to the Euclidean distance of S1 location from the standard pipeline (across the three datasets). If Hypothesis B is true, the S1 locations from the highest-ranking pipelines on day 1, 2 and 3 should be more focal (as determined by smaller Euclidean distances between each day's result and the average result) than the S1 locations from the standard pipeline.

Although the hypotheses only apply to subject-level analysis, some group analysis was done to gain an understanding of how ROCr works with this data. Pipeline rankings and S1 locations were compared across subjects as well.

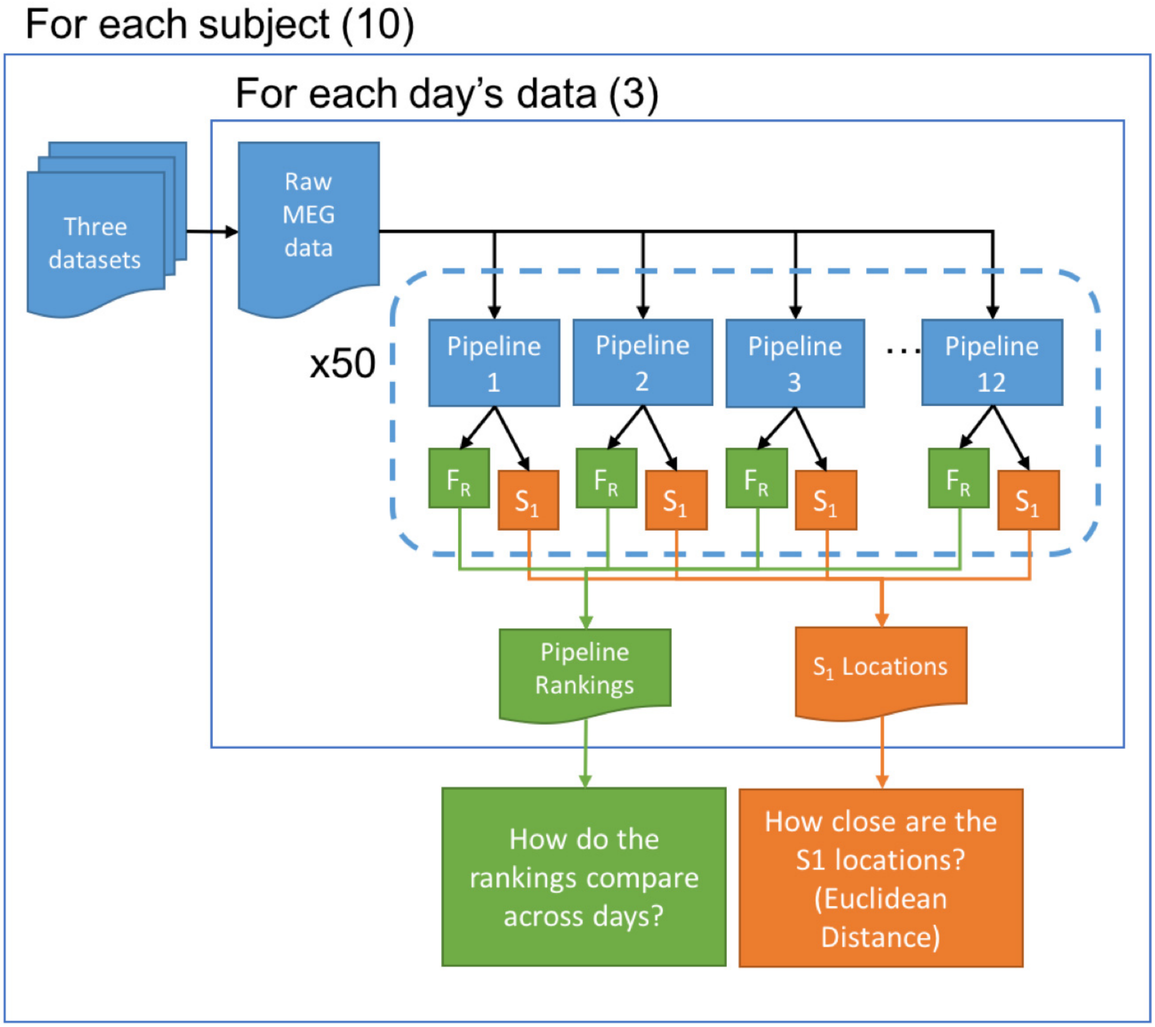


Figure 9: The Big Picture.
Each subject has three datasets. For each dataset, the data is processed with each pipeline and the ROCr framework, giving a reliable fraction and S1 location for each pipeline. This step is repeated 50 times to give a distribution of reliable fractions and S1 locations. At the subject level, the reliability measures and S1 locations are compared to address the hypotheses.

2.3.3 Software Packages

MEG data analysis was done using MaxFilter (MaxFilter, 2.2.15 [2017], Elekta AB, Stockholm, Sweden) and MNE-Python (MNE-Python, Version 0.13.1, [2017]) (20). Specifically, motion correction and environmental noise reduction were done using MaxFilter, and the remaining processing steps were done using MNE-Python. General analysis scripts were written in Python (Python 2.7.11, Anaconda 2.3.0) Additional packages used in Python scripts included SciPy (Version 0.19.0), numpy (Version 1.12.0) and nibabel (Version 2.0.2). Plots were produced using matplotlib (Version 1.4.3). AFNI (Version AFNI_2011_05_26_1457) was used for browsing and clustering of activation maps. Statistical analyses were done using IBM SPSS and the SciPy (Version 0.19.0) package.

2.3.4 Pipeline Details

The pipeline covers all the steps of data analysis following data acquisition until the source localization is complete. Each step of the pipeline has parameters that can be modified as discussed in detail in the Introduction. The parameters that were modified to generate the 12 pipelines used in my study are discussed in more detail in Table 1, and the final pipelines used for analysis are shown in Table 2.

One of the pipelines (Pipeline 6) was termed the standard pipeline, with the goal of mimicking the clinical standard pipelines proposed by Sharma et al. (32) as closely as possible. The clinical pipeline could not be directly used with my data, as they were not recorded at a high enough sampling rate. The clinical SEF paradigm records data at 2500 Hz to allow detection of the N20m peak (32), but the data used in my study were recorded at 1000 Hz and downsampled to 250 Hz. To overcome this issue, I instead tried to match the standard pipeline proposed by Solomon et al. which was previously used to analyze these data and localized the P35m peak (11). All parameters were matched except for the LPF cut-off frequency. The highest LPF cut-off frequency used in my pipelines was 55 Hz, which is lower than the 70 Hz cut-off frequency used by Solomon et al.

Step	Parameters		
Motion Correction			
Method	Automated	None	
Environmental Noise Reduction			
Method	tSSS	SSS	
Filter			
HPF	1 Hz	No additional filtering	
LPF	Conservative (55 Hz)	Moderate (35 Hz)	Aggressive (15 Hz)
ICA			
Use ICA	Yes	No	
# of Components	N = Sufficient components to explain 99% of variance		
Method	Fast ICA		
Noise Covariance			
t_{min}	-100 ms		
t_{max}	0 ms		
Data Covariance			
t_{min}	0 ms		
t_{max}	500 ms		
Source Localization			
Method	Beamformer		

Table 1: Parameters used in analysis pipelines.
The noise and data covariances, and source localization methods, were consistent across pipelines. The other pipelines (motion correction, environmental noise reduction, high pass filtering, low pass filtering and ICA) were varied across pipelines. Although ICA has parameters that can be modified, all ICA pipelines used the FastICA method and N = 99% of variance.

Pipeline	tSSS	MC	HPF	LPF	ICA
1 (SSS)	No	No	No	55 Hz	Yes
2 (MC)	Yes	Yes	No	55 Hz	Yes
3	Yes	No	No	55 Hz	No
4	Yes	No	No	15 Hz	Yes
5	Yes	No	No	35 Hz	Yes
6 (Standard)	**Yes**	**No**	**No**	**55 Hz**	**Yes**
7	Yes	No	1 Hz	15 Hz	Yes
8	Yes	No	1 Hz	35 Hz	Yes
9	Yes	No	1 Hz	55 Hz	Yes
10	Yes	No	1 Hz	15 Hz	No
11	Yes	No	1 Hz	35 Hz	No
12	Yes	No	1 Hz	55 Hz	No

Table 2: Analysis pipelines (alternative names shown in brackets).
Pipelines 1, 2 and 3 were designed to match the standard pipeline except for one parameter, to allow testing on the effect of said parameter. Pipelines 4, 5, and 6 all use ICA and do not use a HPF, but vary in terms of LPF cut-off frequency. Pipelines 7, 8 and 9 use ICA and a 1 Hz HPF, but vary in terms of LPF cut-off frequency. Pipelines 10, 11 and 12 do not use ICA and use a 1 Hz HPF, but vary in terms of LPF cut-off frequency. This set-up allows for paired t-tests to determine the effects of certain parameters. For example, to determine the effects of using ICA, Pipelines 7, 8 and 9 can be compared to Pipelines 10, 11 and 12, because the only unbalanced difference in these pipelines is the use of ICA.

Once the analysis pipelines were finalized, they were used on the data to generate the test and retest datasets required for ROCr analysis. For a given subject on a given day's raw data, each pipeline was used separately to analyze the data. While each pipeline varied in terms of parameters used, the overall procedure was the same: initially, the data was pre-processed, epoched, randomly split into two subsets, and the remaining processing steps were completed on each subset to generate two activation maps. The following paragraphs discuss the pipeline parameters that were used in this experiment.

2.3.4.1 Motion Correction

One of the twelve pipelines used a motion correction algorithm. Head movement during an MEG scan can reduce data quality, in some cases making the data unusable. Motion correction algorithms can improve data quality if the head movement is within acceptable limits. My dataset includes participants who stayed within the head movement limits for all scans and some who exceeded the limits for head movement in one of their sessions. All datasets were included in the analysis, allowing for the comparison of pipelines with and without motion correction for subjects with and without large head movement.

2.3.4.2 Environmental Noise Reduction

The standard method of environmental noise reduction is temporal signal space separation (tSSS). Preliminary data analysis showed that using tSSS gave consistently higher reliability and localization accuracy than not using environmental noise reduction. However, not using any environmental noise reduction resulted in data that was incompatible with many of the analysis pipelines. Signal space separation (SSS) is a method of environmental noise reduction that was developed before tSSS. It was expected that the SSS pipeline would perform worse than the pipelines using tSSS, but better than not using any environmental noise correction. As a comparison, one pipeline used SSS instead of tSSS. At this stage, data were also downsampled to 250 Hz with a low-pass filter of 70 Hz to minimize disk usage and increase processing speed.

2.3.4.3 Filtering

All filters used in this study were finite impulse response (FIR) filters using Hamming windowing. The filters were zero-phase with a transition band of 0.5 Hz. For the analysis pipelines in this study, either a) additional high-pass filtering was applied with a cut-off frequency of 1 Hz or b) no additional high-pass filtering was applied (DC HPF). Low-pass filtering was used in all pipelines, but the cut-off frequency varied. To determine what these frequencies should be, a "grand average" somatosensory evoked field was created by averaging the SEF data for each sensor across all participants. A time-frequency representation of the grand average data was created using a Morlet wavelet analysis, and plotted as a spectrogram between 0 and 75 Hz to explore a combined temporal-spectral representation of the data. From this plot, three frequencies were selected for the low-pass filter cut-off that trade-off between suppressing noise and potentially attenuating signal of interest: a conservative frequency (55 Hz), an aggressive frequency (15 Hz), and a midrange frequency (35 Hz).

2.3.4.4 Independent Component Analysis (ICA)

For the pipelines used in this study, ICA was either used or not used. If ICA was used, the MNE-Python implementation of the FastICA algorithm (20) was used to deconstruct the MEG data into independent components (ICs). ICs were rejected based on three criteria. First, any ICs exceeding the rejection threshold (4 pT for magnetometers, 400 pT/m for gradiometers) were rejected. Next, the ICs were compared to the EOG signal, and up to three sources were rejected based on their correlation. In the same manner, ICs were rejected based on their correlation with the ECG signal.

2.3.4.5 Data Epoching

For all pipelines used in this study, data was epoched from 100 ms pre-stimulus to 500 ms post-stimulus, to separate the data into data segments ("trials") synchronized such that time zero is the onset of the stimulus in each trial. A baseline interval of 100 ms gives sufficient data to calculate the baseline activity level and to calculate the noise covariance. Because the inter-stimulus interval was 1-2 s, any post-stimulus time under 1s would be sufficient. For this study, a 500 ms interval was chosen because previous literature on the SEF shows the response ends before 200 ms (14,28,33).

2.3.4.6 Data Splitting: Generating Test and Retest Datasets

Once epoched, the data was randomly split into two sets (test epochs and retest epochs) by randomly shuffling the epoch indices and splitting the list of indices into two subsets. The data was shuffled and split 50 times to produce 50 pairs of epoch subsets, in order to generate a distribution of reliability measures per subject. Initially these steps were repeated 100 times per pipeline, but it was observed that the reliability measure converged after 50 splits, so the number was reduced to speed up analysis.

The split was performed using the Python “random” module (34). The epochs were shuffled using the random.shuffle() function, and split in half. The first half of the shuffled epochs became the test epochs, and the second half became the retest epochs. To determine if the random module could be used to randomly split the data without repetition, the total number of list permutations was calculated. All datasets have under 100 epochs. This gives a maximum of approximately 2^{22} possible combinations of list orders, much less than the period of the “random” module, which is $2^{19937-1}$. Therefore, the module was considered suitable for use with this data.

2.3.4.7 Data Averaging

For each subject’s dataset, there were 68 to 99 MNS pulses. Thus, each split-half had 34 to 50 epochs. Each channel’s time course was averaged together across the epochs, resulting in a 2-D matrix of channels by time, producing single waveforms per channel for each condition. The inter-trial average attenuates signals that are not temporally correlated to the event marker used for epoching (i.e., the median nerve stimulation). The test and retest epochs were averaged separately, giving two sets of evoked data for a given dataset.

2.3.6 Generating Activation Maps

Finally, source estimation was done to generate brain activation maps. For this project, the LCMV Beamformer method was used. For this study, the noise window was the same as the baseline: -100 ms to 0 ms. The data window was 0 ms to 500 ms. Because no MRI data were collected for the participants, the head was modelled as a single sphere based on the 150-200 point digitized head shape. The epoched data was assessed to calculate the noise and data covariance, then the LCMV Beamformer method was used to produce the brain activation maps. For this study, the activation maps produced were four-dimensional (3D + time) data, with 5mm voxels and 150 time samples from –100 ms to 500 ms. This analysis resulted in two 3D+time activation maps (one for test and one for retest) for each of the 50 random split-halves, and for each analysis pipeline.

2.3.7 Optimizing MEG Data Analysis for Computation Time

This analysis procedure was very time intensive as each pipeline was used 50 times per dataset. To reduce the time required for this analysis, I optimized the processing for computational speed by doing certain pipeline steps before the split, if possible for that step. The steps that only had to be done once were MC, filtering, ICA, and calculation of the forward solution. This resulted in a pre-processed dataset that required only a few steps to be done after the split: calculating the noise and data covariances, and the LCMV Beamformer method to generate the activation maps.

2.3.8 MEG Analysis Outcome Measures

By the end of the MEG analysis, several outcome measures were obtained for each subject, per day, per pipeline. The first was an evoked MEG dataset, containing 306 channels with 150 time points per channel. The second outcome was a whole-dataset activation map, with 5mm voxel size and 150 time points. The third outcome was 50 pairs of Test/Retest activation maps, with 5mm voxel size and 150 time points. MEG data was also averaged at the group level, producing a grand average.

2.4 ROCr Data Analysis

Once the 50 pairs of test and retest brain activation maps were produced, the ROCr method was used to assess test-retest reliability for each pipeline on each day and for each participant. As mentioned previously ROCr analysis compares two datasets. In normal ROC analysis, one dataset is known to be true and the other dataset is compared to it. When used with functional neuroimaging data, however, it is not known what the true activation is. Because of this, ROCr analysis is done twice, once with the first dataset considered to be the true activation, and once with the second dataset considered to be the true activation. The results are combined to give a single set of results representative of the reliability of the data as a whole. These procedures are discussed in more detail in the following paragraphs.

2.4.1 ROCr Logic

Each pair of activation maps was analyzed using the following steps to produce a measure of the data's reliability. First, a list of twenty thresholds was generated. The thresholds are evenly spaced from 0 to the maximum activation level of the dataset. The test set activation map was thresholded at each threshold in the list. The retest set activation map was also thresholded at each threshold in the list. Each of the 20 thresholded test set maps were overlapped with each of the 20 thresholded retest set maps. The 400 overlaps were characterized by calculating the true positive rate and false positive rate for each overlap (Equation (10) and Equation (11)). For each of the 20 test set thresholds, the corresponding TPR and FPR values were plotted to generate an ROC curve (20 curves in total, each with 20 TPR/FPR points). The area under each of the 20 ROC curves was calculated, and plotted against the test set threshold used to generate the ROC curve. The resulting plot is the AUC vs. Threshold Plot. From this plot, the reliable fraction was determined by calculating the percentage of test set threshold values for which the AUC curve is above the mid-range value.

These steps were repeated, with dataset 1 and 2 switched. At this point, two reliable fraction measures have been made, and these two measures are averaged together to give a single measure representative of the reliability of the data. These data were output in text and graph form for each dataset/pipeline pair.

For this project, reliability was assessed in two ways: on the data as a whole ("Overall Reliability") and for each time point ("Reliability vs. Time").

2.4.1.1 Overall Reliability

Overall reliability is the reliability calculated on the activation map over all time (-100 to 500 ms with respect to the stimulus). Each activation map is represented as a 4D array: 3D voxel arrays of activity at each time point. For this reliability measure, the array is flattened into a 1-D x*y*z*time vector, and the procedure outlined above was done on the entire dataset at once.

2.4.1.2 Reliability vs. Time

As shown by Stevens et al., the Reliability vs. Time data can be used as a quality assurance measure and can recommend an ideal time to perform source localization (28). Reliability vs. Time data was generated by calculating reliability of the data at each time point. The 3D dataset at each time point was flattened into an x*y*z vector before being analyzed. In this case, instead of a single reliable fraction at the end, this method produced a reliable fraction for each time point. From this list of reliable fractions, the peak reliable fraction was selected to represent the reliability of the dataset for pipeline ranking and performance purposes. The timing of the peak reliable fraction was also recorded to observe the effect of different pipelines on peak reliability latency.

2.4.2 ROCr Analysis Outcome Measures

By the end of the ROCr analysis, several outcome measures were obtained for each subject, per day, per pipeline. The first was 50 pairs of AUC vs. Threshold plots for the Overall Reliability method. The second was a distribution of 50 F_R values. The third outcome measure was 50 Reliability vs. Time lists, each with 150 F_R values.

2.5 Localization of S1

S1 was localized by finding the peak voxel at the time of interest in the activation maps produced by each pipeline. In clinical situations, the early 20 ms peak is localized, but the data for this study was not recorded at a high enough sampling rate to properly measure this peak. For this study, the localization time was determined from an activation map made using the whole dataset, without splitting the epochs in half. The voxel location with the highest activation level is found, and the time at which the activation occurred is recorded. This time is then used as the localization time for each split half. Each dataset/pipeline combo may have a different time, but each set of 50 split halves will use the same time.

Localization was done using the AFNI 3dclust function. This function takes a dataset, thresholds it, and returns information on any clusters of voxels above the threshold. For this study, each dataset was thresholded at 70% of its maximum value. For each cluster, the outputs included: centre of mass coordinates, bounding box coordinates, mean intensity value, maximum intensity value, and maximum intensity coordinates. The output values for each dataset were written to a file, and the maximum intensity coordinates were used as the S1 location in Study 2 to calculate Euclidean distances.

2.6 Study 1: Pipeline Performance and Rankings

The goal of this study was to assess the performance of ROCr as a tool for automated pipeline selection. This study addresses Hypothesis A and Hypothesis C. Briefly, Hypothesis A states that ROCr should pick the same pipeline for each of a subject’s three datasets (since there should be very little variance between the three days’ datasets), and Hypothesis C states that restricting the dataset temporally or spatially will improve ROCr reliability scores.

2.6.1 ROCr Pipeline Selection

Once the ROCr analysis was complete, the outcome was 50 measures of reliable fraction for each pipeline and for each dataset. Analyses were done on these data to examine the most reliable pipeline and reliability rankings of all pipelines. These analyses were done twice per subject, once using the Overall Reliability measure and once using Reliability vs. Time measure.

2.6.1.1 Most Reliable Pipeline

Pipeline performance was analysed on a per-subject level. For each session, the mean reliable fraction of each pipeline across splits was calculated. The outcome measure for this section was the pipeline with the highest mean reliable fraction, which was designated as the ROCr-selected pipeline. The ROCr-selected pipeline was compared across days to see if the same pipeline was selected on each of the three sessions, as per Hypothesis A.

2.6.1.2 Pipeline Rankings

On a given day, a subject's pipelines can be ranked based on average F_R. These rankings give additional insight into how well ROCr is performing. Instead of simplifying a day's data down to a single pipeline, rankings are less dependent on the number and nature of pipelines used for evaluation. If ROCr is performing as hypothesized, the pipeline rankings should be similar across sessions, not just the top pipeline. Some variance is acceptable, but the rankings should be similar unless there is something abnormal in the data, such as excessive head movement or noise artefacts.

Pipeline rankings were calculated by assigning a weight from 1-12 to each pipeline based on its reliable fraction. The top-ranked pipeline received a weight of 1, and the bottom-ranked pipeline received a weight of 12. Intra-session means and variances of the ranking scores were calculated for each pipeline. Low variances in weights indicates that ROCr is providing a similar ranking across days for that pipeline. High variances indicate that ROCr is providing rankings that vary across sessions for that pipeline.

2.6.2 Effects of Spatial Restriction on ROCr Performance

A common technique in MEG analysis is restricting data to an area of interest (35). To observe the effects of this technique on ROCr performance, some of the analysis was repeated using only the hemisphere where activation was expected. For right MNS, activation of S1 occurs in the left hemisphere. Two main modifications were done to the ROCr procedure outlined above. First, instead of generating a new set of split halves, the indices used for each of the split halves in the original analysis were recorded and used in the one hemisphere analysis. This ensures that differences in results are caused only by the absence of right hemisphere data. Second, a left hemisphere mask was applied to the activation maps prior to ROCr analysis to remove all right hemisphere activation. This modified analysis was done for each subject using the standard analysis pipeline, and the reliable fractions were compared to the standard pipeline results for the whole head.

2.6.3 Effects of Temporal Restriction on ROCr Performance

MEG data can be restricted to a time of interest. This is common in clinical paradigms where certain magnetic field deflection peaks (such as the 20 ms SEF peak) are used for localization. It is possible that the Overall Reliability and Reliability vs. Time methods are thrown off by reliable noise or artefacts in the data, but by restricting the analysis to a time window where the activity of interest occurs, the results may be more relevant. The first step in this analysis was identifying the time window of interest. The goal was to avoid the 0 ms artefact caused by the stimulation, as well as the later downstream SEF responses.

The ideal way to test this hypothesis would be to restrict the activation maps to the time of interest, and re-run ROCr on this reduced dataset. However, as this investigation is not crucial to my thesis, and because the ROCr calculations are time-intensive, an approximative method was used instead. The time window was selected by creating a grand average evoked field plot and finding the time window that met the criteria above. Next, the Reliability vs. Time values that fell within this window were averaged together to give a representative reliability value for the dataset. This modified analysis was done for each subject using the standard analysis pipeline, and the reliable fractions were compared to the unrestricted standard pipeline results.

2.6.4 Study 1 Outcome Measures

Several outcome measures were obtained in Study 1 at the subject level. A distribution of F_R is obtained using the Overall Reliability approach for each pipeline on day 1, 2 and 3 (50 x 12 x 3). A one-way ANOVA (analysis of variance) was done on each day's data (50 x 12) to calculate the mean F_R on each day and determine which means were statistically different from each other.

SPSS was used to perform the one-way ANOVA. The first step was calculating the Levene statistic to test for homogeneity of variances. If the variances of the F_R distributions are not homogeneous, we must use an alternative to ANOVA to determine if the means are not identical. The Welch test can be used if the variances are not equal. If the Welch test shows that the means are not identical, then the ANOVA post-hoc tests can be used to determine which means are homogenous. The last step was performing an ANOVA with the Tukey HSD (honestly significant difference) post-hoc test. The Tukey HSD test indicates which means cannot be statistically distinguished, or in other words, which means are homogenous.

The pipeline(s) with the highest F_R on each day were compared within subjects to address Hypothesis A. Each day's set of pipelines were sorted to create pipeline rankings on each day. For each pipeline, the across-days average, standard deviation and variance were calculated on the ranking. T-tests were then done to compare rankings of certain pipelines. These outcome measures were also obtained using the Reliability vs. Time approach.

For the effects of temporal and spatial restriction, the outcome measures were distributions of F_R on each day, for each subject. These distributions were compared to the standard pipeline F_R distributions obtained using the Overall Reliability approach and the Reliability vs. Time approach. A repeated measures factorial ANOVA was done to compare the effects of these methods. The repeated measures ANOVA was also done in SPSS. The first step was Maulchy's test for sphericity. If the assumption of sphericity was violated, the Greenhouse-Geisser estimate was used to correct for the violation. A pairwise comparison was then done to determine which means were significantly different.

At the group level, outcome measures included a summary of the pipelines with the highest F_R, the across-subjects average pipeline rankings (12), and the variance across pipelines (10).

2.7 Study 2: Relationships between Reliability and Inter-Session Variability

The ROCr framework may give a high reliability measure to an activation map, but that does not necessarily mean that activation map does a good job of localizing S1. As discussed previously, accuracy of functional neuroimaging data cannot be assessed without invasive methods, but other measures can. The Euclidean distance (ED) between S1 locations (see Figure 10) in each session can give a measure of how similar the localization results of activation maps are. Specifically, in this context the Euclidean distance is a measure of intra-session variability. ED has also been referred to as a measure of inter-session reliability by Solomon et al. (11), but to avoid confusion with intra-session reliability (the reliable fraction, F_R) this thesis will refer to ED as a measure of inter-session variability. A lower Euclidean distance indicates that inter-session variability is low, or in other words, the estimated S1 location is not changing much from one scan to the next (11).

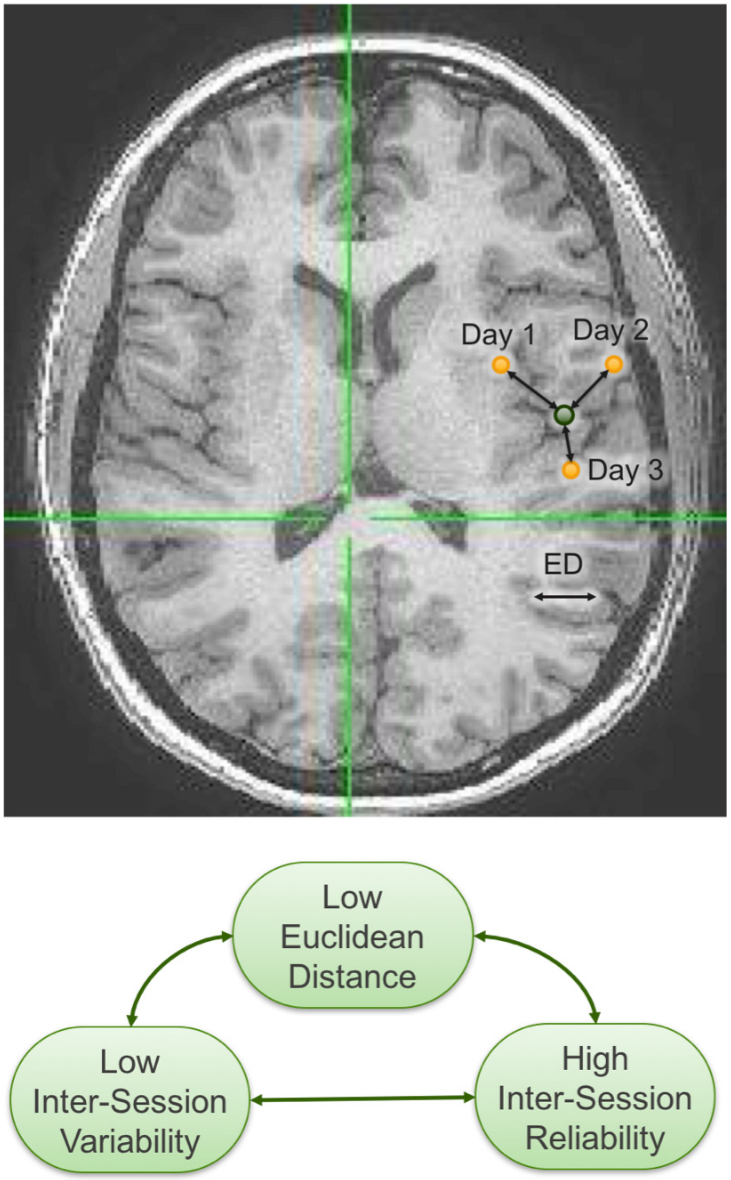


Figure 10: Euclidean distance diagram.
The orange dots indicate the S1 locations on Day 1, Day 2, and Day 3. The green dot indicates the across-days average S1 location. The black arrows connecting the dots indicate Euclidean distances. For each day, the Euclidean distance is defined as the distance between that day's S1 location and the average S1 location. Smaller EDs indicate higher intra-session reliability and lower intra-session variability. In the context of this thesis, Euclidean distance is a measure of inter-session variability. A low ED represents low inter-session variability.

Study 2 addresses Hypothesis B, wherein I expected that increasing intra-session reliability should reduce inter-session variability. If this hypothesis is true, it indicates intra-session reliability is a clinically useful metric. I test this hypothesis by determining if, for each subject, the pipelines with the highest intra-session reliability on each day have lower inter-session variability (as characterized by a low ED) than the standard pipeline.

Once the S1 locations were determined for each pipeline and each session, calculations were done to determine EDs for each session's S1 location as shown in Figure 10. Briefly, the EDs were calculated by averaging the S1 locations on each of the three days, then calculating the distance between each day's location and the average location. This data was used in two different ways. First, I compared the EDs for the ROCr-selected pipelines with the EDs for the standard pipelines. Secondly, I examined the relationship between intra-session reliability (F_R) and inter-session variability (ED) across all data collected in the study.

2.7.1 ROCr vs. Standard Pipeline

Hypothesis B speculates that, by using the ROCr-selected pipeline to localize S1, the inter-session variability should be lower than if the standard pipeline were used. The Euclidean distance is a way to quantify the inter-session variability.

The first step to test this hypothesis for a given subject was calculating the average ED for the standard pipeline. Each day's data was analyzed using the standard pipeline, and an S1 location was obtained for each day. These S1 locations were averaged together, and for each day an ED was calculated (see Figure 10). These three EDs were averaged together to obtain the standard pipeline average ED for that subject.

The next step was calculating the ROCr-selected pipeline average ED. This was done in a similar manner to the standard pipeline average ED. The difference was that, instead of using the standard pipeline on each day, each day used the activation maps associated with the most reliable pipeline, as determined in Study 1. This meant that each day may have used a different pipeline.

Once the ED data were collected, statistical analyses were done on the data. For a given day, there are two distributions of 50 EDs: one for the ROCr-selected pipeline and one for the standard pipeline. These distributions were compared using a t-test. This step is repeated for each day, for each subject. Next, for each subject, the ED distributions were combined across days, giving two distributions of 150 EDs (one distribution from the ROCr-selected pipelines, one from the standard pipeline). These distributions were also compared using a t-test. At the group level, distributions were combined across subjects, giving two distributions of 1800 EDs (one distribution from the ROCr-selected pipelines, one from the standard pipeline). These distributions were compared using a paired t-test.

2.7.2 Intra-Session Reliability vs. Inter-Session Variability

For the second part of Study 2, inter-session variability was calculated for all pipelines, not just the ROCr-selected or standard pipelines. For each subject and pipeline, the ED was calculated for each of the 50 splits on each day, which required a different definition of the average S1 location. Instead of an across-session average S1 location, an across-splits average S1 location was calculated on each day. Following this analysis, each split had a measure of intra-session reliability (F_R) and inter-session variability (ED).

2.7.3 Study 2 Outcome Measures

Several outcome measures were obtained in Study 2 at the subject level. On each day, two distributions of S1 locations were obtained: one from the standard pipeline, and one from the ROCr-selected pipeline. For each subject and day, a distribution of inter-session variability measures (EDs) was calculated for the standard pipeline and ROCr-selected pipeline. A t-test was done on each day’s data to compare the EDs from the ROCr-selected pipeline and the standard pipeline. The EDs were then pooled across days, and another t-test was done to determine if there was a difference between the ROCr-selected pipeline EDs and the standard pipeline EDs for that subject. At the group level, a paired t-test was done on the subject-level data to determine if there was a difference between the ROCr-selected pipeline EDs and the standard pipeline EDs at the subject level. These outcome measures address Hypothesis B.

Additional outcome measures were obtained to investigate the relationship between F_R and ED. EDs and F_Rs were obtained for all pipelines (not just the standard and ROCr-selected pipelines), giving 360 distributions of each (12 pipelines, 3 days, 10 subjects). A linear regression was done on these data to determine if there is a relationship between intra-session reliability (F_R) and inter-session variability (ED).

CHAPTER 3 RESULTS

3.1 Pre-experiment Results

Some basic investigations were done before the experiment to determine which parameters to use. The first investigation was done to determine which cut-off frequencies to use for low-pass filtering. A TFR plot was produced as shown in Figure 11.

From this plot, key signal frequencies were identified (solid green lines). The strongest group of frequencies was around 10Hz, followed by the groups around 25 Hz and 40Hz. The final group identified was around 60Hz, indicating noise in the data from power lines. Cut-off frequencies were identified (dotted blue lines) to retain or remove the key signal frequencies. The conservative cut-off frequency was 55 Hz, removing the known 60Hz noise but not affecting the rest of the signal. The aggressive cut-off frequency was 15 Hz, retaining only the group of frequencies around 10Hz. The normal cut-off frequency was at 35 Hz, half way between the aggressive and conservative frequencies.

The next test done was determining how many splits were required before the reliable fraction would converge. Figure 12 shows the results of this analysis. After approximately 50 splits, the average reliable fraction converged to the same value as when 100 splits were averaged together.

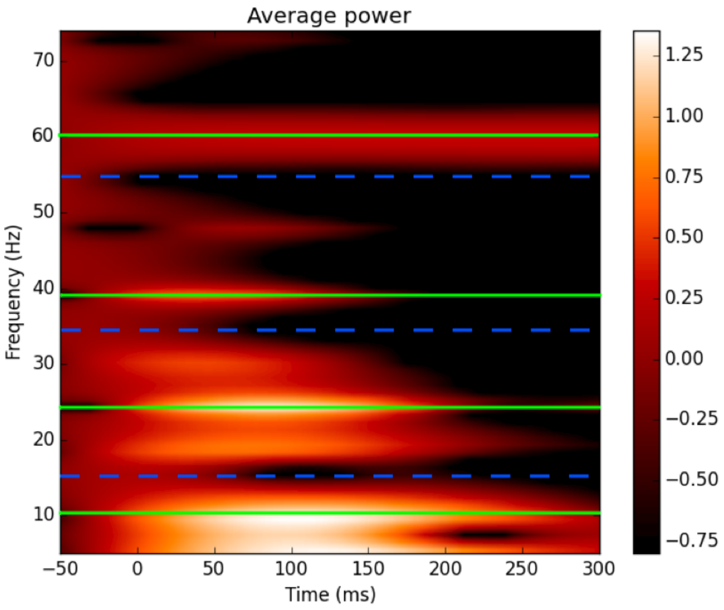


Figure 11: TFR plot of grand average data.
X-axis is time in ms. Y-axis is frequency. Dark areas indicate low average power, light areas indicate high average power. Green lines indicate key signal frequencies (10Hz, 25 Hz, 40Hz, 60Hz). Blue dotted lines indicate the selected cut-off frequencies (15 Hz, 35 Hz, 55 Hz).

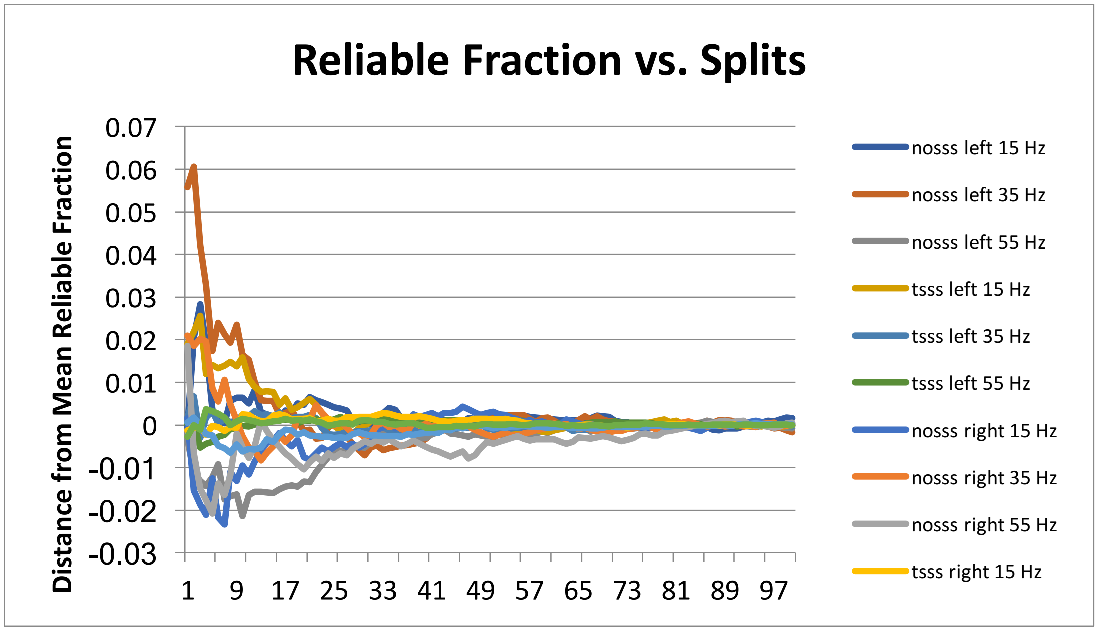


Figure 12: Reliable fraction convergence (preliminary pipelines).
X-axis indicates number of splits. Y-axis indicates the difference from the mean reliable fraction for each pipeline. All reliable fractions converged after approximately 50 splits. This chart was created with six preliminary pipelines used to analyze left and right hemisphere data.

3.2 Preliminary Results, Single Subject

The representative subject for this study was found by calculating the average reliable fraction across sessions for each subject, and comparing to the group average. Subject rt006 had an average F_R closest to the group average, and a visual inspection confirmed that this subject did not have abnormal results. All single-subject results discussed below are from subject rt006.

3.2.1 MEG Analysis Results

Figure 13 a) and b) show the MEG data results for the standard pipeline on day 1. Figure 13 a) is the topography plot, and Figure 13 b) is the averaged epochs at each sensor over time (called butterfly plots). The topography plot shows SEF activation in the left hemisphere, which is to be expected. The SEF appears contralateral to the stimulus, so right hand MNS should produce an SEF in the left hemisphere. This subject had peak magnetic field deflections of 321 to 429 fT. For all datasets (across subjects and sessions) magnetic field deflections varied from 23 fT to 295 fT. On average, each dataset's magnetic field deflections varied by 110 fT across pipelines.

The initial spike in magnetic field deflection seen in the butterfly plot around 0 ms in is the electric stimulus applied at the wrist being detected by the MEG sensors. The 20 ms peak (N20m) can be seen, but the 35 ms peak (P35m) is not clear. There is a clear deflection at 60 ms peak in the SEF, but for this subject the strongest dipolar magnetic field deflections occurred around 80 ms. Similar trends are seen with other pipelines that have an LPF cut-off frequency of 55 Hz. These waveforms are similar to SEF waveforms obtained for clinical use in pre-surgical mapping, but do not have a strong enough N20m response to map S1. This is especially true for pipelines with a LPF value of 15 or 35 Hz, such as Pipeline 10 shown in Figure 13 c) which has a LPF value of 15 Hz. With this pipeline, all the peaks are smoothed and blurred. Only the later response around 80 ms remains. Figure 13 d) shows the grand average MEG data across all subjects and sessions. With enough epochs, the N20m and P35m are clearly visible.

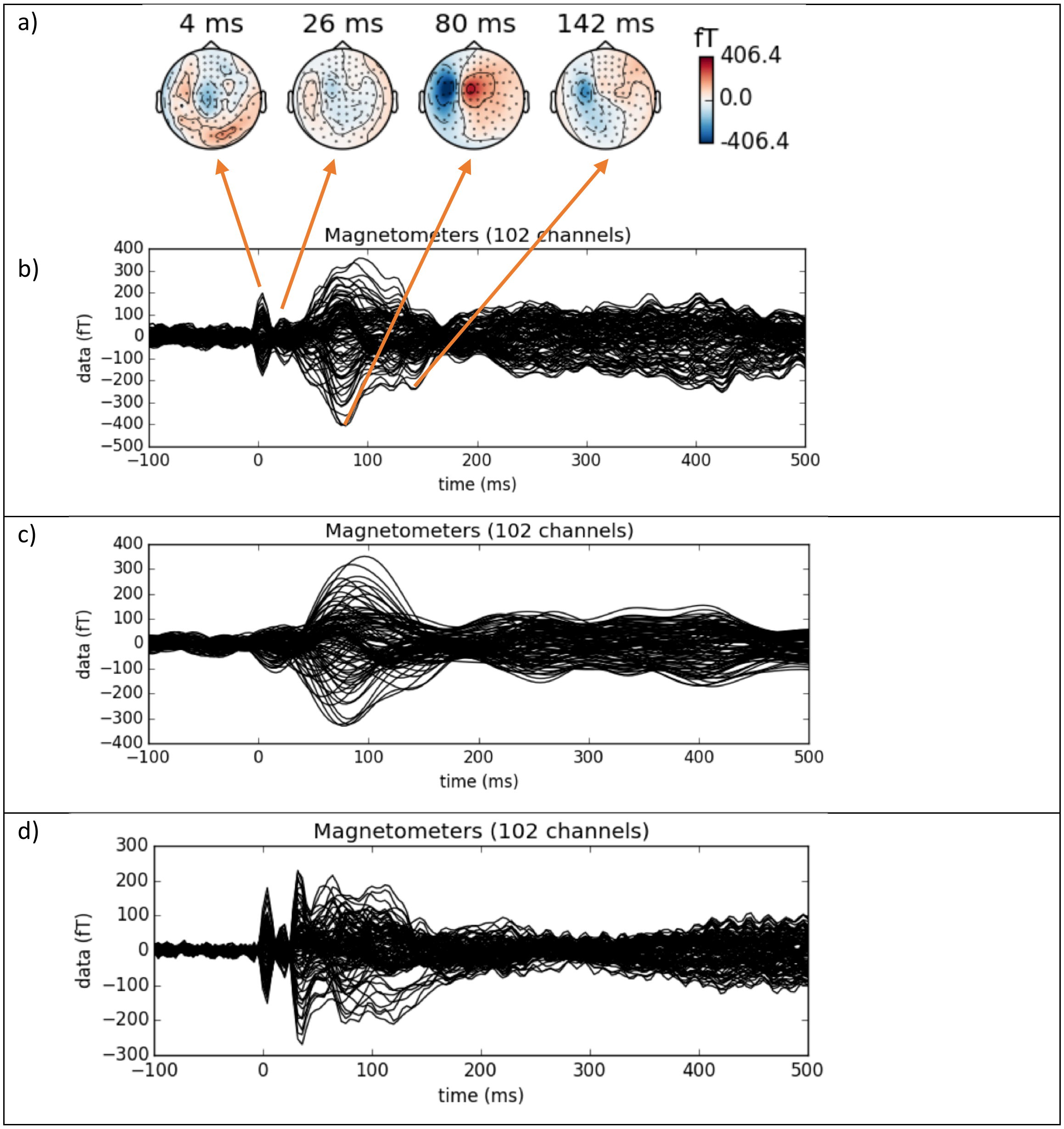


Figure 13: Session 1 MEG magnetometer data, subject rt006.
(a) Topography plots showing magnetic field deflections measured by 102 magnetometers at times of interest. Magnetometer locations are indicated by dots. Neurological convention is used for these plots (the left hemisphere is shown on the left, and the right hemisphere is shown on the right). The nasion is at the top of each image. Red indicates magnetic field deflections exiting the head, and blue indicates magnetic field deflections entering the head. Topography plots at 26 ms and 80 ms exhibit a dipolar pattern, indicating the location of S1. (b), (c), (d) Averaged magnetic field deflections measured at each sensor. 102 sensor measurements are included in each plot. Figure (b) shows results for the standard pipeline (tSSS, DC HPF, 55 Hz LPF). The N20m peak is visible, but weak. The later response around 80 ms has the largest peak. Figure (c) shows results for Pipeline 10 (tSSS, 1 Hz HPF, 15 Hz LPF). With this pipeline, the N20m peak is not visible. The entire waveform is smoothed due to the aggressive filtering. Figure (d) shows the grand average across all subjects and sessions. With a sufficient number of epochs, the early responses (N20m and P35m) become sharp.

Preliminary investigation into the MEG data revealed that the magnetic field deflection waveforms varied not only across pipelines, but also across sessions. Figure 14 shows butterfly plots for three of the pipelines used to analyze data from subject rt006. Each row shows results for each session, and each column shows results for each pipeline. Compared to the standard pipeline, Pipeline 12 did not have a large effect on magnetic field deflections; the waveforms are nearly identical. This trend was observed across all 55 Hz LPF pipelines. Pipeline 4 had a much more noticeable effect on the waveforms, likely due to the aggressive LPF. This trend was observed across all 15 Hz LPF pipelines. Surprisingly, the session seems to have had a larger effect on the deflections than some of the pipelines did. The change in waveforms between a), d) and g) is much more noticeable than the change in waveforms between a) and b), d) and e), and g) and h). This indicates that inter-session variance is not as low as assumed.

The end result of the MEG analysis was an activation map for each pipeline and day. Each 5mm voxel has an assigned probability that the voxel is active at a temporal resolution of 250 Hz. These activation maps over time can be used to estimate the location of S1 by finding the voxel with the highest probability of activation. Figure 15 shows the activation map for session 1, subject 6, using the standard analysis pipeline. The area with the highest probability of activation (shown in red) is in the left hemisphere, around the expected region for the primary somatosensory cortex.

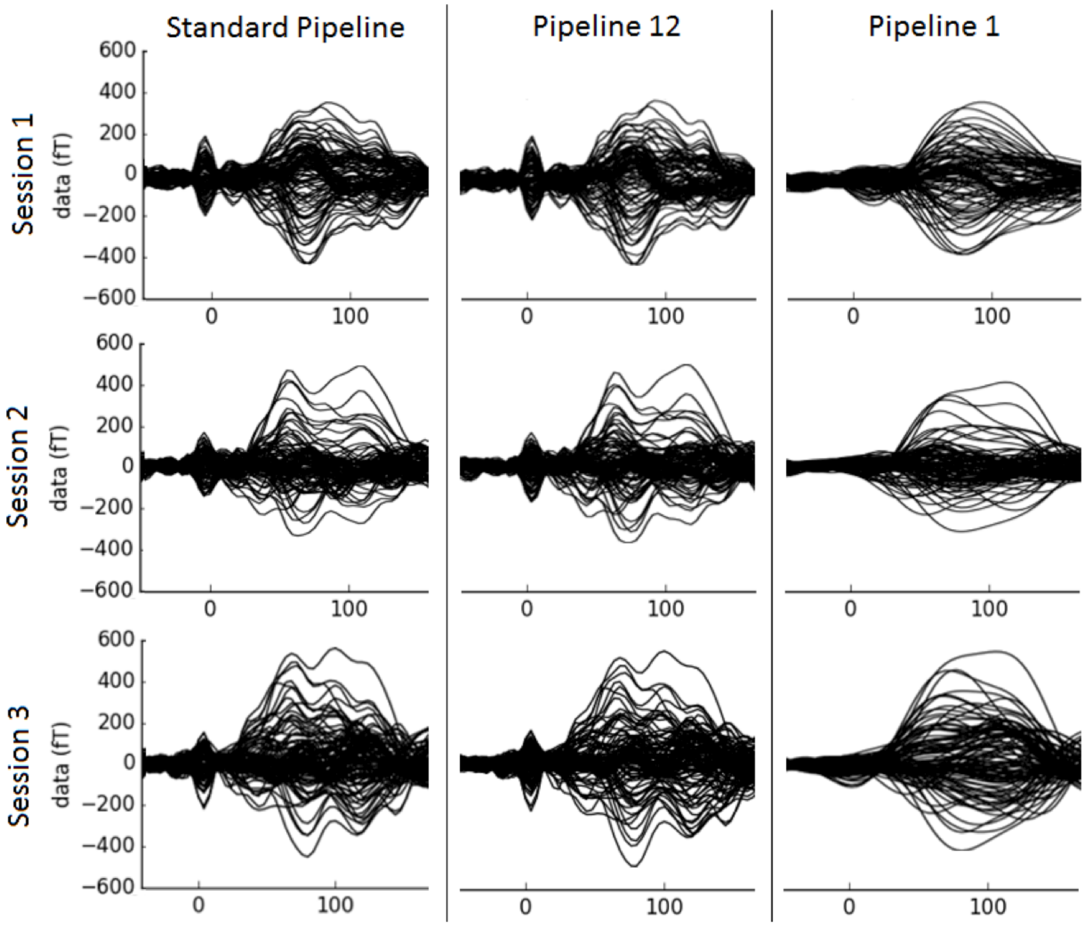


Figure 14: Comparison of MEG data across sessions and pipelines, subject rt006.

Each column represents a different pipeline, and each row represents a different session. The standard pipeline and Pipeline 12 vary in terms of HPF and ICA. The difference in magnetic field deflections across these two pipelines is minimal. The difference across sessions appears to be larger. This behavior is typical of pipelines with a 55 Hz LPF. The standard pipeline and Pipeline 4 are the same except for the LPF frequency. In this case, the difference across pipelines is larger than before, but each session is still distinct and shares characteristics with other pipelines on the same session. This behavior is typical of pipelines with a 15 Hz LPF.

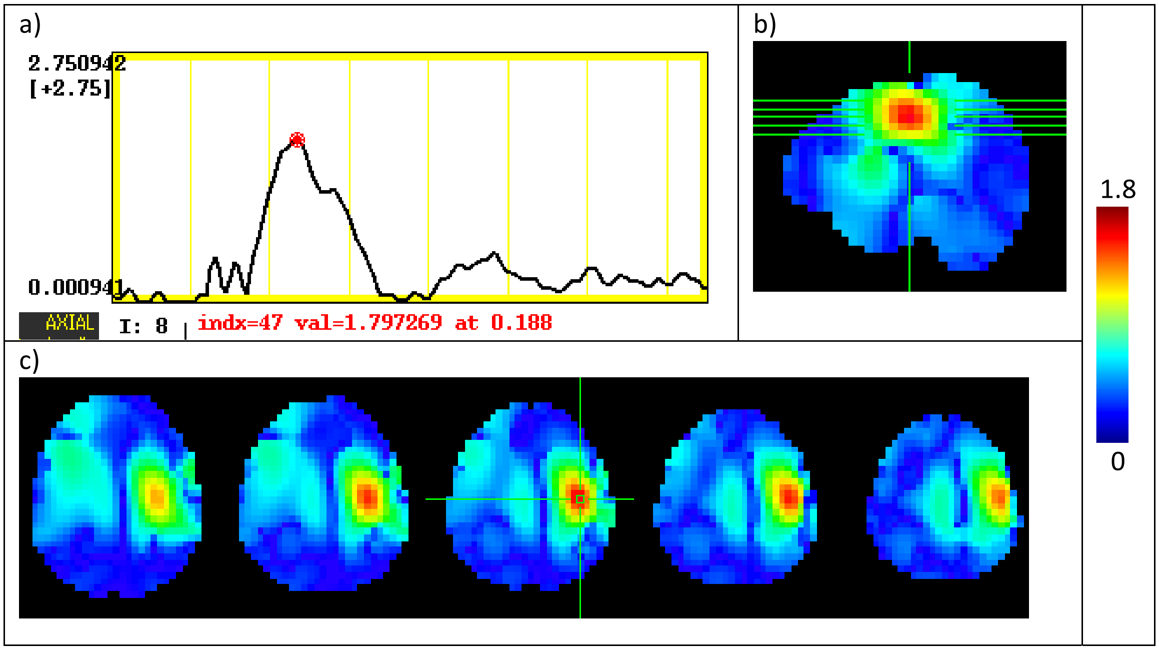


Figure 15: Session 1 Activation Map, Subject rt006.
Voxel size = 5mm. The crosshairs indicate the estimated location of S1, at the voxel with the highest activation probability. (a) shows the activation vs. time for the estimated S1 location. The peak activation occurs at 88 ms. Smaller peaks occur earlier. (b) shows a sagittal view of the activation. The red area indicates the area of highest activation. (c) Axial slices of the activation map (locations shown as the horizontal green lines in the sagittal view).

Figure 16 shows activation maps for the standard pipeline on session 1 at different latencies. Figure 16 a) shows the activation at 4 ms, shortly after the MNS was applied to the wrist. The activation shown here is an artefact of the electrical stimulus detected by the highly sensitive MEG sensors. The stimulus is administered at time zero, and it takes time (about 20 ms) for the signal to reach the brain and for the brain to respond. Also, any brain activity that is not related to the stimulus was averaged out in the analysis pipeline.

Figure 16 b) shows the activation map at 26 ms, around the time of the first SEF peak. The white crosshairs at x, y, z coordinates (40, 0, 10) indicate the peak voxel, which is the estimated S1 location at this time. The estimated location is in the left hemisphere, as expected. The activation is 0.872; low compared to the other latencies, but that is to be expected as these data were not recorded at a high enough sampling rate to accurately localize the 25 ms or 35 ms peak. Even though the activation is weak, it appears to be more focal than Figure 16 c), and less noisy.

Figure 16 c) shows the activation map at 88 ms, the time with the highest activation. The estimated S1 location is indicated by the white crosshairs at (40, 0, 20). Again, the estimated location is in the left hemisphere. This time the peak activation is 1.79, more than double the peak activation at 26 ms. The estimated S1 location at 88 ms is located superior to the estimated location at 26 ms, and there is also activation in the right hemisphere. This difference in location and laterality could indicate that we are seeing downstream activation of S1 and connected areas. With more sources of activation appearing in the data, it may be challenging to localize the area expected by clinicians.

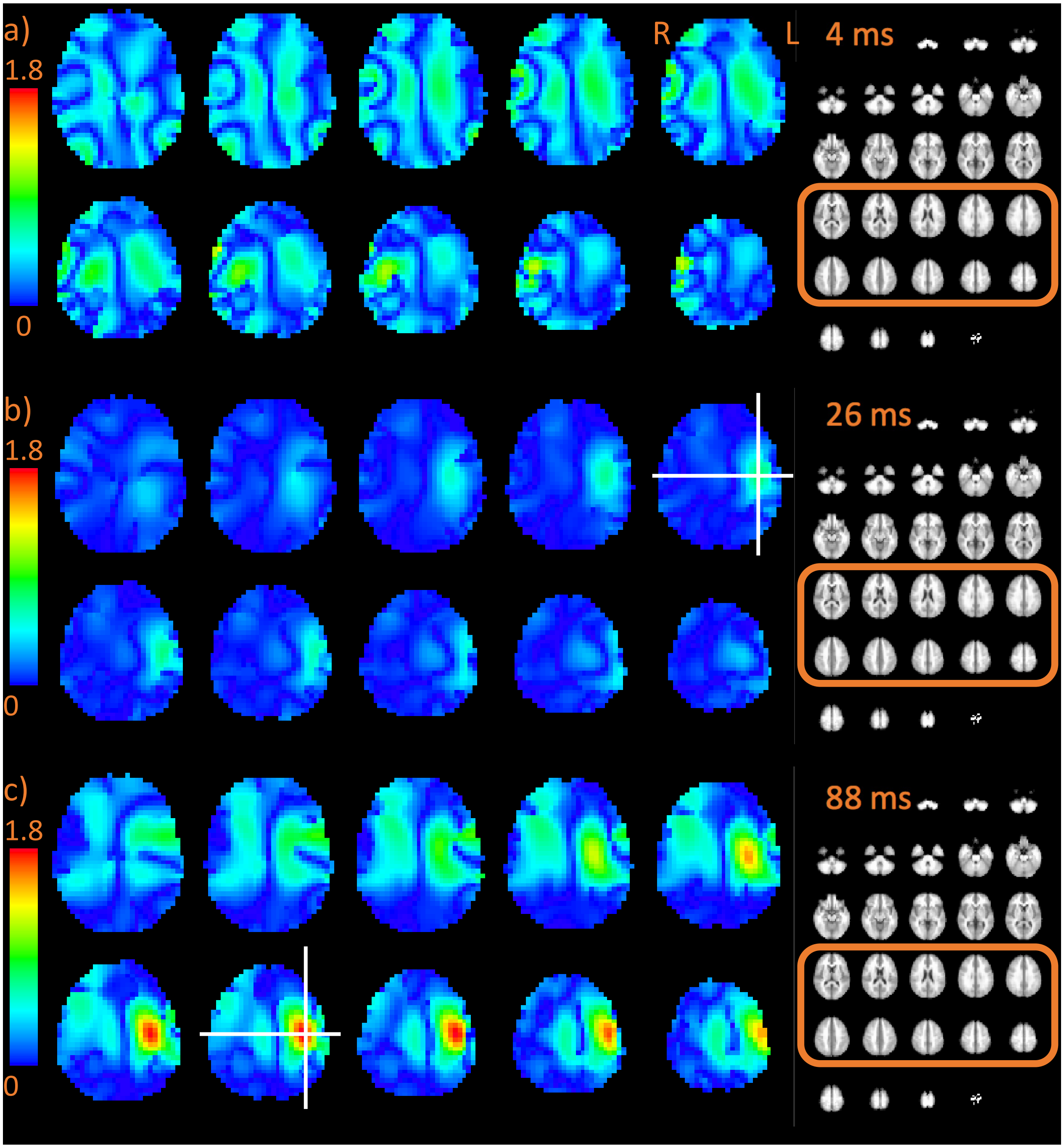


Figure 16: Session 1 Activation Maps at Times of Interest, Subject rt006.
Axial slices are shown, using the radiologial convention. Slices are 5mm apart. The colour of each voxel inidcates its probability of being active. Dark blue areas indicate zero probability, red areas indicate high probability. a) Activation at 4 ms (caused by MNS pulse being detected by MEG sensors). b) Activation at 26 ms, the time with highest activation around the expected N20m peak. Estimated S1 location indicated by crosshairs. The activation is low, but focal. c) Activation at 88 ms, the time with the highest activation. Estimated S1 location indicated by crosshairs. The activation is at its peak here, but is less focal than at 26 ms. There is low activation in the right hemisphere, indicating a bilateral response.

3.2.2 ROCr Analysis Results

3.2.2.1 Overall Reliability

For ROCr analysis, two activation maps were produced for each of the 50 dataset splits. These maps were then compared using the ROCr framework. For Overall Reliability, one AUC plot was produced for each split. For each of the 50 AUC plots, a reliable fraction was calculated, giving a distribution of 50 F_R values. The 50 plots were averaged together, giving the session 1 AUC plot and F_R shown in Figure 17 a). Likewise, the 50 F_R values were averaged together, giving a reliable fraction of 0.748 for session 1. Figure 17 b) and Figure 17 c) show the AUC curves and F_R values for sessions 2 and 3 respectively. Compared to sessions 1 and 3, session 2 has a slow initial increase in AUC for an increase in threshold, and has a correspondingly lower F_R. Figure 17 d) shows a histogram of the 150 F_R values represented in a) b) and c). The distribution has three peaks, indicating that each day has its own distribution of reliability.

Figure 18 shows average AUC plots for each pipeline on each day for subject rt006. The AUC curves on day 1 and day 3 (Figure 18 a) and c) respectively) are relatively similar, but the curves on day 2 (Figure 18 b)) have a variety of rise times. Figure 18 d) shows a histogram of all the F_R values for subject rt006. At this level, the F_R values form a left skewed distribution, and the upper limit of reliability can be estimated for this subject. Subject rt006 had only a handful of splits with F_R values above 0.825.

Figure 19 shows a histogram of all reliable fractions, across all sessions and pipelines. As with subject rt006, the distribution is left-skewed, but to a lesser degree. Since the reliable fraction cannot exceed 1, this slight skew is to be expected. There are outliers around 0.90, and some shelf-like behaviour around 0.85. These abnormalities may be the result of a distinct pipeline or session within an individual subject.

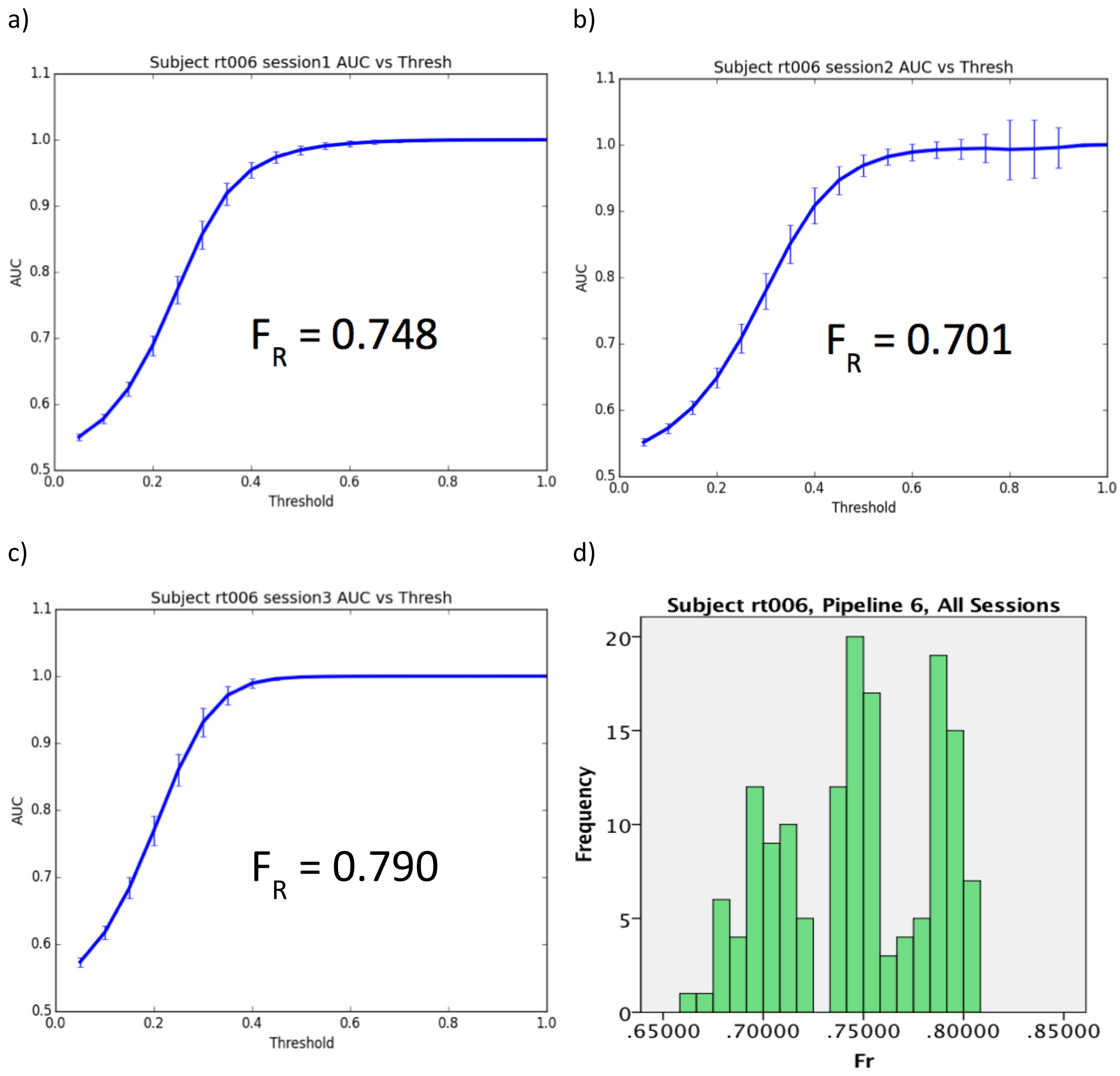


Figure 17: AUC curves and reliable fractions, standard analysis pipeline, subject rt006.
(a) Session 1 (b) Session 2 (c) Session 3. d) Histogram of reliable fraction values across days. Error bars on plots a), b) and c) indicate the standard deviation at that data point. The histogram shown in (d) has three peaks around the average reliable fractions from sessions 1, 2 and 3.

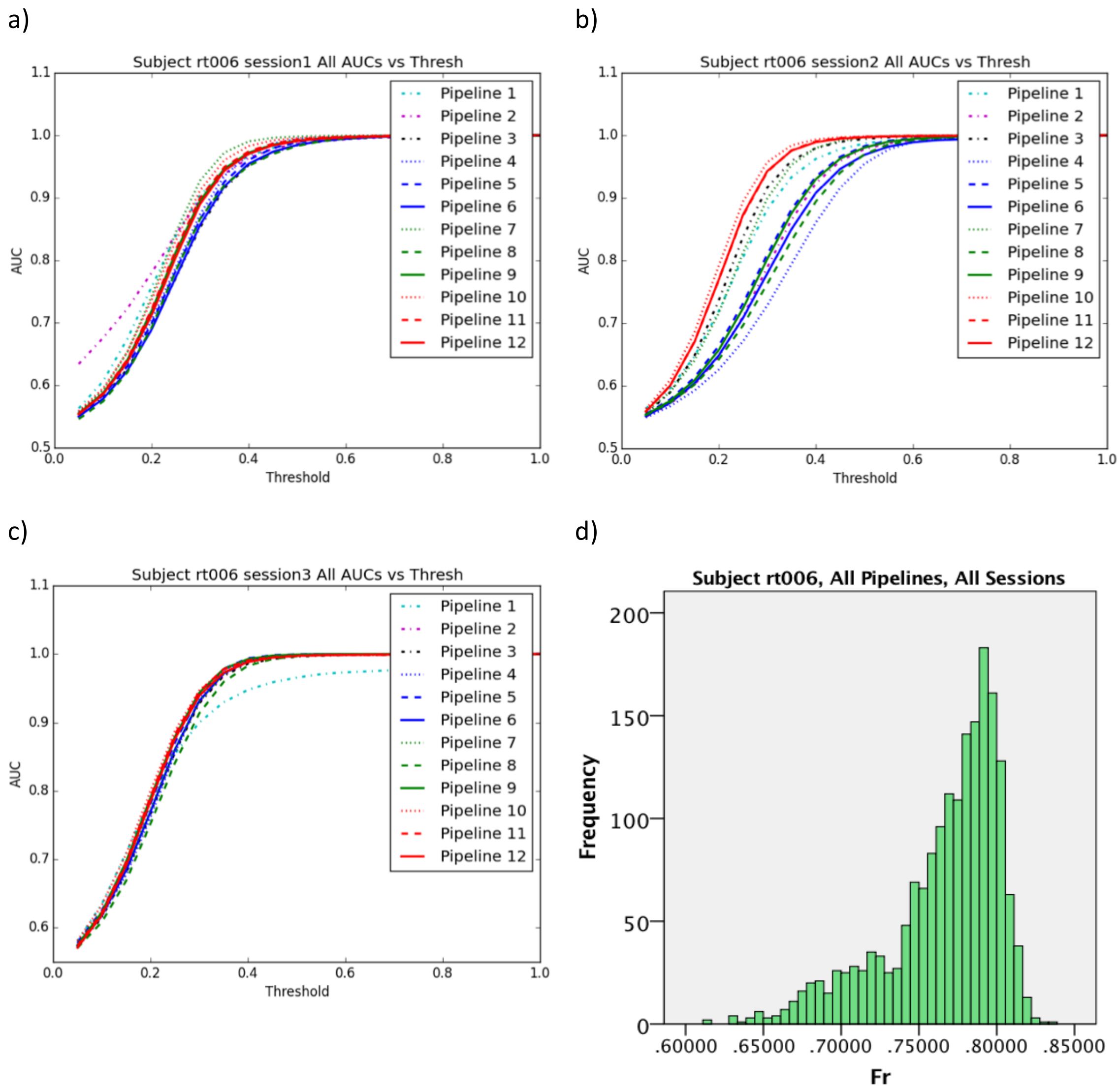


Figure 18: AUC curves, all pipelines, subject rt006.

a) Session 1 (b) Session 2 (c) Session 3. d) Histogram of reliable fraction values across days. Error bars on plots a), b) and c) indicate the standard deviation at that data point. The histogram has a peak around 0.79, and is skewed to the left.

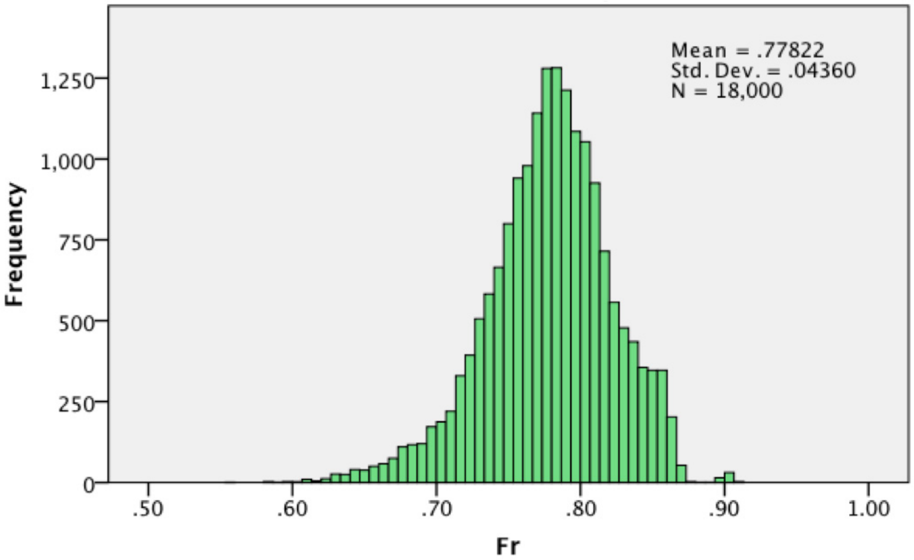


Figure 19: Histogram of all reliable fractions, across all subjects and pipelines.
The distribution is slightly left-skewed, and there are outliers around 0.90. There is shelf-like behaviour around 0.85.

3.2.2.3 Reliability vs. Time

For Reliability vs. Time, a distribution of 50 F_R values was produced at each time point in the dataset. These F_R values were averaged together and plotted versus time, giving the plot shown in Figure 20 a) below. Figure 20 b) and c) show the plots for sessions 2 and 3 respectively. Recall that the standard pipeline used to analyze these data uses a LPF of 55 Hz (rather than 35 Hz or 15 Hz), so we expect the F_R vs. Time plots to have high frequency components up to 55 Hz.

The baseline interval from -100 to 0 ms has low reliability. Each plot shows a spike in reliability around 4 ms, immediately after the stimulus was delivered. After this initial spike, the plots differ, but each has its peak between 80 and 120 ms, and by 200 ms the reliability decreases to near baseline levels. Figure 20 a) has a peak in reliability around 25 ms. It is common to see variance in the basic waveform shape across subjects and across sessions within subjects. For example, there is no peak in reliability around 25 ms in Figure 20 b) and c). However, in these figures the reliability is quite high around 25 ms and remains high. The absence of a peak does not necessarily mean that the N20m or P35m is not reliable if reliability is high over a wide time range.

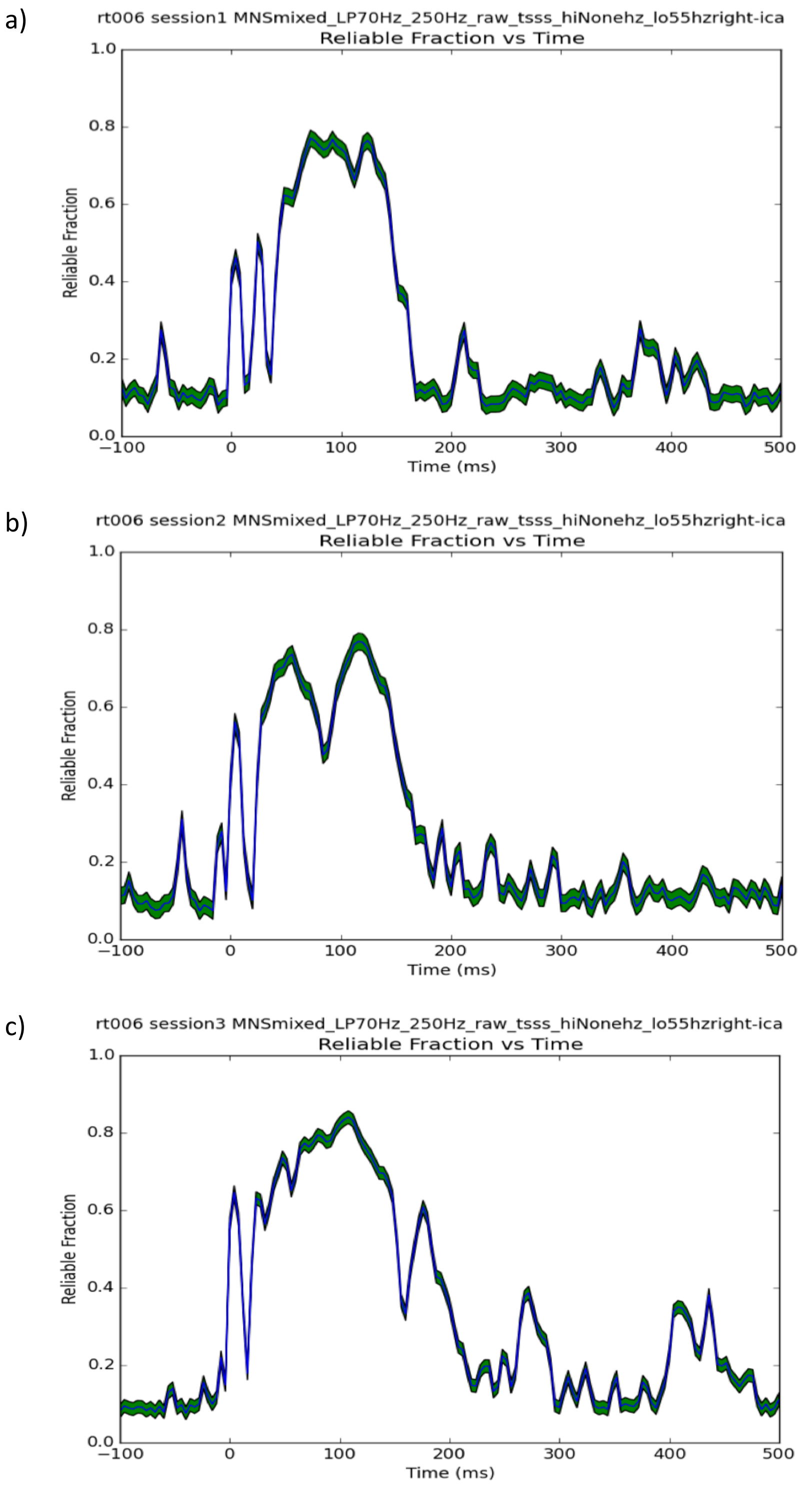


Figure 20: Reliability vs. Time Charts for subject rt006, standard pipeline.
The x axis is time, from -100 to 500 ms. The y axis is the reliable fraction, from 0 to 1. Vertical line thickness indicates standard deviation. The waveform shape varies across days, even when the same pipeline is used. A) Session 1. In this plot there is a peak in reliability that may correspond to the N20m or P35m. Peak reliability occurs around 80 ms. B) Session 2. c) Session 3

Reliability vs. Time charts were obtained for each pipeline, and give insight into how each pipeline affects the data. Figure 21 shows the Reliability vs. Time charts for each pipeline. There are common features across all pipelines: there is a spike in reliability around 0 ms resulting from the median nerve stimulation applied at the wrist, and there is a period of prolonged high reliability from around 50 ms to 150 ms.

The exact shape of the waveform depends on the pipeline parameters. For example, in each of the pipelines with a LPF cut-off of 55 Hz (1, 2, 3, 6, 9, 12), there is a reliability peak that may correspond to the N20m or P35m. As the LPF cut-off frequency lowers, the waveform becomes smoother, and this peak is less visible or absent.

The SSS pipeline differs from all other pipelines in that the reliability is much higher on the 0-40 ms range. This may indicate that this pipeline does not reduce the effects of the wrist stimulation as effectively as the other pipelines, or that this pipeline better captures the early SEF responses. Fifteen out of 30 datasets exhibited higher reliability in the -100 to 50 ms time range compared to the other pipelines. Subjects exhibiting this behaviour in one or more of their three sessions also had one or more instances of Pipeline 1 being selected by ROCr.

Figure 22 a) shows a histogram of the times of peak reliability across all subjects and pipelines, with a grand average shown in Figure 22 b) for reference. Very rarely was the MNS artefact selected as the peak reliability. The P35m peak was selected often, but the N20m was rarely selected. The histogram has two main peaks around 60 ms and 125 ms.

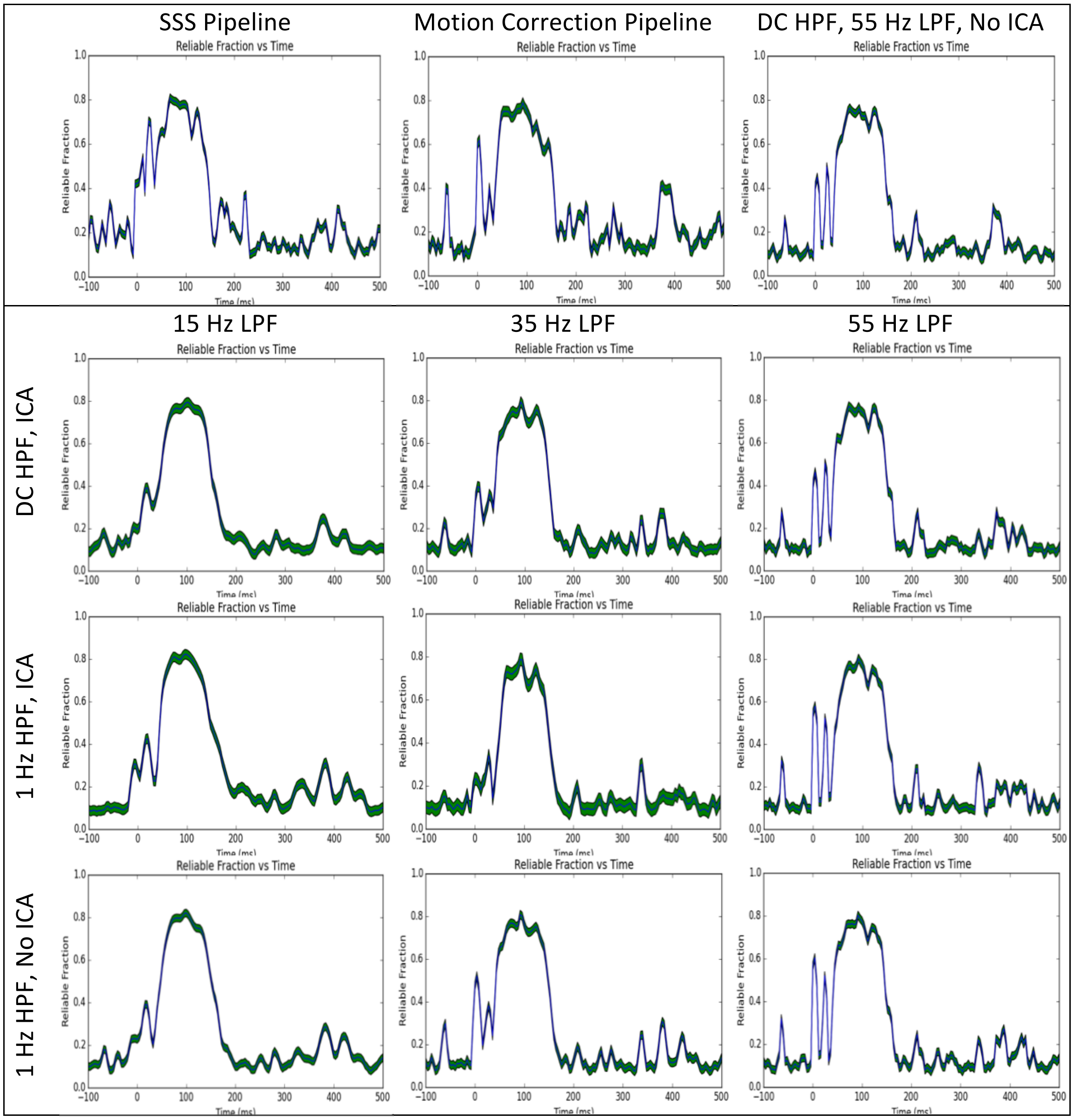


Figure 21: Reliability vs. Time plots, Subject rt006, Session 1, All pipelines.
The x axis is time, from -100 to 500 ms. The y axis is the reliable fraction, from 0 to 1. The first three pipelines are unique, but the other nine pipelines can be grouped based on LPF, HPF and ICA. Moving from left to right is an increase in LPF cut-off frequency. Pipelines 3-6 have DC HPF and use ICA. Pipeline 6 is the standard pipeline. Pipelines 7-9 use 1 Hz HPF and use ICA. Pipelines 10-12 use 1 Hz HPF, but do not use ICA. Each pipeline has a spike in reliability around 0 ms, corresponding to the median nerve stimulation applied to the wrist at t=0. From around 50 ms to 150 ms is a period of high reliability. This basic waveform shape is modified by different pipeline parameters, especially low-pass filtering. As the cut-off frequency is decreased, the waveforms become smoother and lose temporal resolution.

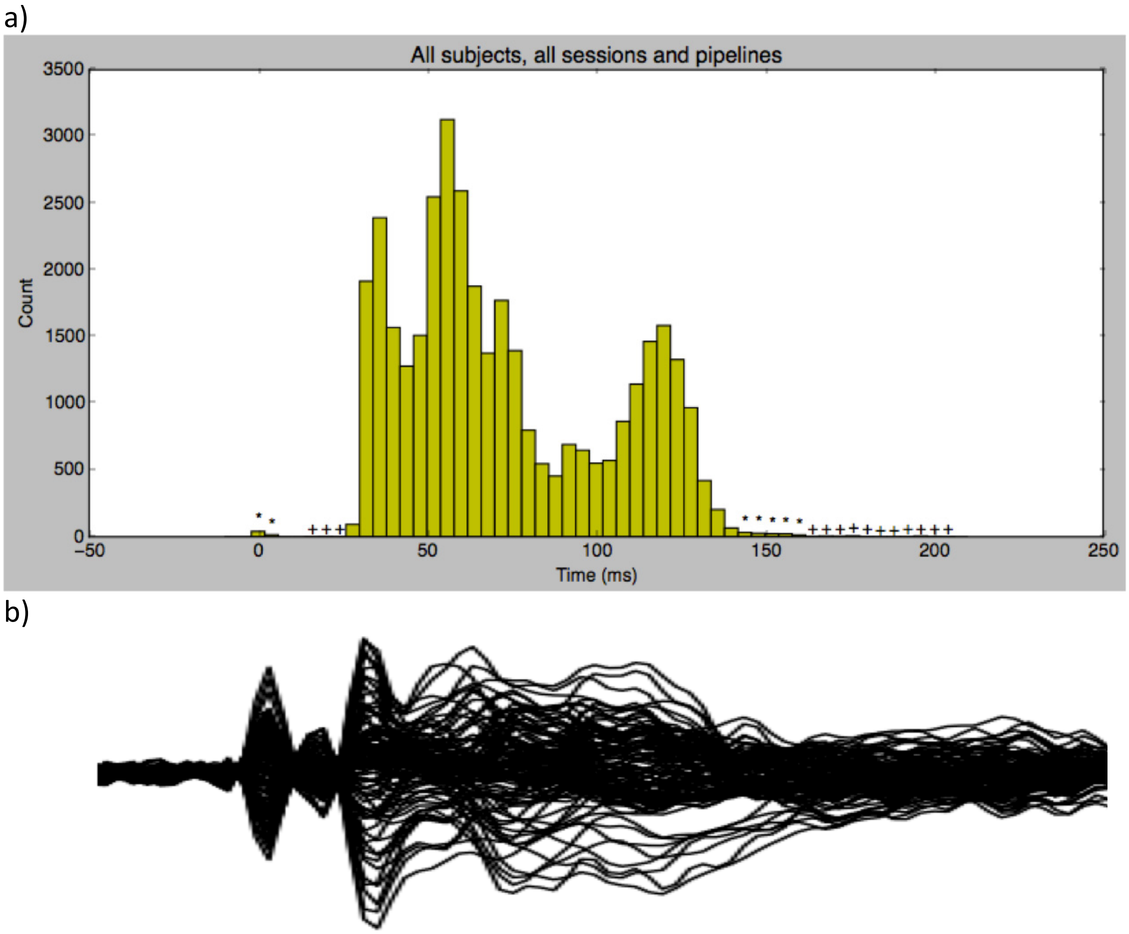


Figure 22: Peak reliability timing.
(a) Histogram showing times of peak reliability in each split. Results are included for all pipelines and subjects. X-axis indicates time in ms. Y-axis indicates the number of splits where peak reliability occurred at a given time bin. '*' indicates that there are 50 or fewer counts in the bin. '+' indicates that there are 10 or fewer counts in the bin. Peaks occurred around 35 ms, 58 ms and 120 ms. (b) Grand average evoked data.

3.4 Study 1: Pipeline Performance and Ranking

3.4.1 ROCr Pipeline Selection

ROCr was used to select the pipeline(s) with the highest reliable fraction on each day. For each dataset, the average reliability was calculated for each pipeline, and the pipeline(s) with the highest average reliable fraction was termed the "ROCr-selected pipeline." This was done with two approaches: the "Overall Reliability" approach and the "Reliability vs. Time" approach.

3.4.1.1 Overall Reliability

Figure 23 shows the average pipeline reliabilities for subject 6. Each pipeline has a reliable fraction for day 1, 2 and 3, shown in blue, red and green, respectively. The highest-performing pipeline on each day is indicated by a blue dot. On day 1, Pipeline 1 was the best. On day 2 and day 3, Pipeline 11 was the best. Hypothesis A was not supported by this subject because ROCr did not select the same pipeline on each day.

Day 3 had higher reliability across all pipelines, whereas Day 2 had low reliability results across many pipelines. The data session had a significant effect on the mean reliable fraction across pipelines. The average reliable fraction for Day 3 was 0.052 greater than for Day 2, and 0.029 greater than for Day (?). This difference in performance could indicate that Day 3's data was higher quality. Because Day 2 has low reliability on many pipelines, this could indicate a) low data quality, but with the right pipeline parameters we can get reliable results, or b) certain pipelines have a negative effect on data quality for this day. This is an interesting case, especially given the poor performance from the standard pipeline.

At the group level, session once again had a significant effect on the mean reliable fraction across pipelines. The average reliable fraction for Day 3 was 0.014 greater than for Day 2, and 0.018 greater than for Day 3. The difference between Day 2 and Day 3 was also statistically significant. This may indicate that the variance in data between days is much larger than initially hypothesized.

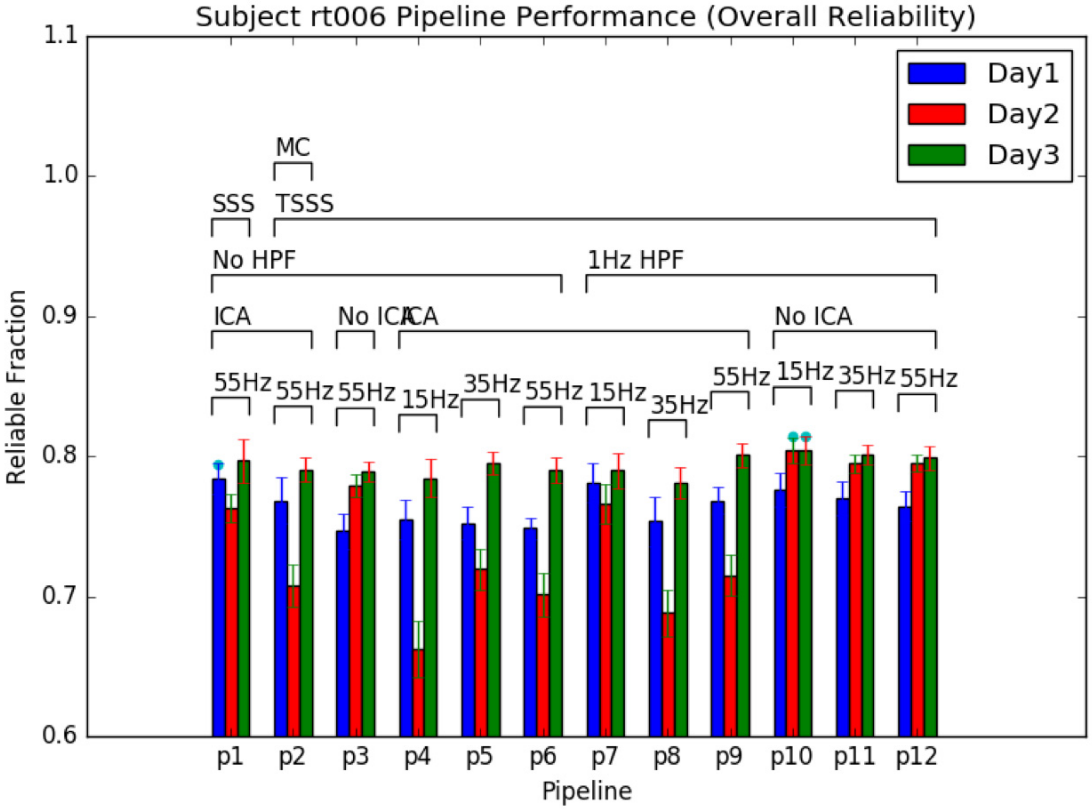


Figure 23: Pipeline Performance, Single Subject (rt006), Overall Reliability.
The x-axis shows the pipeline used, and the y-axis shows the average reliable fraction. Blue, red and green bars indicate average reliability on day 1, 2 and 3 respectively. Blue dots indicate the pipeline with the highest average reliable fraction. Error bars indicate reliable fraction standard deviation. Day 2 and Day 1 results have significantly lower (p < 0.001) average reliability than Day 3.

This procedure was repeated for each subject. Additionally, a one-way ANOVA was done for each session to identify which pipelines were not significantly different from the pipeline with the highest reliability. Table 3: summarizes the pipelines with highest reliability on each day, for each subject. Pipeline 9 was selected the most frequently (15/30), followed by Pipelines 12 and 1 (13/30). Pipeline 5 was never selected by ROCr, and Pipelines 4 and 6 were only selected twice each.

Scores were assigned based on how consistent the highest-reliability pipelines were across days. A score of 4 (worst) indicates that there was no overlap in the ROCr-selected pipelines across days. A score of 3 indicates that ROCr selected the same pipeline on two of the three days. A score of 2 indicates that there were at least two ROCr-selected pipelines on two of the three days. A score of 1 indicates that there was a common ROCr-selected pipeline across days. Subjects with a score of 1 meet Hypothesis A. Three subjects had a score of 1: rt001, rt002 and rt010. Three subjects had a score of 2, two subjects had a score of 3, and the remaining two subjects had a score of 4.

Subject	Day 1	Day 2	Day 3	Score
rt001	9**	9**, 7, 8*, 2	8*, 9**	1
rt002	9*, 12, 11**	8*, 1, 11**, 9*, 12	8*, 11**	1
rt003	1, 9*	12*, 3, 6	9*, 12*	2
rt004	12, 9, 1*	10	1*	3
rt005	1	12	10, 7	4
rt006	1, 7, 10**	10**	10**, 11, 9, 12	1
rt010	1*	1*	2, 9, 12, 8	3
rt011	12, 9	1	11, 8	4
rt012	1, 12*, 9*	12*, 8*, 11, 1	9*, 8*	2
rt015	12*, 9*, 1*, 2*, 3*, 4*, 11	12*, 9*, 3*, 6, 1*, 2*	7, 10, 4*	2

Table 3: ROCr-selected pipelines, Overall Reliability.
The ROCr-selected pipelines for day 1, 2 and 3 are shown for each subject. '*' indicates that the pipeline was selected on 2/3 days. '' indicates that the pipeline was selected on 3/3 days. Hypothesis A was met for subjects who obtained a score of 1: rt001, rt002 and rt006.**

3.4.1.2 Reliability vs. Time

Figure 24 shows the results for pipeline performance using the Reliability vs. Time approach for subject rt006. On day 1, Pipeline 7 was the best. On day 2 and 3, Pipeline 10 was the best. This subject did not meet Hypothesis A because ROCr did not select the same pipeline on each day. The average reliabilities for the Reliability vs. Time approach appear to be higher compared to the Overall Reliability approach, and the second session seems to have better rankings than with the Overall reliability approach.

Day 3 had higher reliability across all pipelines. The data session had a significant effect on the mean reliable fraction across pipelines. The average reliable fraction for Day 3 was 0.021 greater than for Day 2, and 0.026 greater than for Day 1. The data session also had a significant effect on the mean reliable fraction at the group level. The average reliable fraction for Day 3 was 0.0075 greater than for Day 1, and 0.0078 greater than for Day 2.

These trends observed at the subject and group level for the Reliability vs. Time approach are similar to the findings for the Overall Reliability approach, with two main differences. First, at the subject level, rt006 Day 2 had much better results using the Reliability vs. Time approach (mean = 0.8248750, min = 0.7370, max = 0.8830) than the Overall Reliability group (mean = 0.7413255, min = 0.61183, max = 0.82168). Second, at the group level, the across-session differences were much smaller. This might suggest that the Reliability vs. Time approach is more resilient to noise artefacts than the Overall Reliability approach, as long as the artefacts are not during a time of expected activity.

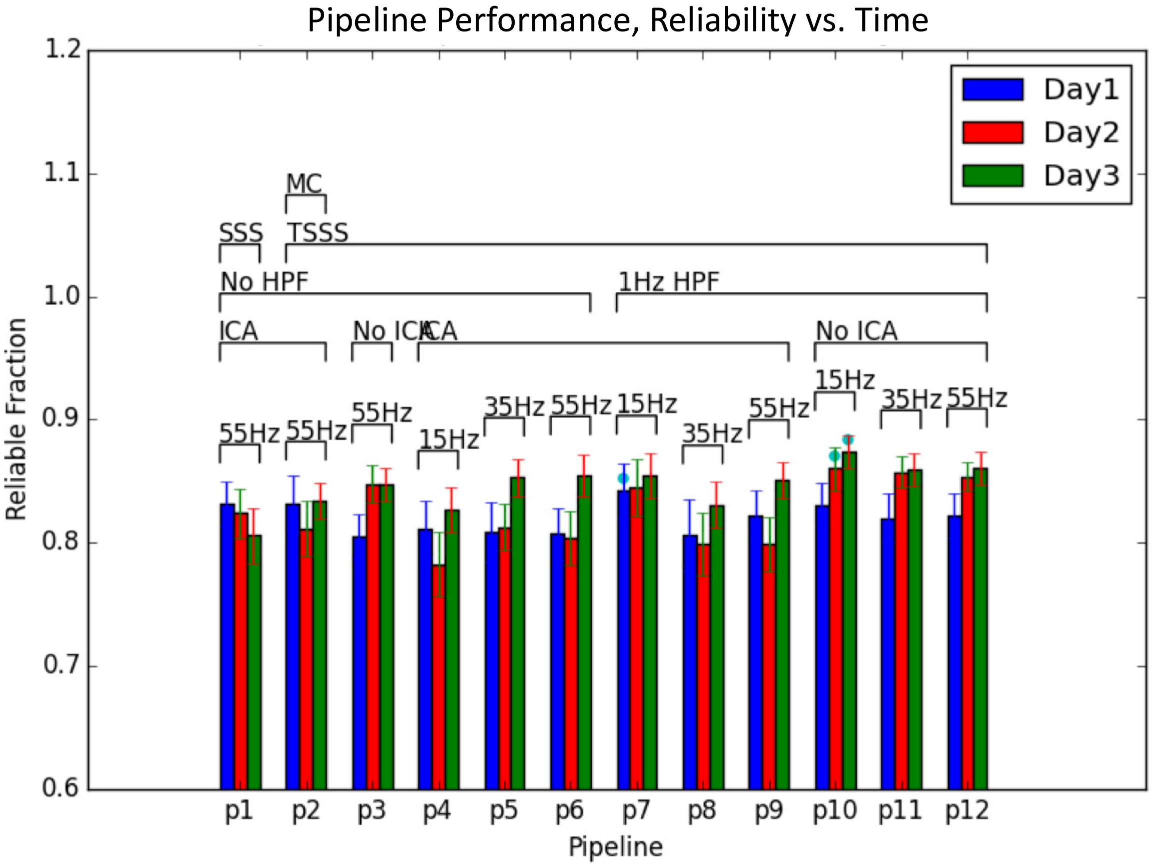


Figure 24: Pipeline Performance, Single Subject (rt006), Reliability Vs Time.
The x-axis shows the pipeline used, and the y-axis shows the average reliable fraction. Blue, red and green bars indicate average reliability on day 1, 2 and 3 respectively. Blue dots indicate the pipeline with the highest average reliable fraction. Error bars indicate reliable fraction standard deviation. Day 2 and Day 1 results have significantly lower average reliability than Day 3.

Table 4: summarizes the ROCr-selected pipelines on each day, for each subject. Pipeline 9 was selected the most frequently (15/30), followed by Pipeline 12 (14/30) and 11 (13/30). Pipeline 4 was selected the least (2/30).

Scores were assigned to each subject as discussed in section 3.4.1.1. Scores range from 4 (worst) to 1 (best). Subjects with a score of 1 meet Hypothesis A. Three subjects had a score of 1: rt003, rt005 and rt015. All other subjects had a score of 2 or 3. Both methods had three (different) pipelines that met Hypothesis A, but the Reliability vs. Time approach had better scores overall: the scores were summed, and the Overall Reliability method had a combined score of 23 whereas the Reliability vs. Time method had a combined score of 19. This could suggest that the Reliability vs. Time method is less susceptible to artefacts in the data.

Subject	Day 1	Day 2	Day 3	Score
rt001	9*, 7*	7*, 2, 8*	9*, 8*	2
rt002	11*, 7, 12, 9*, 10	8*, 1, 9*, 11*	8*	2
rt003	9**, 8*, 1	6, 12*, 8*, 11*, 3, 9**, 2*, 5	12*, 9**, 10, 11*, 7, 4, 2*, 6	1
rt004	12*,9, 11*	12*, 11*, 7, 2, 6, 8	1	2
rt005	1, 12**	12**	10, 7, 12**, 9, 4, 3, 6, 11	1
rt006	7	10*, 11, 12	10*	3
rt010	1*	1*	10, 7, 11	3
rt011	12*, 9*	1	9*, 8, 11, 12*	2
rt012	7	5*, 12, 8*, 6*, 11	5*, 9, 2, 8*, 6*	2
rt015	11*, 12*, 1, 10, 9**, 3*, 7*	6, 9**, 12*, 3*, 2*, 5, 11*	9**, 2*, 6, 8, 7*	1

Table 4: ROCr-selected pipelines, Reliability vs. Time.
The ROCr-selected pipelines for day 1, 2 and 3 are shown for each subject. '*' indicates that the pipeline was selected on 2/3 days. '' indicates that the pipeline was selected on 3/3 days. Hypothesis A was met for subjects who obtained a score of 1: rt003, rt005 and rt015.**

3.4.3 Pipeline Rankings

Pipeline rankings can give us additional insight to reliability findings for each pipeline. A limitation of this approach is that it does not show the difference in reliable fraction between each rank. These differences vary, and as indicated by the ANOVAs done in section 3.4.1, many of the differences in reliable fraction across pipelines are not statistically significant.

3.4.3.1 Overall Reliability

Table 5 shows the rank of each pipeline on day 1, day 2 and day 3, as well as the pipeline's average rank and standard deviation across days. A ranking of 1 indicates the pipeline with the highest reliable fraction, and a ranking of 12 indicates the pipeline with the lowest reliable fraction. Pipeline 10 was, on average, the top-ranking pipeline for this subject. Pipeline 8 was the lowest-ranking pipeline.

At the group level, the best- and worst-ranked pipelines differ. Figure 25 shows the average pipeline rankings across all subjects and sessions. Pipeline 12 had the lowest (best) average ranking, and Pipeline 4 had the worst ranking. Standard deviations were quite large for each pipeline, but trends were observed in the effects of different pipeline parameters on pipeline rankings, such as: a) as LPF increases, ranking improves, b) a HPF of 1 Hz gives better rankings than a DC HPF, and c) not using ICA gives better rankings than using ICA.

Pipeline	Session 1	Session 2	Session 3	**Average**
SSS	2.68*	5.46	5.82	**4.65**
MC	5.70	8.98	7.54	**7.41**
P3	9.90**	4.24	8.16	**7.43**
P4	8.38	11.64**	8.72	**9.58**
P5	8.84	7.88	6.14	**7.62**
Standard	9.92	9.56	7.66	**9.05**
P7	2.96	5.18	7.72	**5.29**
P8	8.52	10.60	10.00**	**9.71****
P9	5.68	8.30	4.26	**6.08**
P10	3.88	1.38*	3.50*	**2.92***
P11	5.28	2.38	3.76	**3.81**
P12	6.26	2.40	4.72	**4.46**

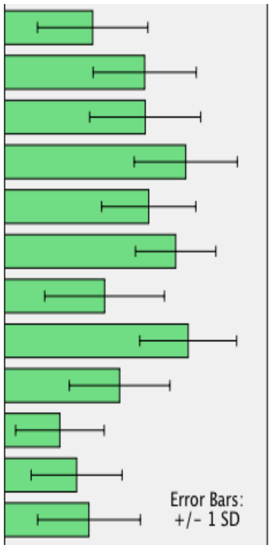


Table 5: Mean Pipeline Rankings Across Sessions (Overall Reliability Method, Subject rt006).
'*' indicates the best ranked pipeline in the session. '' indicates the worst ranked pipeline in the session. Green bars are a visual representation of the across-sessions average rankings. On average, Pipeline 8 was the worst for subject rt006, and Pipeline 10 was the best.**

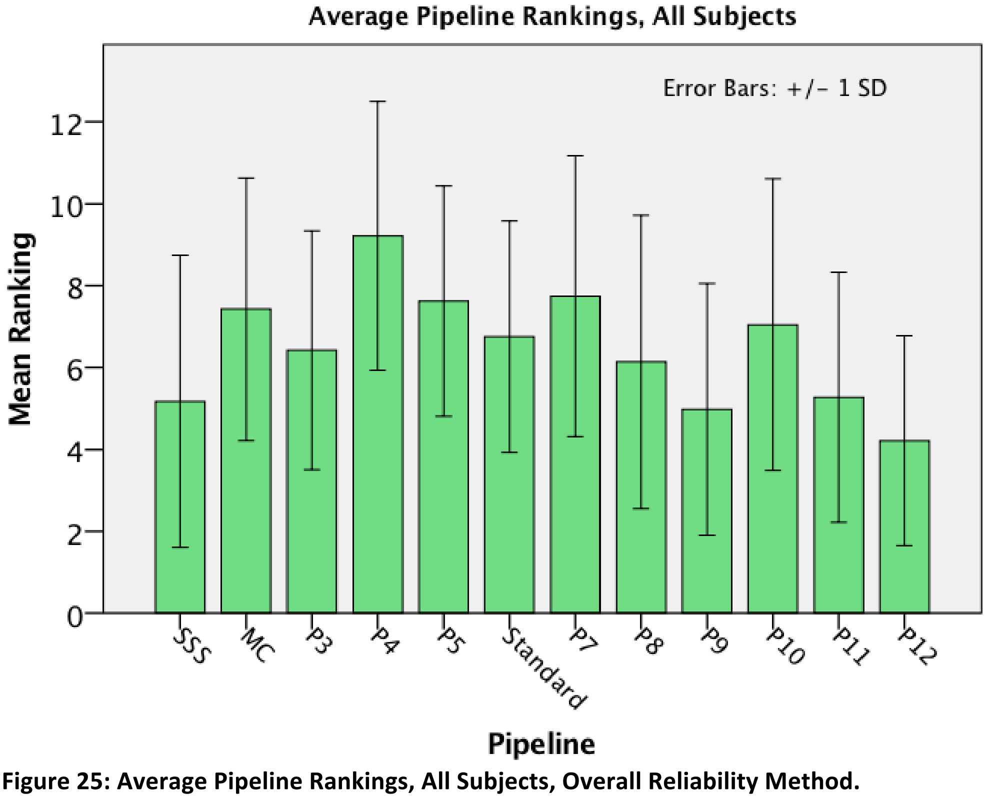
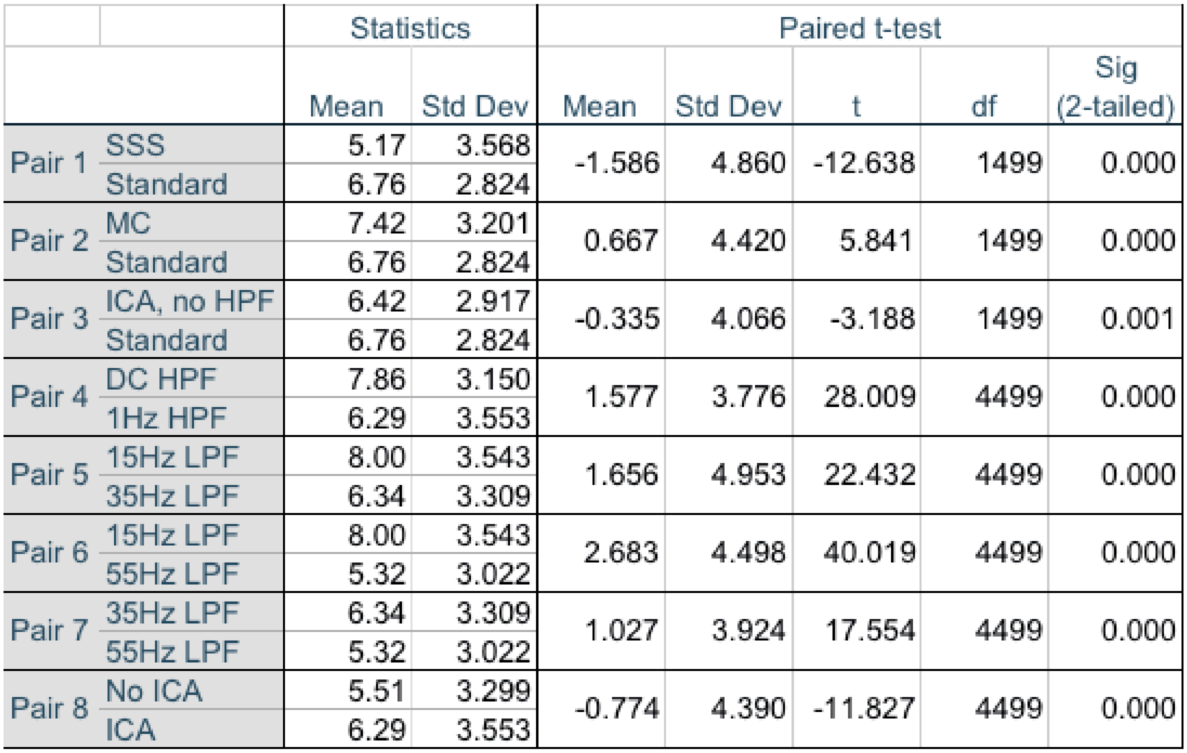


Figure 25: Average Pipeline Rankings, All Subjects, Overall Reliability Method.
The x-axis indicates the pipeline and the y-axis indicates the mean ranking. Error bars indicate standard deviation. Several trends can be observed from this plot. As LPF increases, ranking improves. A HPF of 1 Hz gives better rankings than a DC HPF. Not using ICA gives better rankings than using ICA.

Paired t-tests were done to quantify the observed trends. The first test was done to determine the effect of using SSS instead of tSSS. Pipelines 1 and 6 have identical parameters except for environmental noise reduction. The second test was done to determine the effects of using motion correction. Pipelines 2 and 6 have identical parameters except for motion correction. The third test was done to determine the effects of using ICA in combination with a DC HPF. Pipelines 3 and 6 have identical parameters except for ICA and HPF. The fourth test was done to determine the effects of using DC or 1 Hz HPF. Pipeline groups 4-6 and 7-9 have identical parameters except for HPF. The fifth, sixth and seventh tests were done to determine the effects of the LPF frequency. Pipeline groups 4, 7, 10; 5, 8, 11; and 6, 9, 12 have LPF frequencies of 15 Hz, 35 Hz and 55 Hz respectively. All other parameters are balanced in the comparison. Finally, the eighth test was done to determine the effects of using ICA.

Table 6 shows the results of these paired t-tests. Each comparison was significant. Using SSS yielded better rankings. This matches the high occurrence of Pipeline 1 among the ROCr-selected pipelines. Motion correction resulted in worse rankings. Using a HPF of 1 Hz gave better rankings than using the default HPF. For each LPF frequency pair tested, the higher LPF frequency gave better rankings. Using ICA gave worse rankings than not using ICA. Pipeline 12, which had the best average ranking, used most of these parameters: it did not use SSS, but it had a 1 Hz HPF, a 55 Hz LPF and did not use ICA.

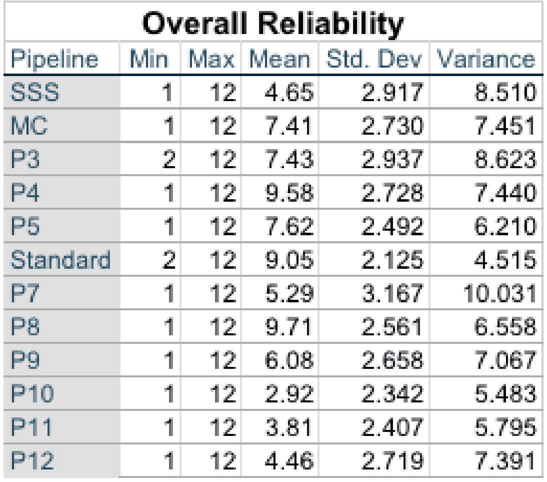


		Statistics		Paired t-test				
		Mean	Std Dev	Mean	Std Dev	t	df	Sig (2-tailed)
Pair 1	SSS	5.17	3.568	-1.586	4.860	-12.638	1499	0.000
	Standard	6.76	2.824					
Pair 2	MC	7.42	3.201	0.667	4.420	5.841	1499	0.000
	Standard	6.76	2.824					
Pair 3	ICA, no HPF	6.42	2.917	-0.335	4.066	-3.188	1499	0.001
	Standard	6.76	2.824					
Pair 4	DC HPF	7.86	3.150	1.577	3.776	28.009	4499	0.000
	1Hz HPF	6.29	3.553					
Pair 5	15Hz LPF	8.00	3.543	1.656	4.953	22.432	4499	0.000
	35Hz LPF	6.34	3.309					
Pair 6	15Hz LPF	8.00	3.543	2.683	4.498	40.019	4499	0.000
	55Hz LPF	5.32	3.022					
Pair 7	35Hz LPF	6.34	3.309	1.027	3.924	17.554	4499	0.000
	55Hz LPF	5.32	3.022					
Pair 8	No ICA	5.51	3.299	-0.774	4.390	-11.827	4499	0.000
	ICA	6.29	3.553					

Table 6: Paired t-test results for pipeline parameters (all subjects, Overall Reliability method). Basic statistics (mean and standard deviation) are shown for rankings of each group being compared. Paired t-test results (mean, standard deviation, t statistic, degrees of freedom, 2-tailed significance) are shown for each pair. A significant difference was found between each pair. Pair 1 tests the effect of using SSS versus tSSS. Pair 2 tests the effects of using motion correction. Pair 3 tests the effect of using ICA and no HPF. Pair 4 tests the effect of using HPF. Pairs 5, 6 and 7 test the effects of different LPF frequencies. Pair 8 tests the effects of using ICA.

The variance of pipeline rankings within subjects was also investigated. Figure 24 a) shows some descriptive statistics for subject rt006 rankings data. Figure 26 b) shows the ranking variance for each pipeline for subject rt006. The standard pipeline had the lowest variance (4.515), and Pipeline 7 had the highest variance (10.031). Low variance indicates that the pipeline's ranking, good or bad, was relatively consistent across days. If ROCr is working as expected, ranking variance should be low, as this would indicate the rankings are very similar across sessions and splits. Figure 27 shows the pipeline ranking variance at the group level. Pipeline 1 had the highest average variance, and Pipeline 12 had the lowest average variance.

a)



Overall Reliability					
Pipeline	Min	Max	Mean	Std. Dev	Variance
SSS	1	12	4.65	2.917	8.510
MC	1	12	7.41	2.730	7.451
P3	2	12	7.43	2.937	8.623
P4	1	12	9.58	2.728	7.440
P5	1	12	7.62	2.492	6.210
Standard	2	12	9.05	2.125	4.515
P7	1	12	5.29	3.167	10.031
P8	1	12	9.71	2.561	6.558
P9	1	12	6.08	2.658	7.067
P10	1	12	2.92	2.342	5.483
P11	1	12	3.81	2.407	5.795
P12	1	12	4.46	2.719	7.391

b)

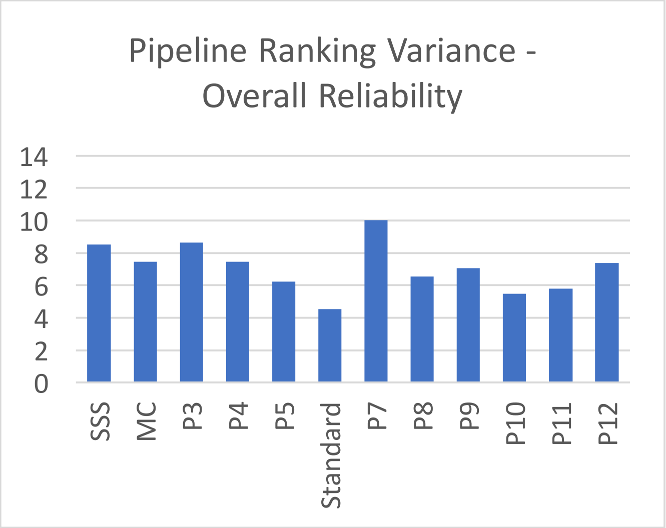


Figure 26: Pipeline ranking variance, single subject, Overall Reliability method.
(a) Table of statistical measures for each pipeline's rankings: minimum value, maximum value, mean ranking, standard deviation, variance. (b) Graph of pipeline ranking variance. X-axis indicates the pipeline. Y-axis indicates the variance in pipeline ranking. The standard pipeline had the least variance (its ranking did not change much across splits). Pipeline 7 had the highest variance (its ranking changed the most across splits).

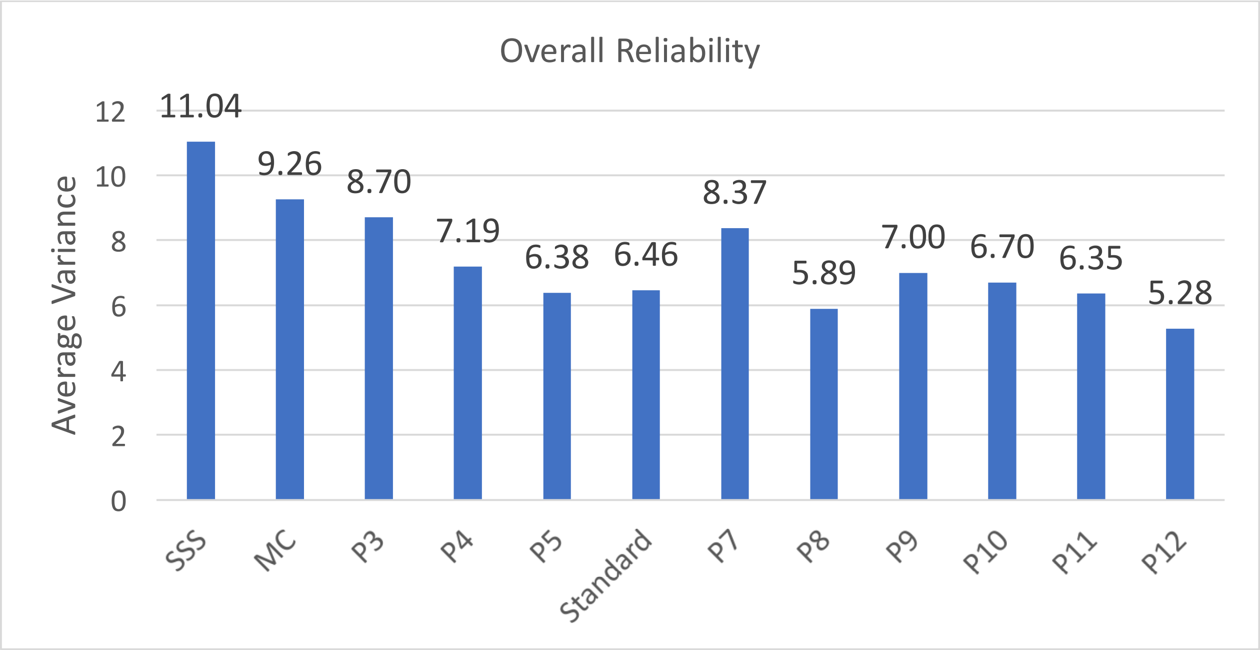


Figure 27: Average variances across subjects, Overall Reliability approach.
X-axis indicates pipeline. Y-axis indicates sum of variance. The SSS pipeline and the MC pipeline had the highest average variances. Pipelines 8 and 12 had the lowest average variances.

3.4.3.2 Reliability vs. Time

Table 7 shows the rank of each pipeline on day 1, day 2 and day 3, as well as the pipeline's average rank and standard deviation across days, using the Reliability vs. Time approach for subject rt006. A ranking of 1 indicates the pipeline with the highest reliable fraction, and a ranking of 12 indicates the pipeline with the lowest reliable fraction. Pipeline 10 was, on average, the top-ranking pipeline for this subject. Pipeline 4 was the lowest-ranking pipeline. In contrast to the results obtained using the Overall Reliability approach, Pipeline 1 had much worse rankings, and was the worst pipeline on one of the days.

At the group level, the worst-ranked pipeline is the same (Pipeline 4). The best-ranked pipeline is different. Figure 28 shows the average pipeline rankings across all subjects and sessions. Pipeline 12 had the lowest (best) average ranking, and Pipeline 4 had the worst ranking. Standard deviations were once again quite large, but trends were observed in the effects of different pipeline parameters on pipeline rankings. These trends are the same as for the Overall Reliability approach: a) as LPF increases, ranking improves, b) a HPF of 1 Hz gives better rankings than a DC HPF, and c) not using ICA gives better rankings than using ICA.

Pipeline	Session 1	Session 2	Session 3	Average
SSS	4.42	6.66	11.54**	**7.54**
MC	5.30	8.18	9.28	**7.59**
P3	9.06**	3.82	6.62	**6.50**
P4	7.68	10.72**	9.50	**9.30****
P5	8.18	8.26	5.38	**7.27**
Standard	8.46	9.38	5.20	**7.68**
P7	3.04*	4.02	5.28	**4.11**
P8	8.38	9.64	9.08	**9.03**
P9	6.56	9.46	6.00	**7.34**
P10	4.54	2.24*	1.78*	**2.85***
P11	6.22	2.48	4.12	**4.27**
P12	6.16	3.14	4.22	**4.51**

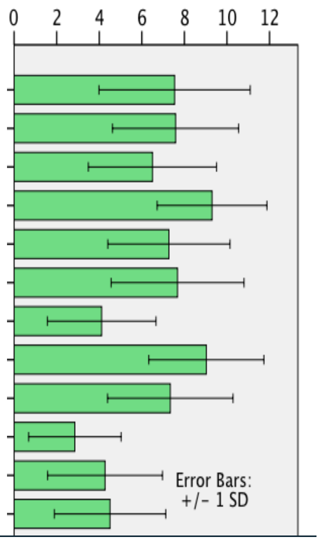


Table 7: Mean Pipeline Rankings Across Sessions (Subject rt006, Reliability vs. Time method).
'*' indicates the best ranked pipeline in the session. '' indicates the worst ranked pipeline in the session. Green bars are a visual representation of the across-sessions average rankings.**

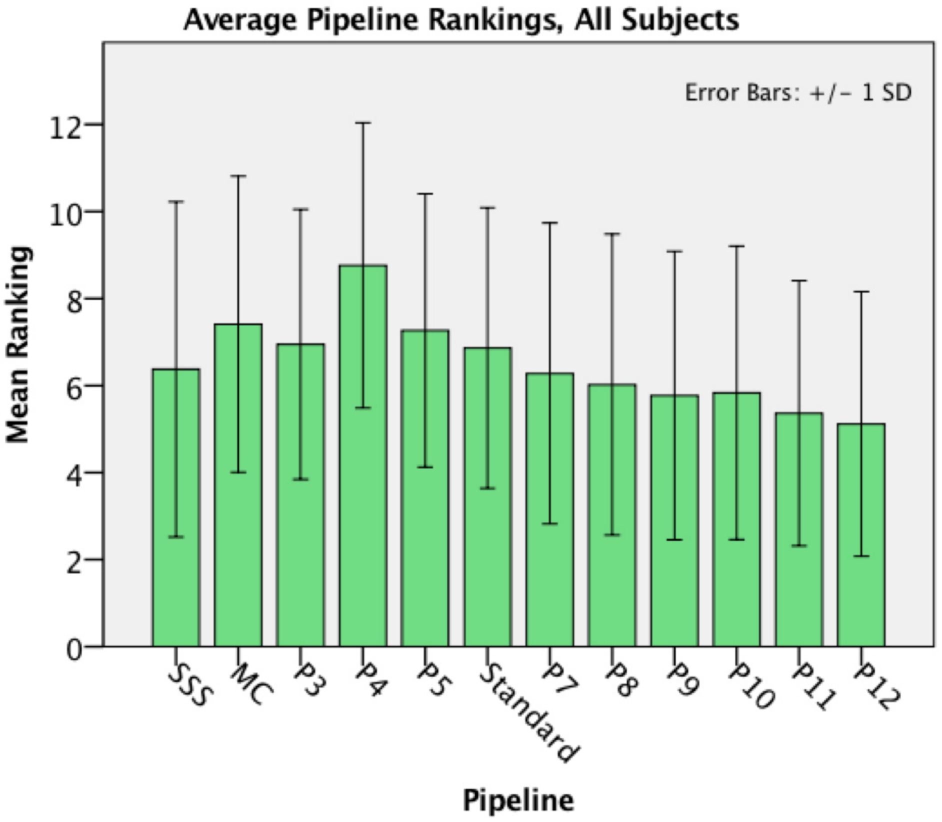
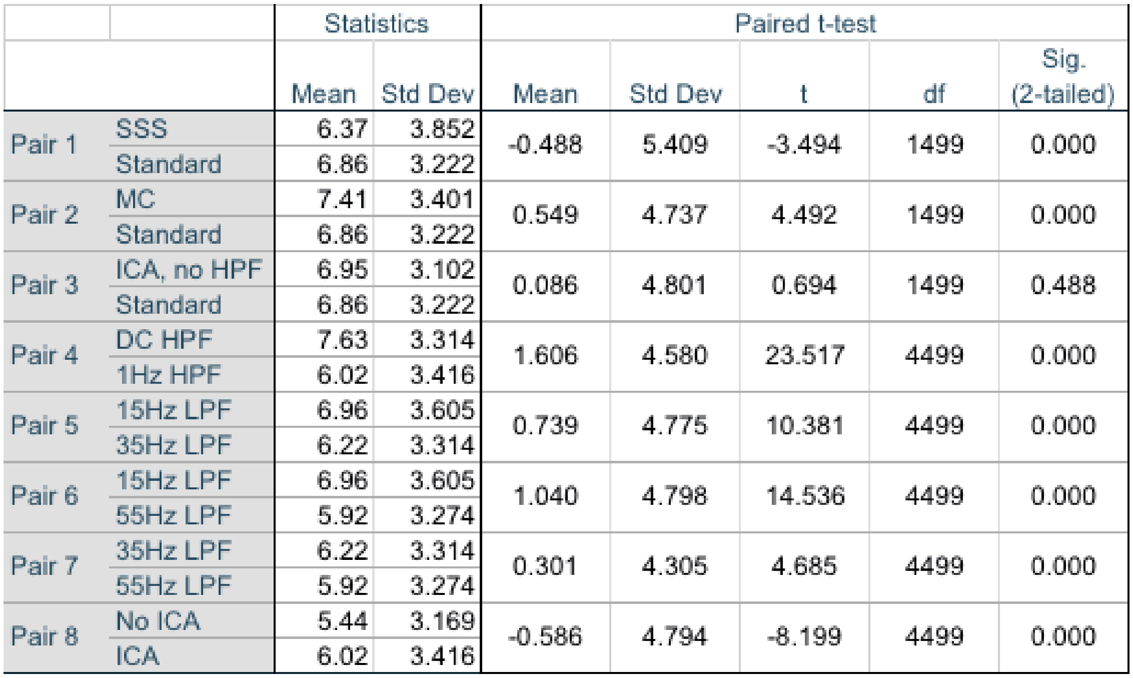


Figure 28: Average Pipeline Rankings, All Subjects, Reliability vs. Time method.
The x-axis indicates the pipeline and the y-axis indicates the mean ranking. Error bars indicate standard deviation. Several trends can be observed from this plot. As LPF increases, ranking improves. A HPF of 1 Hz gives better rankings than a DC HPF. Not using ICA gives better rankings than using ICA.

Paired t-tests were done to quantify the observed trends. The tests were the same as for the Overall Reliability approach. The first test was done to determine the effect of using SSS instead of tSSS. Pipelines 1 and 6 have identical parameters except for environmental noise reduction. The second test was done to determine the effects of using motion correction. Pipelines 2 and 6 have identical parameters except for motion correction. The third test was done to determine the effects of using ICA in combination with a DC HPF. Pipelines 3 and 6 have identical parameters except for ICA and HPF. The fourth test was done to determine the effects of using DC or 1 Hz HPF. Pipeline groups 4-6 and 7-9 have identical parameters except for HPF. The fifth, sixth and seventh tests were done to determine the effects of the LPF frequency. Pipeline groups 4, 7, 10; 5, 8, 11; and 6, 9, 12 have LPF frequencies of 15 Hz, 35 Hz and 55 Hz respectively. All other parameters are balanced in the comparison. Finally, the eighth test was done to determine the effects of using ICA.

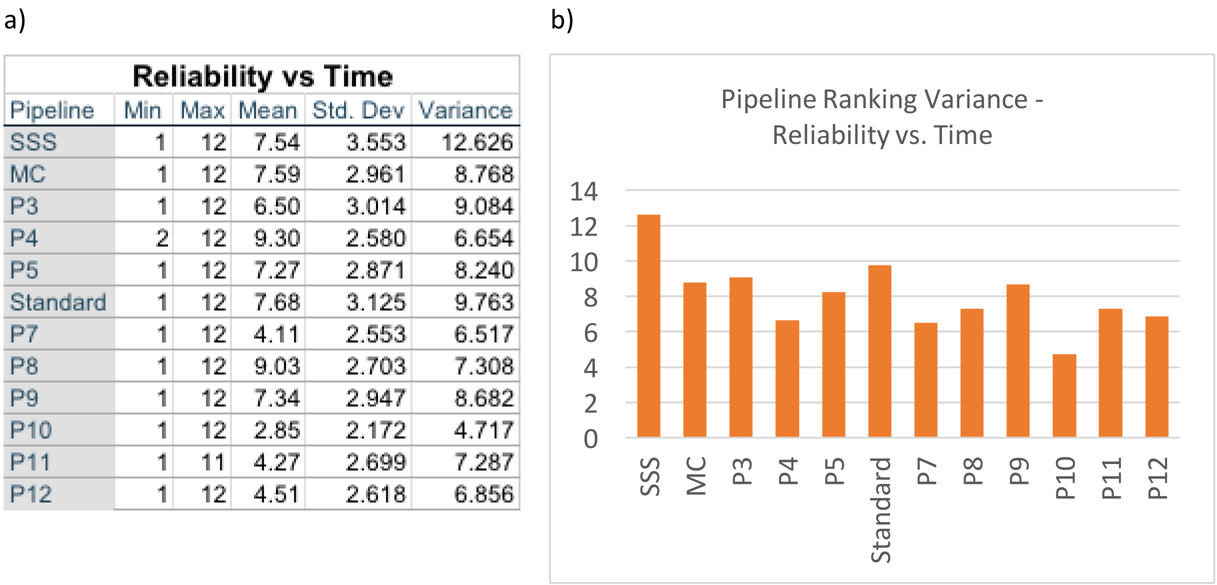
Table 8 shows the results of these paired t-tests. Each comparison was significant except for the third test. Using SSS yielded better rankings. This matches the high occurrence of Pipeline 1 among the ROCr-selected pipelines. Motion correction resulted in worse rankings, and using a HPF of 1 Hz gave better rankings than not applying additional high-pass filtering. For each LPF frequency pair tested, the higher LPF frequency gave better rankings. Using ICA gave worse rankings than not using ICA. Pipeline 12, which had the best average ranking, used most of these parameters: it did not use SSS, but it had a 1 Hz HPF, a 55 Hz LPF and did not use ICA.



		Statistics		Paired t-test				
		Mean	Std Dev	Mean	Std Dev	t	df	Sig. (2-tailed)
Pair 1	SSS	6.37	3.852	-0.488	5.409	-3.494	1499	0.000
	Standard	6.86	3.222					
Pair 2	MC	7.41	3.401	0.549	4.737	4.492	1499	0.000
	Standard	6.86	3.222					
Pair 3	ICA, no HPF	6.95	3.102	0.086	4.801	0.694	1499	0.488
	Standard	6.86	3.222					
Pair 4	DC HPF	7.63	3.314	1.606	4.580	23.517	4499	0.000
	1Hz HPF	6.02	3.416					
Pair 5	15Hz LPF	6.96	3.605	0.739	4.775	10.381	4499	0.000
	35Hz LPF	6.22	3.314					
Pair 6	15Hz LPF	6.96	3.605	1.040	4.798	14.536	4499	0.000
	55Hz LPF	5.92	3.274					
Pair 7	35Hz LPF	6.22	3.314	0.301	4.305	4.685	4499	0.000
	55Hz LPF	5.92	3.274					
Pair 8	No ICA	5.44	3.169	-0.586	4.794	-8.199	4499	0.000
	ICA	6.02	3.416					

Table 8: Paired t-test results for pipeline parameters (all subjects, Reliability vs. Time method). Basic statistics (mean and standard deviation) are shown for rankings of each group being compared. Paired t-test results (mean, standard deviation, t statistic, degrees of freedom, 2-tailed significance) are shown for each pair. A significant difference was found between each pair. Pair 1 tests the effect of using SSS versus tSSS. Pair 2 tests the effects of using motion correction. Pair 3 tests the effect of using ICA and no HPF. Pair 4 tests the effect of using HPF. Pairs 5, 6 and 7 test the effects of different LPF frequencies. Pair 8 tests the effects of using ICA.

The variance of pipeline rankings within subjects was also investigated. Figure 29 a) shows the descriptive statistics for subject rt006 rankings data. Figure 29 b) shows the ranking variance for each pipeline for subject rt006. Pipeline 10 had the lowest variance (4.717), and Pipeline 1 had the highest variance (12.626). Figure 30 shows the pipeline ranking variance at the group level. As with the Overall Reliability approach, Pipeline 1 had the highest average variance, and Pipeline 12 had the lowest average variance.



a)

Reliability vs Time					
Pipeline	Min	Max	Mean	Std. Dev	Variance
SSS	1	12	7.54	3.553	12.626
MC	1	12	7.59	2.961	8.768
P3	1	12	6.50	3.014	9.084
P4	2	12	9.30	2.580	6.654
P5	1	12	7.27	2.871	8.240
Standard	1	12	7.68	3.125	9.763
P7	1	12	4.11	2.553	6.517
P8	1	12	9.03	2.703	7.308
P9	1	12	7.34	2.947	8.682
P10	1	12	2.85	2.172	4.717
P11	1	11	4.27	2.699	7.287
P12	1	12	4.51	2.618	6.856

Figure 29: Pipeline ranking variance, single subject, Reliability vs. Time method.
(a) Table of statistical measures for each pipeline's rankings: minimum value, maximum value, mean ranking, standard deviation, variance. (b) Graph of pipeline ranking variance. X-axis indicates the pipeline. Y-axis indicates the variance in pipeline ranking. Pipeline 10 had the least variance (its ranking did not change much across splits). Pipeline 1 had the highest variance (its ranking changed the most across splits).

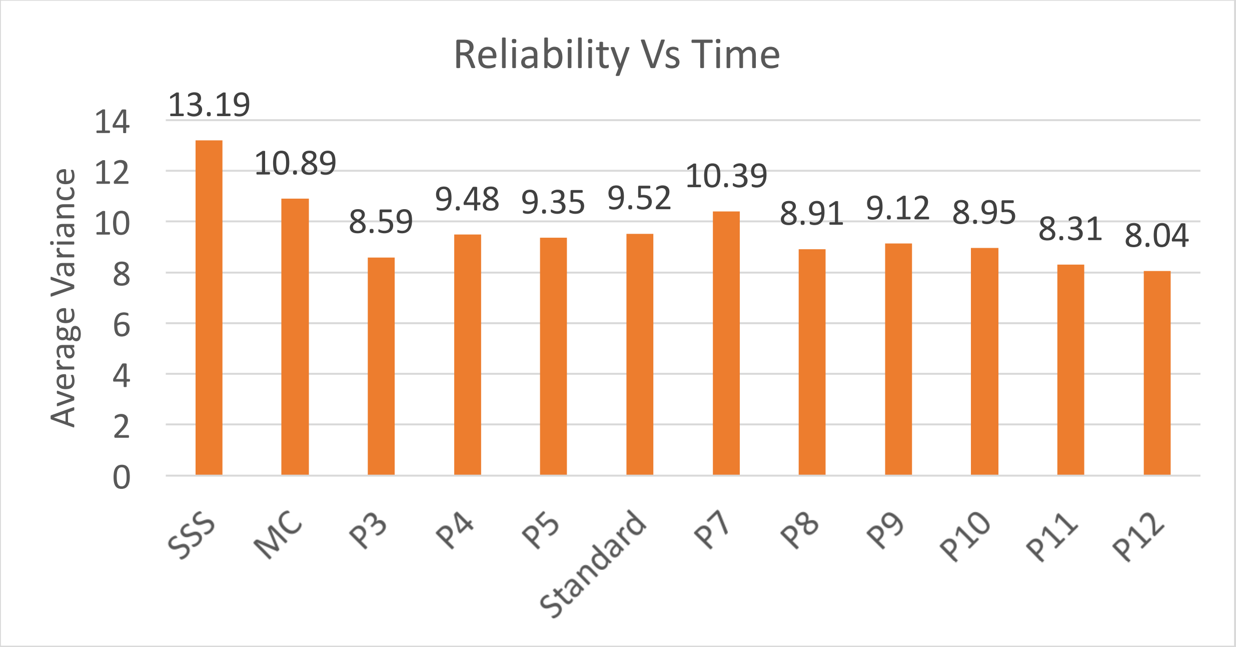


Figure 30: Average variances across subjects, Reliability vs. Time approach.
X-axis indicates pipeline. Y-axis indicates average variance. The SSS pipeline and the MC pipeline had the highest average variances, and Pipeline 12 had the lowest average variance. The range of variances is smaller for the Reliability vs. Time approach.

Comparison

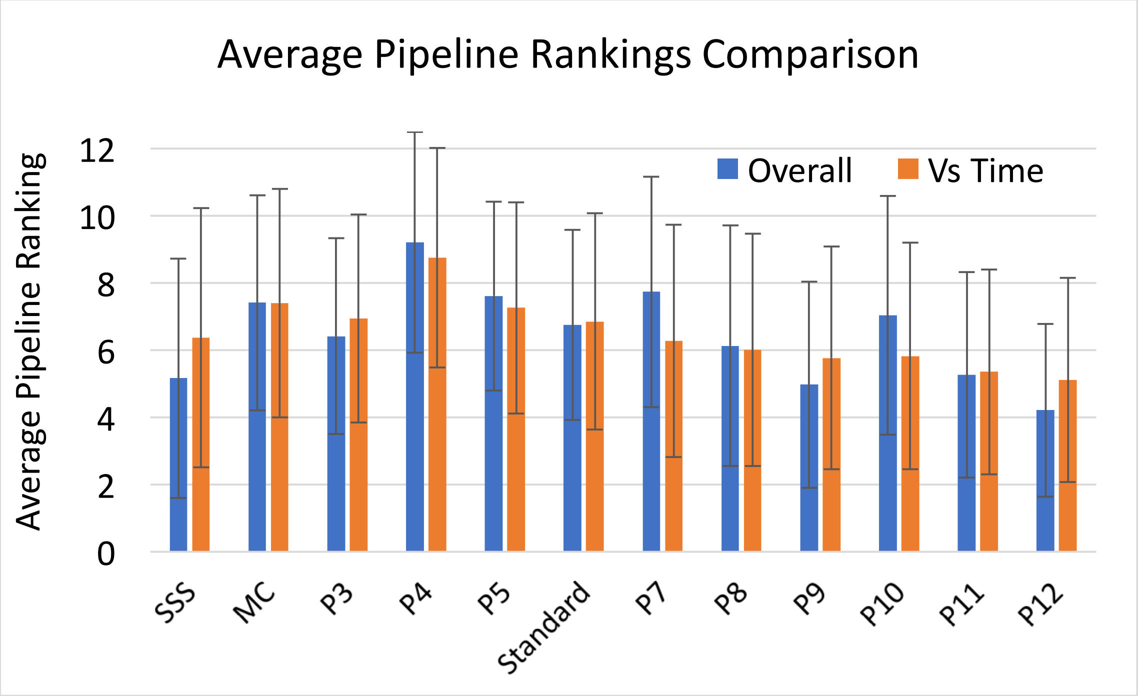


Figure 31 compares the average pipeline rankings obtained with the Overall Reliability method and the Reliability vs. Time method. Pipelines 1 and 9 performed moderately well with the Overall Reliability approach, but their performances worsened when the Reliability vs. Time approach was used. This indicates that these pipelines produced activation maps with fairly reliable features (noise, artefacts, features of interest) overall, but there was no time where reliability was exceptionally high.

Pipelines 7 and 10 demonstrate the opposite effect: their performances improved when the Reliability vs. Time approach was used. This indicates that the activation maps had low Overall Reliability, but for at least one time point, there was a very reliable feature. Ideally this feature should be one of the SEF peaks, but it may also be an artefact such as the MNS pulse or a motion artefact.

Figure 32 shows the average ranking variance for each pipeline, across all subjects. Figure 33 shows the total ranking variance for each subject, across all pipelines. For each figure, the variances appear to be larger for the Reliability vs. Time approach. This makes sense because the Reliability vs. Time approach is more sensitive to local fluctuations in reliability. Figure 22 is a histogram indicating the time in each split where reliability was the highest. The majority of these points are in the 35-125 ms range, with peaks around 35 ms, 58 ms and 120 ms. There were a few occasions where a very late time had a high reliability, and a few occasions where the MNS artifact was the most reliable point in the data. Because the Overall Reliability approach looks at the whole dataset, not just the peak reliability, these differences in timing and reliability do not have as large of an effect.

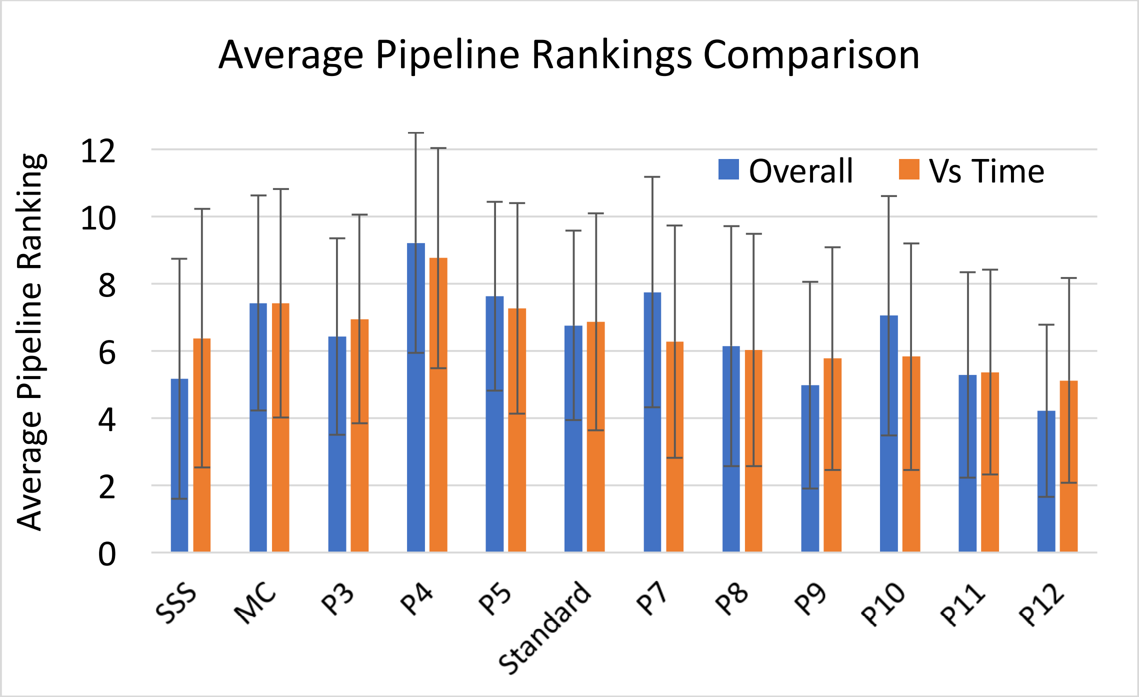


Figure 31: Average pipeline rankings comparison, all subjects.

X-axis indicates pipeline. Y-axis indicates sum of variance. Blue bars indicate results for the Overall Reliability method. Orange bars indicate results for the Reliability vs. Time method. Certain pipelines had better rankings with different methods. For example, the SSS pipeline had a better average ranking with the Overall Reliability method, and Pipeline 10 had a better average ranking with the Reliability vs. Time method.

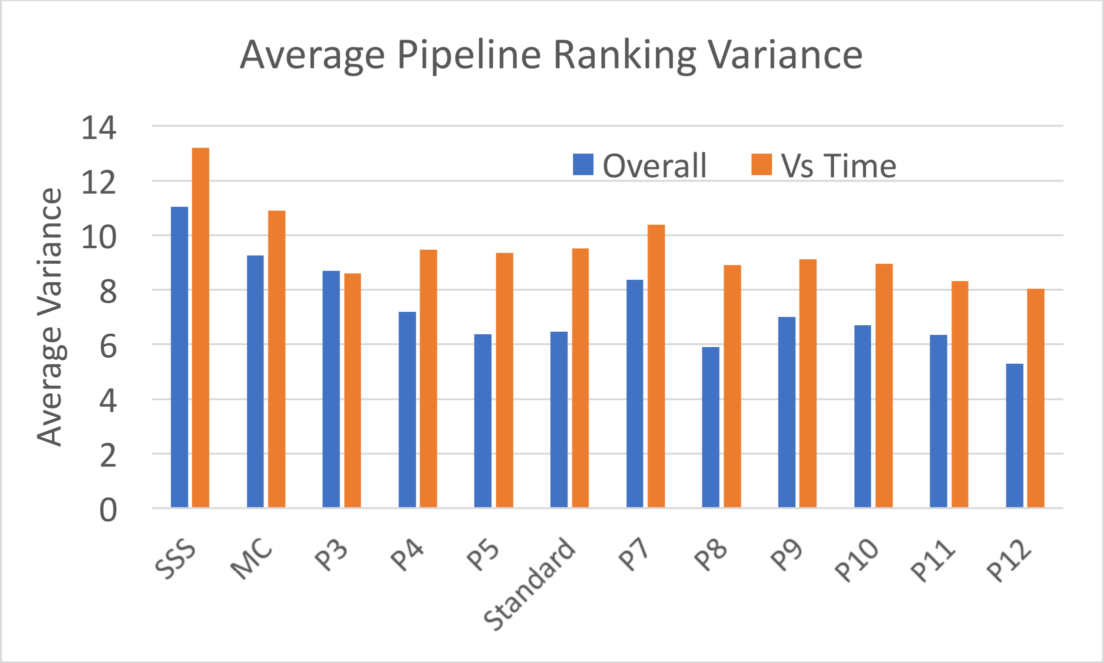


Figure 32: Average pipeline ranking variance across all subjects and sessions.
X-axis indicates pipeline. Y-axis indicates average ranking variance. Blue bars indicate results for the Overall Reliability method. Orange bars indicate results for the Reliability vs. Time method. On average, the Overall Reliability method resulted in lower average variance in pipeline rankings.

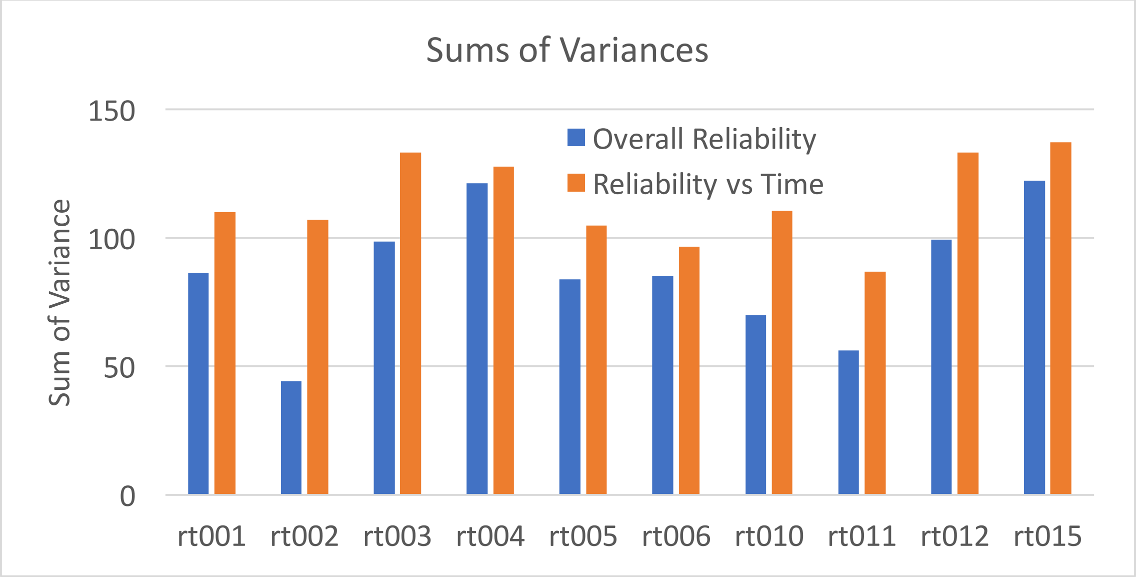


Figure 33: Total variance for each subject, across all pipelines.
X-axis indicates subject. Y-axis indicates the sum of variance. Blue bars indicate results for the Overall Reliability method. Orange bars indicate results for the Reliability vs. Time method. On average, the Overall Reliability method resulted in lower total variance across pipelines. Subjects rt002 and rt011 had the lowest total variance.

3.4.4 Effects of Spatial and Temporal Restriction

The results of spatial and temporal restriction are shown below.

A repeated measures ANOVA was done to determine the effect of the reliability method used on the reliable fractions across subjects and sessions. Only the standard pipeline was used in this test. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(5) = 1506.457$, $p < 0.001$, therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.488$). The results show that there was a significant effect of the method used on the reliable fraction across sessions, $F(1.464, 876.723) = 11452.436$, $p < 0.001$.

A pairwise comparison indicated that each method gave a significantly different average reliability. The Reliability vs. Time method gave the highest average reliability, and the windowed (temporally restricted) method gave the lowest average reliability. This is to be expected, as the windowed method takes the average reliable fraction across a time window where reliability was generally lower. The windowed method would be better evaluated in terms of ranking rather than a direct comparison to the other methods. The one hemisphere (spatially restricted) method had a significantly higher average reliability than the Overall Reliability method, which indicates that restricting the data to an area or hemisphere of interest may improve reliability.

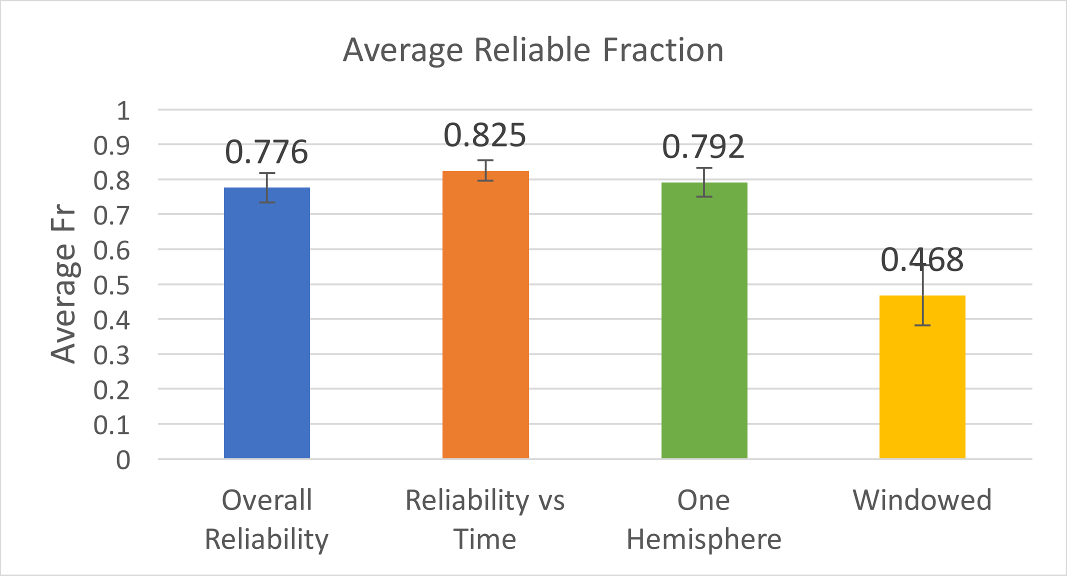


Figure 34: Comparison of reliability methods, all subjects and sessions, standard pipeline. X-axis indicates the method used. Y-axis indicates the average reliable fraction. Each method had a significantly different average. The Reliability vs. Time method had the highest average reliable fraction, followed by the one hemisphere (spatially restricted) method, the Overall Reliability method, and finally the windowed (temporally restricted) method.

3.6 Study 2: Relationships between Reliability and Inter-Session Variability

For this study, a typical subject was selected to display the results. Seven out of ten subjects had lower inter-session variability for the ROCr-selected pipeline. Of these seven subjects, two (rt001 and rt003) had inter-session variability measures similar to results published by Solomon et al. (11). All single-subject results discussed below are for subject rt001.

3.6.1 ROCr vs. Standard Pipeline

Figure 35 shows the Euclidean distances, or "focality" of the S1 localizations obtained using the ROCr-selected pipeline and the standard pipeline on each day and averaged across days. This figure shows results for subject rt001. On each day, the ROCr-selected pipeline had a statistically significantly smaller ED than the standard pipeline. The average ED for the ROCr-selected pipeline was also significantly smaller than the average ED for the standard pipeline. Briefly, Hypothesis B states that the average ED for the ROCr-selected pipelines should be lower than for the standard pipeline, so Hypothesis B is met for subject rt001.

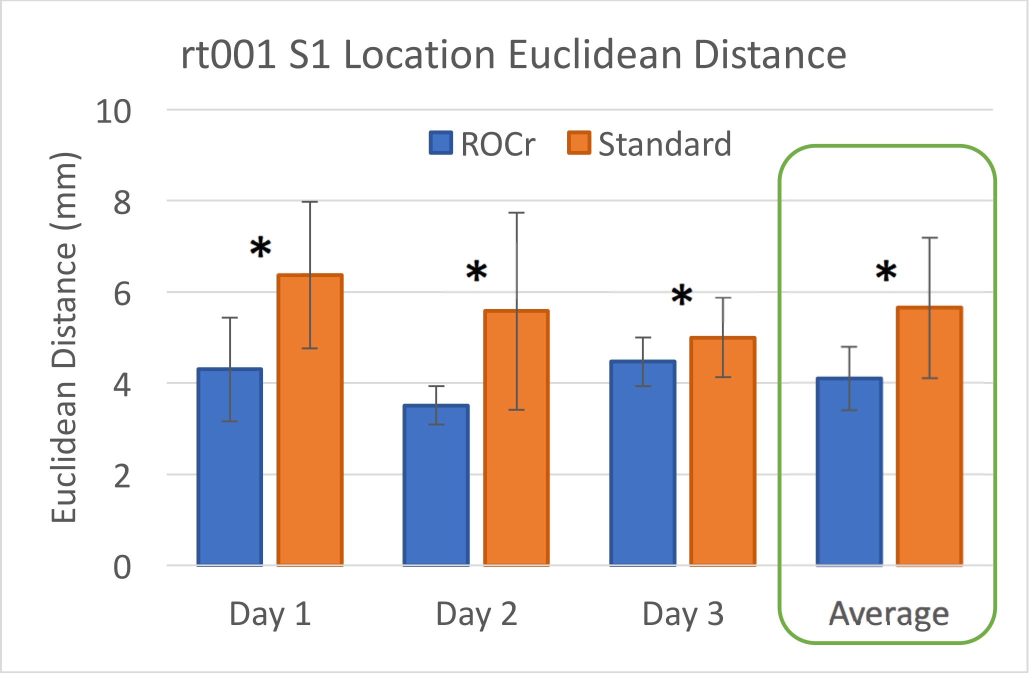


Figure 35: Euclidean Distances from ROCr-selected and Standard pipelines, subject rt001. Blue bars indicate the Euclidean distance for the ROCr-selected pipeline on the indicated day, and orange indicate the standard pipeline (Pipeline 6). Error bars show standard deviation. Asterisks indicate that there was a statistically significant difference between the ROCr-selected pipeline ED and the standard pipeline ED. The difference between average EDs (outlined in green) was used to test Hypothesis B. For this subject, Hypothesis B was met because the ROCr-selected pipelines gave a lower average ED than the standard pipeline.

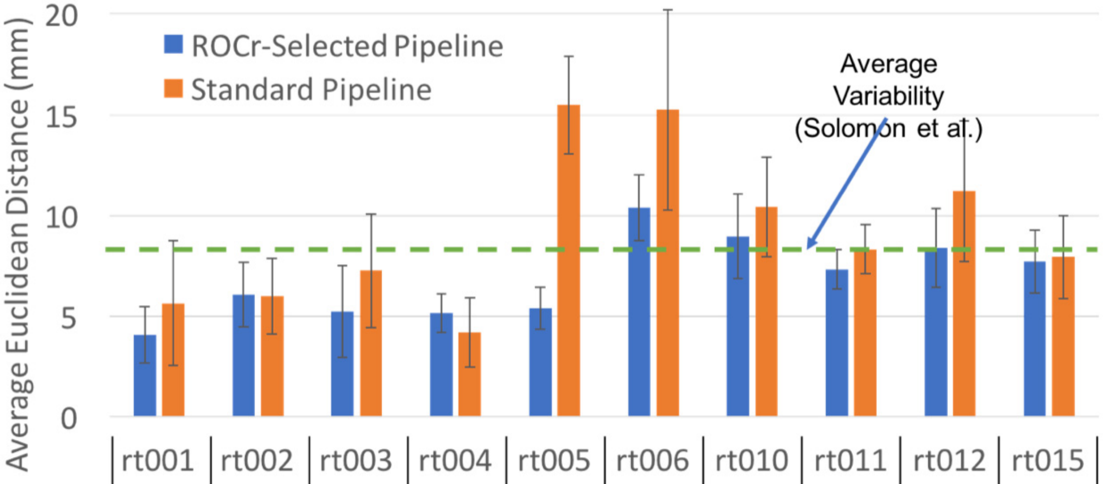
Table 9 shows the average Euclidean distances for each subject, obtained using the ROCr-selected pipelines and the standard pipeline on each of the three days. This table also shows if Hypothesis B was met for each subject. Figure 36 displays this same information graphically. In terms of this study, Hypothesis B states that the average ED (inter-session variability) from the ROCr-selected pipeline should be better than or equal to the average ED from the standard pipeline. Only one subject did not support the hypothesis. Of the nine subjects that did meet the hypothesis, three had average EDs that were equivalent between the ROCr-selected and standard pipelines, and seven had better EDs when the ROCr-selected pipeline was used. On average, the ROCr-selected pipelines' EDs are 2mm smaller than for the standard pipelines.

Previous work with these data have shown that, on average, the S1 location should not vary more than 8mm across days. When the ROCr-selected pipelines were used, seven subjects had EDs that were within 8mm. When the standard pipeline was used, five subjects had EDs smaller than 8mm.

Subject	ROCr Average Euclidean Distance (mm)	Standard Average Euclidean Distance (mm)	Hypothesis Met?	t-test p value
rt001	**4.09***	5.65*	**Yes (less)**	**< 0.001**
rt002	6.09*	**6.01***	Yes (equal)	0.676
rt003	**5.24***	7.27*	**Yes (less)**	**< 0.001**
rt004	5.15*	**4.19***	No	< 0.001
rt005	**5.41***	15.49	**Yes (less)**	**< 0.001**
rt006	**10.39**	15.25	**Yes (less)**	**< 0.001**
rt010	**8.96**	**10.43**	**Yes (less)**	**< 0.001**
rt011	7.33*	8.33	Yes (equal)	0.146
rt012	**8.40**	**11.22**	**Yes (less)**	**< 0.001**
rt015	7.72*	7.94*	Yes (equal)	0.238

Table 9: Average Euclidean Distances.

The second column shows the average of the Euclidean distances (a measure of inter-session variability) obtained using the ROCr-selected pipelines. The third column shows the average ED obtained using the standard pipeline. The fourth column states if the hypothesis was met for each subject. The fifth column shows the p value from the t-test used to test the hypothesis. Only one subject did not meet the hypothesis. "*" indicates that the ED is within the expected 8 mm variance of S1 location across days.



	rt001	rt002	rt003	rt004	rt005	rt006	rt010	rt011	rt012	rt015
ROCr </= Standard?	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
Significance	<0.01	0.68	<0.01	<0.01	<0.01	<0.01	<0.01	0.15	<0.01	0.24

Figure 36: Visual Representation of Average Euclidean Distances.
The x-axis indicates the subject, and the y-axis indicates the average ED (measure of inter-session variability) across sessions. Blue bars indicate the ROCr-selected pipeline, and orange the standard pipeline. The green line indicates the average variability that is expected with these data, as previously reported by Solomon and colleagues. Below each subject, it is indicated whether or not the hypothesis criteria were met for that subject, as well as the significance of the t-test. Only one subject, subject rt004, did not meet the hypothesis. However, both the standard pipeline and the ROCr-selected pipeline had inter-session variability that was within the expected 8 mm limit. Many subjects have inter-session variability that exceeds 8mm, but this is more extreme with the standard pipeline than the ROCr-selected pipeline.

Figure 35 shows histograms of S1 locations in three different planes, comparing the results of the ROCr-selected pipeline to the standard pipeline. Each histogram uses the same colour bar. Figure 35 a), d) and g) show MRI slices of each plane indicating the approximate location of each histogram.

Figure 35 b) and c) show the histograms for the axial plane. Both pipelines localized S1 most frequently at XY coordinate (25, 10), but the ROCr-selected pipeline was more consistent in its localization. It localized the S1 location at (25, 10) around 75% of the time, whereas the localization from the standard pipeline was primarily split between (25, 10) and (20, 10). The standard pipeline had an additional outlying location.

Figure 35 e) and f) show the histograms for the sagittal plane. Both pipelines localized S1 most frequently at YZ coordinate (10, 35), but the ROCr-selected pipeline was more consistent in its localization. It localized S1 at this location 75% of the time, whereas the standard pipeline localized S1 at this location around 65% of the time. The standard pipeline also had more outlying locations than the ROCr-selected pipeline.

Figure 35 h) and i) show the histograms for the coronal plane. Both pipelines localized S1 most frequently at XZ coordinate (25, 35), but the ROCr-selected pipeline was more consistent in its localization. It localized S1 at this location 65% of the time, whereas the standard pipeline S1 localization was primarily split between (25, 35) and (20, 35). In this plane, the ROCr-selected pipeline had a few more outlying locations than the standard pipeline.

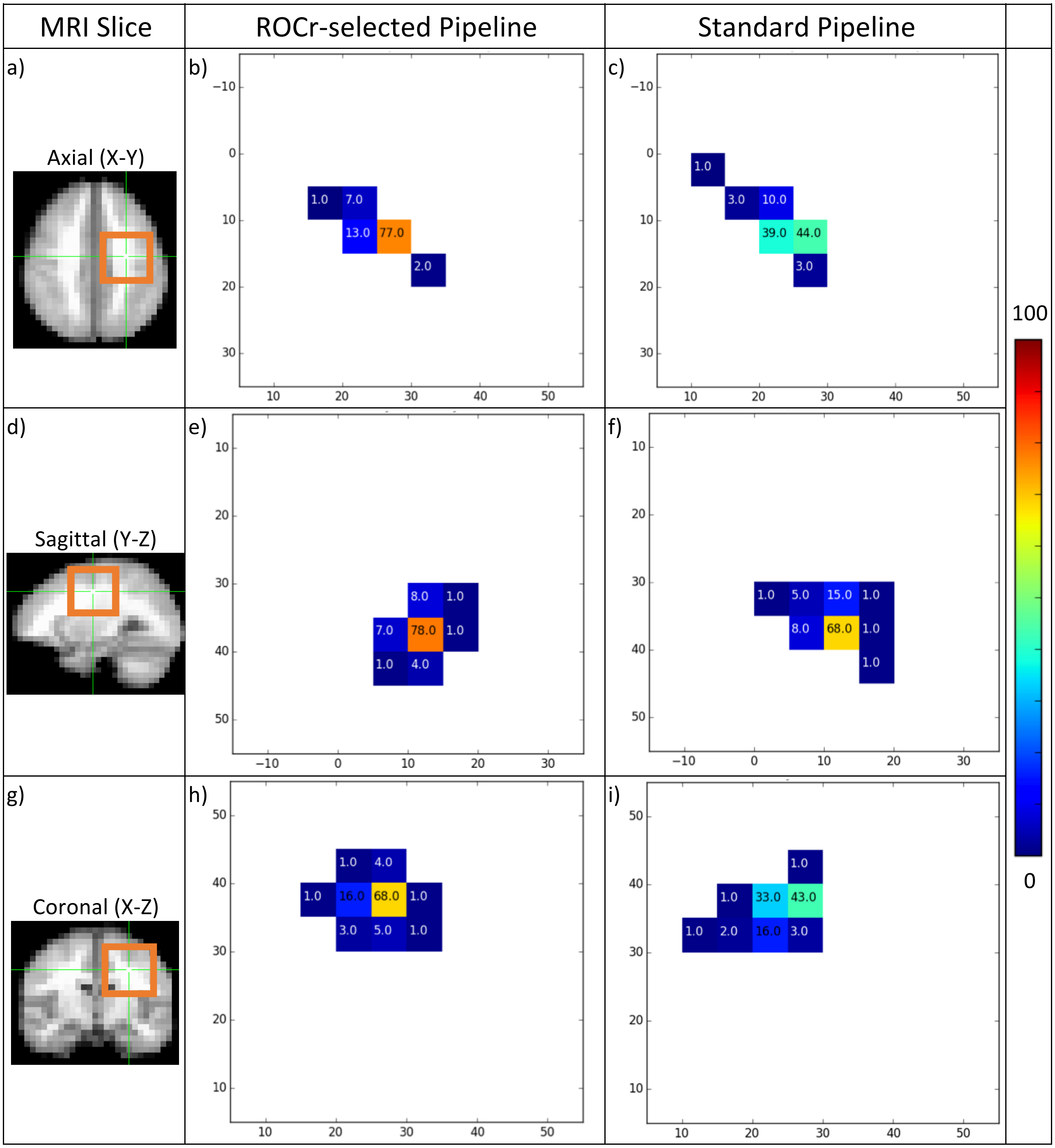


Figure 37: S1 Location Histograms, Subject rt001.
(a) Axial (X-Y) MRI slice. Orange box indicates area shown in histograms to the right. (b) X-Y histogram for the ROCr-selected pipeline. (c) X-Y histogram for the standard pipeline. (d) Sagittal (Y-Z) MRI slice. Orange box indicates area shown in histograms to the right. (e) Y-Z histogram for the ROCr-selected pipeline. (f) Y-Z histogram for the standard pipeline. (g) Coronal (X-Z) MRI slice. Orange box indicates area shown in histograms to the right. (h) X-Z histogram for the ROCr-selected pipeline. (i) X-Z histogram for the standard pipeline. In each plane, the ROCr-selected pipeline had more focal localization than the standard pipeline.

3.6.2 Intra-Session Reliability vs. Inter-Session Variability

Figure 38 is a scatter plot showing the relationship between reliability (F_R) and inter-session variability (ED), with a kernel density estimate (KDE) plot to indicate where the majority of points are. Histograms were also included for the reliable fraction and Euclidean distances to better show the distribution of points. The relationship between intra-session reliability and ED (inter-session variability) is very weak, with an R value of -0.333. As shown in the top histogram, you can get a very good Euclidean distance for a wide range of reliability. However, the very good F_R values (0.85 and above) consistently have an ED within the 8 mm day-to-day variance of S1 localization, as indicated by the local peak in the KDE plot around a reliable fraction of 0.85.

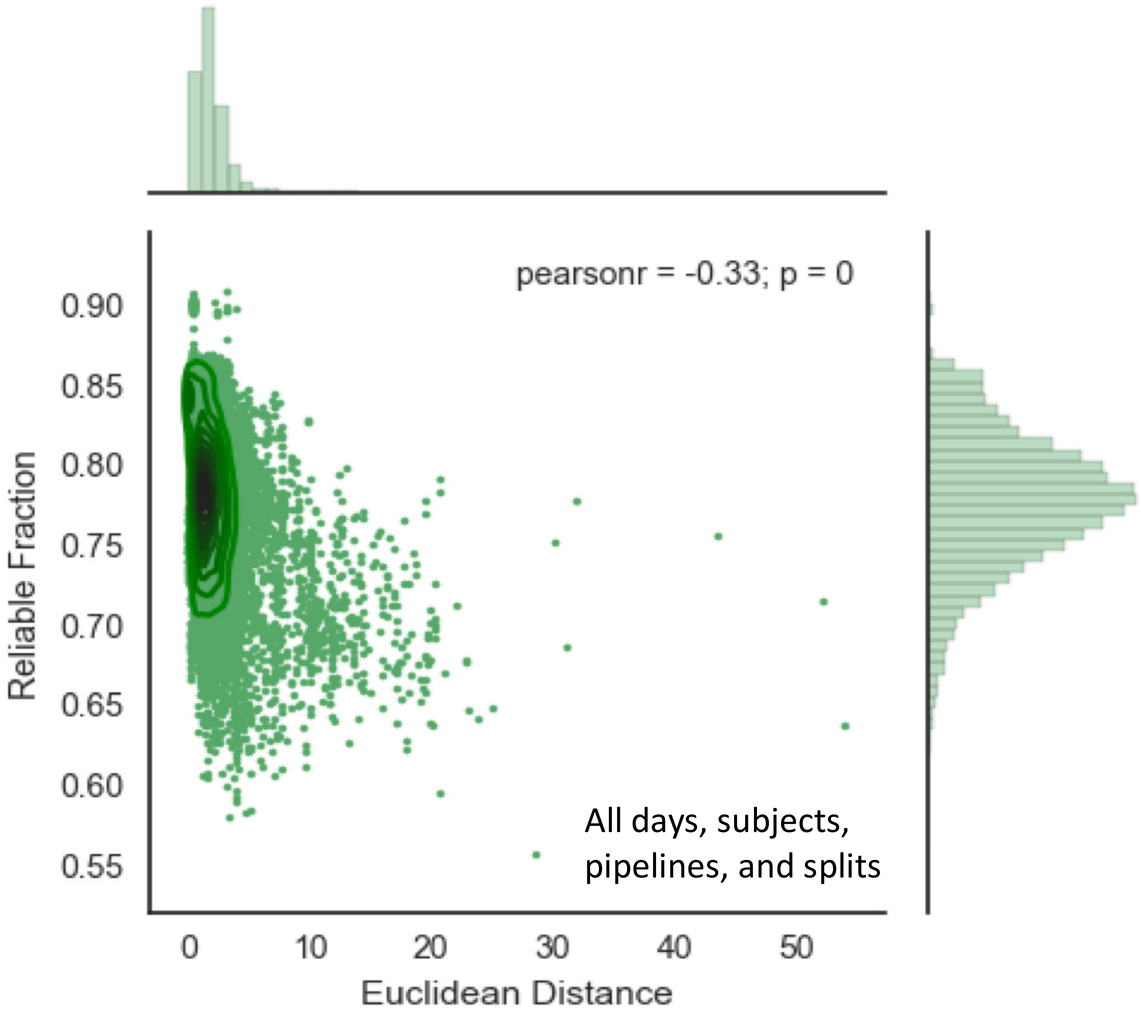


Figure 38: Reliable Fraction vs. Euclidean Distance.
X-axis shows Euclidean distance in mm. Histograms are included to show the distribution of reliable fractions (right) and Euclidean distances (top) in the data. Green lines indicate a KDE (kernel density estimate) plot overlaid on the scatter plot. The majority of points have a Euclidean distance under 8mm. The r value of -0.33 indicates that there is not a relationship between reliable fraction and Euclidean distance.

CHAPTER 4 DISCUSSION

The aim of this thesis was to investigate ROCr as a tool for patient-specific pipeline selection. The investigation demonstrated that ROCr shows promise, despite not meeting Hypothesis A. A key assumption made in the study was proven to be incorrect, and under these circumstances the results from ROCr make sense. The investigation also explored the relationship between inter- and intra-session reliability. Finally, over the course of the investigation several considerations were noted for future studies using ROCr. Before these points can be discussed further, I will provide context for these discussions by summarizing the main findings of my studies and showing how my work builds upon previous studies by my lab group.

4.1 Summary of Main Findings

4.1.1 Preliminary Analysis

Preliminary analysis was done to generate the data needed for Study 1 and Study 2. Each of the ten subjects' raw MEG data was processed 50 times using each of the twelve pipelines, and the resulting activation maps were analysed using ROCr. The processed MEG data showed variance between sessions, pipelines, and subjects. For each pipeline, a distribution of S1 locations and reliable fractions were obtained. The reliability values also showed variance between sessions, pipelines, and subjects. The distribution of all reliable fractions appeared left-skewed. This makes sense because there is a theoretical and practical upper limit to the reliable fraction. The location of S1 was seen to vary over time. Activation maps showed a focal S1 response around the P35m peak, with a stronger but more disperse later response.

4.1.2 Study 1

This study was done to address Hypothesis A and C. Briefly, Hypothesis A states that ROCr will "select" the same pipeline for each of a subject's three data sessions. On a given day, the ROCr-selected pipeline was the pipeline with the highest reliable fraction. Hypothesis A was only met for three of the ten subjects. For many datasets, there were multiple pipelines with reliable fractions that were not significantly different from each other. To investigate these data in a different way, each pipeline was assigned a rank based on its reliable fraction. These rankings were used to identify trends in the pipelines. Hypothesis C briefly states that restricting the ROCr analysis to the contralateral hemisphere will improve the reliable fraction. This hypothesis was tested using only the standard pipeline, across all subjects and sessions. On average, Hypothesis C was met, but additional investigation is required to see if this effect is observed across all pipelines.

4.1.3 Study 2

This study was done to address Hypothesis B, which briefly states that for each subject the average Euclidean distances of S1 locations from the ROCr-selected pipelines should be better than or equal to those from the standard pipeline. The average Euclidean distances were calculated for each subject and session, and averaged together across days. Hypothesis B was met for nine out of the ten subjects. Additional work was done to investigate the relationship between the reliable fraction and Euclidean distance. Euclidean distances were calculated for all pipelines, and compared to the reliable fractions. The resulting relationship was significant, and although there was not a strong linear relationship between intra-session reliability and inter-session variability, the investigation yielded interesting results.

4.2 Continuation of Previous Work with ROCr

As discussed in the introduction, previous work has been done regarding the reliability of functional neuroimaging data. My work extends upon four studies done by my lab group, based on two sets of data. The first dataset (which I did not use) was elicited using a sensorimotor task. Three sessions were recorded, one right after the other. The second dataset (which I used) was generated using median nerve stimulation. Three sessions were recorded, each on separate days. The following paragraphs discuss how my work compares to and builds upon these studies, and provides a basis from which other discussions can take place.

4.2.1 Sensorimotor Task

The first studies done by Stevens et al using ROCr were done on sensorimotor data recorded using fMRI. As previously reported (22), three sessions were recorded, one right after the other. Because the sessions were collected on the same day without having to repeat the set-up, the inter-session variability should be low. The datasets were analyzed using ROCr to characterize their reliability. Because fMRI data cannot be epoched and split like MEG data can, true intra-session reliability could not be calculated for these data. However, since the intra-session variance is estimated to be low, the between-sessions reliability is likely to be similar to intra-session reliability.

The first study using this data was done by Stevens et al.(22). This study used ROCr as a tool to select an image threshold for each subject, as well as a general QA measure. They found that reliability was highly dependent on the analysis pipeline and the subject. Because of the limited number of sessions, they only had three test-retest pairs. My study also investigates ROCr as a QA measure, but primarily investigates it as a tool for automated pipeline selection. Like Stevens et al., I also observed differences in reliability depending on subject and pipeline. A key improvement in my study is the number of test-retest pairs. By taking advantage of the nature of MEG data, I was able to split each dataset in half as many times as needed to achieve converging reliable fractions, and obtain distributions of outcome measures for meaningful statistical analysis.

The second study using this data was also done by Stevens et al. This study involved replicating the fMRI data acquisition with MEG (21). Eight pipelines were selected to analyze these data, and ROCr was used to assess the reliability of each pipeline. The eight pipelines used distinct parameters (such as stimulus timing) that had big impacts on the data, resulting in a wide range of reliable fractions across pipelines. My study uses 12 pipelines instead of 8, and these pipelines focus heavily on frequency filtering. As a result, my pipelines had a narrower range of reliable fractions. In many datasets, several pipelines tied for first place. In fact, there were only 7-9 datasets out of 30 where there was a true “best” pipeline, depending on the ROCr method used. The narrow range of reliability measures is most likely because of how similar my pipelines were.

4.2.2 Median Nerve Stimulation

Recent ROCr studies used median nerve stimulation data, which was collected in the manner discussed in section 2.2. Briefly, subjects underwent MNS on three different days during an MEG scan. Because the scans were done on different days, each session had its own set-up and environmental conditions. The inter-session variance is therefore likely higher than for the sensorimotor experiment, but because the MNS data were recorded using MEG, not fMRI, it is possible to calculate true intra-session reliability as well as inter-session reliability.

In 2015, Solomon et al used this data to measure inter-session reliability (inter-session variance) in MNS data. In this study they localized S1 using the P35m peak, and only one pipeline. The intersession reliability was characterized using the Euclidean distance. They found that the location of S1 may vary up to 8mm across sessions. My study builds upon this work by comparing intra-session reliability to inter-session reliability, using the average Euclidean distance as presented by Solomon et al. Across pipelines and subjects, I found some datasets with inter-session variabilities greater than 8mm. This may be caused by several factors. First, I used twelve pipelines instead of just one optimized pipeline. It was expected that some pipelines would perform poorly, resulting in lower quality localization. Also related to pipelines, most of the time S1 localization was not done using the P35m peak. The later responses I localized are more dispersed, which could increase the average Euclidean distance. Despite the increased Euclidean distances, my study advances the field by providing a comparison between intra- and inter-session reliability.

In 2016, Stevens et al did a study that assessed ROCr as a quality assurance measure, similar to other MEG QA measures. This study also localized S1 using the P35m peak. Only one pipeline was used, but it had an LPF of 70Hz. A key advancement with this study was the ability to assess intra-session reliability by splitting each dataset in half. This was done eight times per dataset. My study used twelve pipelines and most often localized S1 using a late SEF peak (not the P35m). My study extends Stevens et al.'s work farther by splitting each dataset in half 50 times, improving the accuracy of the reliable fraction and providing a distribution of values for statistical analysis.

4.3 Does ROCr Work?

Hypothesis A was disproven, but the results indicate that ROCr still shows potential as a tool for analysis of functional neuroimaging data. This section discusses three main points regarding ROCr performance. First, based on preliminary investigations, inter-session variance in the data appears larger than expected, so the fact that Hypothesis A was disproven does not mean that ROCr is not working. Second, ROCr identified trends in the group-level data, matching trends found in other studies. Finally, the investigation revealed other issues with the experiment that may improve ROCr for future use. These topics are discussed in more detail in the following paragraphs.

4.3.1 Inter-session variance

A key assumption in this study was that inter-session variance (in activation) is minimal. This assumption was partially based on work by Solomon et al., which demonstrated that the S1 location should not vary by more than 8mm between sessions (11). MNS is a very robust paradigm, but the conditions under which the data is recorded can vary from day to day. For example, during each scan, the stimulus may be applied in a slightly different location with a different strength, and the subject’s ability to stay still may vary. Even within a single subject's scan, there can still be variance. Stevens et al. reported that the pipeline that gave the best localization for left MNS was not always the same as the pipeline that gave the best localization for right MNS (21). Evidence of inter-session variance was seen at all stages of the analysis.

The processed MEG data in my study demonstrated high variability between sessions. Looking at the topography and butterfly plots generated in my analysis, there are different temporal and spatial

patterns across sessions for the same pipeline as shown in

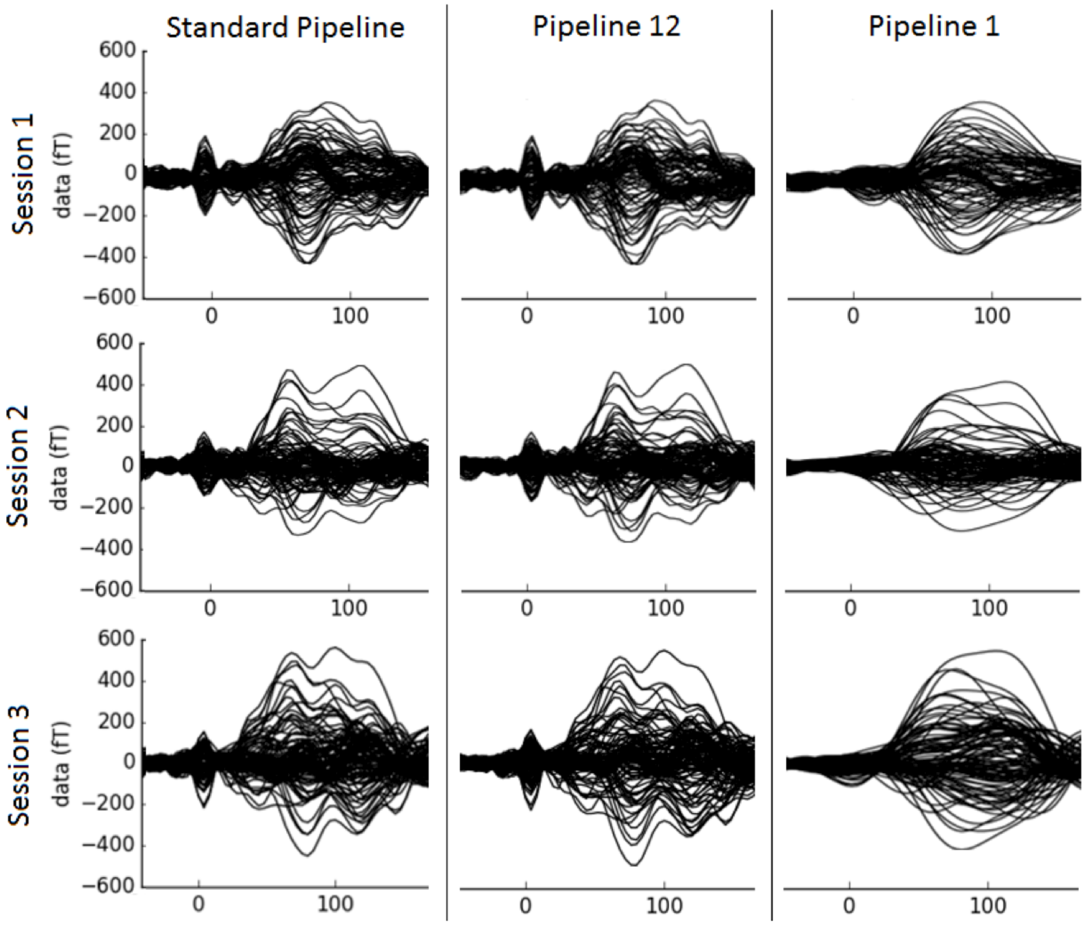


Figure 14. The standard pipeline and Pipeline 12 vary in two parameters (HPF and ICA), but their magnetic field deflections for subject rt006 appear to be identical. When the magnetic field deflections are compared across sessions, the differences are much more noticeable. To verify these claims, additional analysis should be done to compare the inter-session data variance to the inter-pipeline variance in a quantitative way.

Inter-session variance was also seen in the reliable fraction once ROCr had been used to assess the reliability of activation maps. An ANOVA revealed that session had a significant effect on both the reliable fraction and pipeline rankings, using both the Overall Reliability and the Reliability vs. Time methods. Specifically, the third session had, on average, the highest reliable fractions and the best rankings. This may be the result of better performance from the subjects, once they have done several MEG scans and are more used to the procedure. Similarly to the magnetic field deflections, the

Reliability vs. Time plots also revealed inter-session variance as shown in Figure 20 and Figure 21. The variance across sessions often appeared larger or more unpredictable than the variance across pipelines. The variances should be compared in a quantitative manner to validate the qualitative observations.

We expect differences in the average reliable fraction across subjects and pipelines because of inter-subject variance and differences in the analysis pipelines. At the subject level, even though their brain is (hopefully) not re-arranging between scans, the environmental conditions (such as MNS positioning and strength) and behaviour (such as how still the subject is) during the scan will vary. If ROCr is working correctly, it should be able to account for any inter-session variances in the data. At the same time, each session should not be completely different. This behaviour indicates that ROCr is working as it should.

Based on the evidence found in my study and previous work, the assumption of minimal inter-session variance is not correct. Based on my observations, the variance in data between sessions often appears greater than the variance in data between pipelines. If the inter-pipeline data variance were greater than the inter-session data variance, the latter might not have as negative an impact. Hypothesis A was not supported by these data, but this does not discredit ROCr as a tool for automated pipeline selection.

4.3.2 Pipeline Trends

The pipeline trends observed in this investigation also display the benefits of using ROCr, by highlighting why a tool like ROCr is important. Several trends were observed in the average pipeline rankings (see Figure 25 and Figure 28), and paired t-tests verified that these trends were significant as shown in Table 6 and Table 8. However, further investigation revealed that the data were not normally distributed, and a non-parametric test should be used instead to analyze these data. Also, based on the nature of the data used in the paired t-tests, an unpaired t-test would likely be a more appropriate way to analyze the data, and the degrees of freedom may need to be adjusted.

Due to these issues with the statistics, the observed trends can no longer be treated as significant, but are discussed briefly. The first trend was that the SSS pipeline gave better results than the standard pipeline. This trend is discussed in more detail in section 4.5.1. Some other trends matched results from previous studies with this data. My results indicated that increasing the LPF cut-off frequency improved results, not using ICA gave better results than using ICA, and standard analysis options were not always the best. Stevens et al. reported similar findings (21).

The presence of trends such as these in the average reliable fractions may indicate that ROCr can pick up on the expected group level behavior. More importantly, these trends highlight the need for ROCr as a tool in analysis of functional neuroimaging data. Not all subjects matched these trends, and these subjects would be examples of why ROCr is important, because it can select a pipeline that will give better results than just following the trends or using the standard pipeline.

4.3.3 Improvements

Throughout the project, several issues with the experiment design were identified as areas for improvement. These issues include: robustness of the MNS paradigm, similarity of pipelines, late timing for S1 localization, and scheduling of MEG sessions. For each issue, a potential solution is discussed.

The first issue with this experiment is the robustness of median nerve stimulation. MNS is attention and arousal independent, has a direct way to measure stimulus timing, and does not have a task for the subject to perform well or poorly. This resulted in a small range of reliable fractions, that could not be statistically differentiated, so there were many days where more than one pipeline tied for highest reliable fraction. Good localization could be achieved with almost any pipeline because the data was so robust. A more variable task, such as the sensorimotor paradigm used in previous ROCr studies, may be a better choice as it has been demonstrated to give a wider variety of reliability across pipelines.

Another area for improvement was the analysis pipelines. The pipelines used in my investigation were too similar, and twelve pipelines were likely too many. This resulted in a small range of reliable fractions with many tying for highest reliability, which had an impact on Study 1 and Study 2. The eight pipelines in Stevens' thesis had distinct effects on the data in varied ways (21). For example, one of the pipeline parameters affected the stimulus timing, one affected filtering frequency, and one affected ICA. My pipelines varied primarily in terms of frequency, and based on other studies with the same data my LPF frequencies were too low to preserve the P35m peak (this behaviour was not clear from the TFR plot, but should be considered for future work with these data). Selecting pipelines with more distinct parameters would produce a wider range of reliable fractions and increase inter-pipeline variance. If inter-pipeline variance becomes larger than inter-session variance, this may resolve some of the issues with Hypothesis A. Reducing the number of pipelines (especially by eliminating similar pipelines) would increase spacing between the remaining pipelines, allowing for a clear top-ranked pipeline.

The third area for improvement was the timing used to localize S1. Clinical MNS uses the N20m peak, but with a low sampling rate or a low number of epochs, it is challenging to detect a strong enough N20m peak. Previous work by Stevens et al. demonstrated that the P35m peak localizes S1 in the same location as the N20m peak, but is much easier to detect in non-clinical data. Because of the pipeline parameters used in my thesis, I was unable to consistently localize S1 using the P35m peak. Instead, I was often localizing a later S1 response. As shown in Figure 16, the later response is more disperse and bilateral at later times than at around 35 ms, and localizes to a slightly different location. To overcome this issue, localization could have been restricted to a window around 35 ms. Pipelines that could preserve the P35m peak would have more reliable results than pipelines that reduced the peak, such as pipelines with low LPF cut-off frequencies. This would also help to increase the range of reliable fractions, which would have potential benefits for Study 2.

4.4 Relationships between Intra-Session Reliability and Inter-Session Variability

Study 2 aimed to compare intra-session reliability and inter-session variability. I found that, on average, using the pipelines with the highest reliable fraction (highest intra-session reliability) resulted smaller average Euclidean distances (better inter-session variability) than the standard pipeline, which often did not have the highest average reliable fraction. As shown in Table 9, there was only one subject for whom Hypothesis B was not met, and for this subject, both pipelines had average EDs within the 8mm maximum reported by Solomon et al. (11). At the session level though, sometimes the standard pipeline was better.

The average EDs were often larger than those reported by Solomon et al. for these data. Even when the ROCr-selected pipelines were used, four out of the ten subjects had across-session EDs greater than 8mm, indicating higher than expected inter-session variance. At the individual session level, there may be more examples of pipelines with large inter-session variances. It would be beneficial to investigate this in future studies. One possible explanation for the increase in inter-session variability is the fact that the test and retest activation maps used for localization only have half the number of epochs. Because fewer epochs are averaged together, the signal-to-noise ratio is reduced. Specifically, reducing the number of epochs by a factor of two will reduce the SNR by a factor of $\sqrt{2}$. This increased level of noise in the data may reduce the quality of S1 localization.

As an extension of Hypothesis B, I investigated the relationship between intra-session reliability and inter-session variability across all pipelines. The results (shown in Figure 38) showed that the reliable fraction cannot be predicted from ED (r = -0.333), but the relationship was significant. These findings are somewhat at odds with the results from the first part of Study 2. However, a slightly different method had to be used to calculate the average S1 location for all pipelines, so this may have affected the results. Also, as seen in Figure 38, only a small percentage of F_R values were below 0.70. Adding more data points with low intra-session reliability would give new information about the relationship between intra-session reliability and inter-session variability.

Ideally, there would be a linear relationship between reliability, indicating that increasing intra-session reliability reduces inter-session variability. The relationship between these two values may appear weak because of the limited range of reliable fractions. The majority of the values are between 0.75 to 0.80, making it challenging to characterize the relationship between intra-session reliability and inter-session variability across all reliability values.

Despite the limited range of reliable fractions, there are some interesting observations that can be made from the plot. For example, the mid-range of F_R values have a wider distribution of EDs, but the likelihood of inter-session variability greater than 8mm appears very low in this range. The lower F_R values (0.70 and below) appear to have a much larger chance of giving an ED in the 10-20 mm range.

4.5 ROCr Considerations

In this thesis, ROCr has shown potential as a tool for presurgical planning. However, more studies are required before ROCr can be used clinically. Some considerations were noted during the investigation that may be beneficial for future studies involving ROCr.

4.5.1 Reliable Artefacts

A caveat of ROCr is that it cannot distinguish between reliable brain signals and reliable artefacts. An example of this can be seen with the SSS pipeline. The most surprising result from this study was the high occurrence of the SSS pipeline having the highest reliable fraction as shown in Table 3 and Table 4. This pipeline was intended to be a poorly performing pipeline in comparison to the tSSS pipelines, but was one of the highest-reliability pipelines on 13/30 datasets for the Overall Reliability method, and 8/30 datasets for the Reliability vs. Time method. Additional investigation was done to identify possible causes of this behaviour.

Using the Reliability vs. Time data, I observed that, for data processed with the SSS pipeline, reliability is often higher in the -100 to 50 ms range than all other pipelines. An example of this can be seen in Figure 21. The Reliability vs. Time plot for the SSS pipeline shows an average reliability of around 0.2 in the -100 to 0 ms time range, whereas the other pipelines average around 0.1. The SSS pipeline reliability increases at 0 ms due to the stimulus, but does not drop during the 0 to 20 ms time range between stimulus and expected brain activation. This phenomenon was not observed in all subjects and sessions, but high reliability in the -100 to 50 ms range seems to correlate with the SSS pipeline having high reliability and being the "ROCr-selected" pipeline.

One reason for this behaviour could be that the SSS pipeline is better preserving the early SEF responses such as the N20m and the P35m. However, investigation of the topography plots revealed abnormal topographies for these subjects, such as halo-like areas of strong magnetic field deflections. This indicates that the SSS pipeline is likely unable to reduce the effects of the MNS artefact. It would be worthwhile to investigate the correlation between the magnitude of the MNS artefact and the reliable fraction obtained using the SSS pipeline. This may explain the high variance seen in the SSS pipeline ranking, indicating that there are days where the pipeline has a very high reliable fraction and days where the pipeline has much lower reliable fractions.

The frequent occurrence of high reliability seen with the SSS pipeline highlights reliable artefacts as a key limitation of using ROCr to assess reliability. Reliable artefacts may occur in any paradigm. In this study the electrical stimulus caused an issue; in another paradigm a task-related artefact such as speech or finger movement may occur. Another issue highlighted by the SSS pipeline is that ROCr still requires expert knowledge. Someone with no knowledge of SSS or tSSS would not have been surprised that the SSS pipeline performed so well. In this study, the high performance of the SSS pipeline data was only investigated further because it was an unexpected result. Further work is needed to solve the challenging issue of reliable artefacts.

4.5.2 Comparison of Methods

Two main methods were used to determine the reliable fraction in this investigation, and each has its challenges. In general, Overall Reliability appears to be skewed by data quality at times of non-interest, and the Reliability vs. Time approach may over-estimate reliability. The following paragraphs discuss the two methods in more detail.

An ANOVA revealed that the Reliability vs. Time method gave higher reliable fractions across all pipelines and sessions. This is likely because the Reliability vs. Time approach is picking the best possible time in the data, whereas the Overall Reliability method is giving an average measure of reliability. The relationship between the two methods is non-linear and more complicated than described here, but in general that is the main difference between the two methods.

While it might seem like a good thing to get higher reliable fractions, when all pipelines get high reliability it is harder to distinguish between pipelines. This is especially an issue when these high rankings are likely over-estimates of reliability. Another downside of the Overall Reliability method is that the range of reliable fractions is reduced. We see evidence of this when looking at the effect of session on the reliable fraction and pipeline rankings. The differences across sessions were larger for the Overall Reliability method than for the Reliability vs. Time method.

The Reliability vs. Time method has some other issues regarding timing. This method reports the reliability at only one time point, which may not be a clinically relevant time. This behaviour was seen frequently in these data: the clinically relevant 35 ms peak was not usually identified as the most reliable time in the data. For this reason, caution should be taken when using the Reliability vs. Time method to determine when to localize S1 (which was not done in this thesis). Additionally, even if there is not a spike in reliability around a clinically relevant time, it may still be acceptable to localize S1 around that time. A prolonged period of high reliability may obscure a local peak.

The Overall Reliability method also has its limitations. Reliability may be high at clinically relevant times, but this method does not take time into account, so the Overall Reliability may be low. This method is also susceptible to noise that is reliable and persistent, but not large in magnitude. For example, the SSS pipeline likely results in higher magnetic field deflections in the -100 to 50 ms range. The Reliability vs. Time method was not affected by this as much because this method assessed reliability at later time points. As a result, the SSS pipeline had high ranking variance with this method as shown in Figure 32. The Overall Reliability method, however, often ranked the SSS pipeline higher.

Two other methods were tested in the investigation: restricting the data to a hemisphere of interest and restricting the data to a time of interest. For the standard pipeline, reducing data to a hemisphere of interest did improve reliability, but additional investigation is required to determine the benefits of this method. Restricting data to a time of interest gave much worse reliability, but this method was not used to its fullest potential. This method should be tested again, but with the time restriction occurring before ROCr analysis.

4.5.3 Pipeline Rankings vs. Highest F_R Pipeline

Hypothesis A investigated ROCr in a simplistic way. Assessing ROCr based only on the pipeline with the highest reliable fraction does not take advantage of all the information available in the data. Pipeline rankings can give a more representative view of how well ROCr is working. However, a problem with the ranking approach is that the difference between each rank may not be significant. Subject rt015 is a good example of this: the top seven pipelines (in terms of reliable fraction) are statistically equivalent to each other. However, the ranking is calculated for each split and averaged together, which should help account for this effect. If these 7 pipelines are equivalent, they should each have splits where they have the highest reliable fraction, and this will be reflected in the averaged pipeline rankings for that day. Using rankings instead of average F_R values has another benefit: is that it allows for easier comparison of results across subjects. My investigation and other studies demonstrated that average reliability varies across subjects, but rankings puts the data on the same scale.

4.6 Future Work

Further investigation is required before ROCr should be used in a clinical setting. Many suggestions were already made in the discussions on how to continue ROCr investigations in the short-term to address the new hypothesis. Specifically, a lot of benefit would come from improving the pipelines used in the analysis and restricting localization to around 35 ms, regardless of whether there is a peak in the window or not. These suggestions would give a wider range of reliable fractions and S1 locations, which in turn would improve the results of this investigation. A few additional suggestions that were not discussed previously are discussed below.

In this study, scores were assigned to each subject to address Hypothesis A. While these scores gave a new way to look at the data, the scores had a seemingly random distribution. More information could be obtained from the scores by identifying features in the MEG data, and comparing the scores to the presence or absence of these features in the MEG data. For example, does a large magnetic field deflection around the time of stimulus administration correlate with a certain score?

As discussed above, the inter-session variance was much higher than assumed, making it unlikely that the same pipeline would be picked across sessions. The issue of inter-session variance could be reduced by doing all three scans on the same day. This would avoid having to do three set-ups, possibly with different simulation strengths and placements, and reduce the variance of external factors affecting the data. This new method could be improved further by doing one scan that is three times as long as a single scan, and then splitting the data into three parts after.

An alternative method must be found to properly evaluate ROCr. One alternative is presented here. A subject in the sensorimotor study underwent surgery and the primary motor cortex was localized with the clinical gold-standard method. The gold-standard M1 location was then compared to the location obtained using the ROCr-selected pipeline to validate ROCr. This subject's data, and any other similar datasets, provides an excellent opportunity to compare reliability to accuracy on a larger scale, in a way similar to Study 2 of my thesis. The subject's data could be analyzed with a variety of pipelines, including very poor pipelines, to produce activation maps with a wide range of reliable fractions. The motor cortex would then be localized using the activation maps, and the estimated locations compared to the known location. Reliability could then be compared to accuracy. The more datasets available, the more useful this comparison would be.

CHAPTER 5 CONCLUSIONS

This thesis posed the question, “Can ROCr be used as a tool for automated pipeline selection?”. Specifically, three hypotheses were investigated. Hypothesis A stated that ROCr should pick the same pipeline on each of a subject's three datasets. Hypothesis B stated that the localization results obtained with the ROCr-selected pipelines should have smaller inter-session variability (as determined by a smaller average Euclidean distance) than the standard pipeline. Hypothesis C stated that restricting the analysis to the hemisphere of interest should improve the reliable fraction.

My research study used data from ten subjects who underwent median nerve stimulation during three separate MEG scans. I found that hypothesis A was only met for three out of ten subjects. Two subjects had no pipelines in common across sessions, and the remainder had some overlap in pipelines across sessions. Further investigation revealed that the inter-session data variance was much larger than initially assumed. Because this key assumption was violated, Hypothesis A was not an appropriate way to assess ROCr. Based on the characteristics of the data, ROCr appears to perform as expected. In answer to Hypothesis B, all but one subject had lower EDs for the ROCr-selected pipelines. Additional investigation was done to determine the relationship between intra-session reliability (F_R) and inter-session variability (ED), but predictability between these two values was low (r = -0.333). Hypothesis C was proven to be true, but this hypothesis was only tested on one pipeline. Additional investigation is required for this hypothesis.

My work contributes to the literature on ROCr as a tool for quality assurance and pipeline selection by improving the methods used to test ROCr. I have presented an improved procedure of generating a distribution of reliability values for each pipeline, increasing statistical power. I have also identified some key issues with the approach I took to evaluating ROCr. Despite the issues with my experiment, many of my findings concur with previous studies by Stevens et al. and Solomon et al. using ROCr.

These findings could be of interest to anyone planning to use ROCr in future studies. I have highlighted considerations that will be essential for using and validating ROCr. As a result of my thesis, future researchers working with ROCr will be better informed of its limitations and challenges for use. There are still obstacles to be overcome before ROCr can be used clinically. Once ROCr has been validated, however, it has potential to be used as a quality assurance measure for functional neuroimaging data as well as automated pipeline selection. These improvements to functional neuroimaging analysis may improve pre-surgical mapping, leading to better surgical outcomes for patients.

BIBLIOGRAPHY

1. Oishi M, Fukuda M, Kameyama S, Kawaguchi T, Masuda H, Tanaka R. Magnetoencephalographic representation of the sensorimotor hand area in cases of intracerebral tumour. J Neurol Neurosurg Psychiatry [Internet]. BMJ Group; 2003 Dec;74(12):1649–54. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1757408/

2. Abou-Khalil B. Methods for determination of language dominance: The wada test and proposed noninvasive alternatives. Curr Neurol Neurosci Rep [Internet]. Current Science Inc.; 2007 Nov 20 [cited 2016 Sep 10];7(6):483–90. Available from: http://link.springer.com/10.1007/s11910-007-0075-6

3. Enatsu R, Mikuni N. Invasive Evaluations for Epilepsy Surgery: A Review of the Literature. Neurol Med Chir (Tokyo). Scientific Journal Publishing Dept., Medical Tribune, Inc.; 2016;56(5):221.

4. Loddenkemper T, Morris HH, Möddel G. Complications during the Wada test. Epilepsy Behav. 2008;13(3):551–3.

5. Simos PG, Breier JI, Maggio WW, Gormley WB, Zouridakis G, Willmore LJ, et al. Atypical temporal lobe language representation: MEG and intraoperative stimulation mapping correlation. Neuroreport [Internet]. 1999 Jan 18 [cited 2016 Sep 16];10(1):139–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10094150

6. Szymanski MD, Perry DW, Gage NM, Rowley HA, Walker J, Berger MS, et al. Magnetic source imaging of late evoked field responses to vowels: toward an assessment of hemispheric dominance for language. J Neurosurg [Internet]. 2001 Mar [cited 2016 Sep 16];94(3):445–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11235950

7. Fisher AE, Furlong PL, Seri S, Adjamian P, Witton C, Baldeweg T, et al. Interhemispheric differences of spectral power in expressive language: a MEG study with clinical applications. Int J Psychophysiol [Internet]. 2008 May [cited 2016 Sep 16];68(2):111–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18316134

8. Kim JS, Chung CK. Language lateralization using MEG beta frequency desynchronization during auditory oddball stimulation with one-syllable words. Neuroimage [Internet]. 2008 Oct 1 [cited 2016 Sep 16];42(4):1499–507. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18603004

9. Pittau F, Grouiller F, Spinelli L, Seeck M, Michel CM, Vulliemoz S. The role of functional neuroimaging in pre-surgical epilepsy evaluation. Front Neurol [Internet]. Frontiers Media SA; 2014 [cited 2016 Sep 12];5:31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24715886

10. Osorio JA, Aghi MK. Optimizing glioblastoma resection: intraoperative mapping and beyond. CNS Oncol. England; 2014;3(5):359–66.

11. Solomon J, Boe S, Bardouille T. Reliability for non-invasive somatosensory cortex localization: Implications for pre-surgical mapping. Clin Neurol Neurosurg. 2015;139:224–9.

12. Hämäläinen M, Hari R, Ilmoniemi RJ, Knuutila J, Lounasmaa O V. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev Mod Phys [Internet]. American Physical Society; 1993 Apr 1 [cited 2016 Sep 16];65(2):413–97. Available from: http://link.aps.org/doi/10.1103/RevModPhys.65.413

13. Hall EL, Robson SE, Morris PG, Brookes MJ. The relationship between MEG and fMRI. Neuroimage [Internet]. 2014 [cited 2017 Apr 11];102:80–91. Available from: http://www.sciencedirect.com.ezproxy.library.dal.ca/science/article/pii/S1053811913010975

14. Castillo EM, Simos PG, Wheless JW, Baumgartner JE, Breier JI, Billingsley RL, et al. Integrating sensory and motor mapping in a comprehensive MEG protocol: Clinical validity and replicability. Neuroimage [Internet]. 2004 Mar;21(3):973–83. Available from: http://www.sciencedirect.com/science/article/pii/S1053811903006736

15. Korvenoja A, Kirveskari E, Aronen HJ, Avikainen S, Brander A, Huttunen J, et al. Sensorimotor Cortex Localization: Comparison of Magnetoencephalography, Functional MR Imaging, and Intraoperative Cortical Mapping. Radiology [Internet]. Radiological Society of North America; 2006 Oct 1;241(1):213–22. Available from: http://dx.doi.org/10.1148/radiol.2411050796

16. Forss N, Narici L, Hari R. Sustained activation of the human SII cortices by stimulus trains. Neuroimage [Internet]. 2001 Mar [cited 2016 Sep 19];13(3):497–501. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11170814

17. Cheyne D, Bostan AC, Gaetz W, Pang EW. Event-related beamforming: A robust method for presurgical functional mapping using MEG. Clin Neurophysiol [Internet]. 2007 [cited 2017 Apr 11];118(8):1691–704. Available from: http://www.sciencedirect.com.ezproxy.library.dal.ca/science/article/pii/S1388245707002076

18. Taulu S, Hari R. Removal of magnetoencephalographic artifacts with temporal signal-space separation: demonstration with single-trial auditory-evoked responses. Hum Brain Mapp [Internet]. 2009 May [cited 2016 Sep 7];30(5):1524–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18661502

19. Stolk A, Todorovic A, Schoffelen J-M, Oostenveld R. Online and offline tools for head movement compensation in MEG. Neuroimage. 2013;68:39–48.

20. Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, et al. MNE software for processing MEG and EEG data. Neuroimage. 2014;86:446–60.

21. Stevens MTR. Enhancing the Reliability of Functional MRI and Magnetoencephalography for Presurgical Mapping. Dalhousie University; 2015.

22. Stevens MTR, D'Arcy RCN, Stroink G, Clarke DB, Beyea SD. Thresholds in fMRI studies: Reliable for single subjects? J Neurosci Methods. 2013;219(2):312–23.

23. Genovese CR, Noll DC, Eddy WF. Estimating test-retest reliability in functional MR imaging I: Statistical methodology. Magn Reson Med [Internet]. Wiley Subscription Services, Inc., A Wiley Company; 1997 Sep 1 [cited 2017 Jul 4];38(3):497–507. Available from: http://doi.wiley.com/10.1002/mrm.1910380319

24. Gullapalli RP, Maitra R, Roys S, Smith G, Alon G, Greenspan J. Reliability estimation of grouped functional imaging data using penalized maximum likelihood. Magn Reson Med [Internet]. Wiley Subscription Services, Inc., A Wiley Company; 2005 May 1 [cited 2017 Jul 4];53(5):1126–34. Available from: http://doi.wiley.com/10.1002/mrm.20470

25. Chen EE, Small SL. Test-retest reliability in fMRI of language: group and task effects. Brain Lang [Internet]. NIH Public Access; 2007 Aug [cited 2017 Jul 4];102(2):176–85. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16753206

26. Maitra R. A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps. Neuroimage [Internet]. Elsevier Inc.; 2010;50(1):124–35. Available from: http://dx.doi.org/10.1016/j.neuroimage.2009.11.070

27. Le TH, Hu X. Methods for assessing accuracy and reliability in functional MRI. NMR Biomed. England; 1997;10(4–5):160–4.

28. Stevens MTR, Bardouille T, Stroink G, Boe SG, Beyea SD. Fully Automated Quality Assurance and Localization of Volumetric MEG for Single-Subject Mapping. J Neurosci Methods [Internet]. Elsevier B.V.; 2016; Available from: http://dx.doi.org/10.1016/j.jneumeth.2016.03.008

29. Strother SC, Anderson J, Hansen LK, Kjems U, Kustra R, Sidtis J, et al. The Quantitative Evaluation of Functional Neuroimaging Experiments: The NPAIRS Data Analysis Framework. Neuroimage [Internet]. 2002 [cited 2017 Apr 19];15(4):747–71. Available from: http://www.sciencedirect.com.ezproxy.library.dal.ca/science/article/pii/S1053811901910341

30. Churchill NW, Yourganov G, Oder A, Tam F, Graham SJ, Strother SC. Optimizing preprocessing and analysis pipelines for single-subject fMRI: 2. Interactions with ICA, PCA, task contrast and inter-subject heterogeneity. PLoS One [Internet]. Public Library of Science; 2012 [cited 2017 Apr 19];7(2):e31147. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22383999

31. Churchill NW, Oder A, Abdi H, Tam F, Lee W, Thomas C, et al. Optimizing preprocessing and analysis pipelines for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods. Hum Brain Mapp [Internet]. Wiley Subscription Services, Inc., A Wiley Company; 2012 Mar [cited 2017 Apr 19];33(3):609–27. Available from: http://doi.wiley.com/10.1002/hbm.21238

32. Sharma R, Pang EW, Mohamed I, Chu B, Hunjan A, Ochi A, et al. Magnetoencephalography in children: Routine clinical protocol for intractable epilepsy at the hospital for sick children. Int Congr Ser [Internet]. 2007 Jun [cited 2017 Jun 8];1300:685–8. Available from: http://linkinghub.elsevier.com/retrieve/pii/S053151310700266X

33. Kakigi R, Hoshiyama M, Shimojo M, Naka D, Yamasaki H, Watanabe S, et al. The somatosensory evoked magnetic fields. Prog Neurobiol. 2000;61(5):495–523.

34. Foundation PS. 9.6. random - Generate pseudo-random numbers [Internet]. The Python Standard Library. [cited 2016 Nov 4]. Available from: https://docs.python.org/2/library/random.html

35. D'Arcy RCN, Bardouille T, Newman AJ, Mcwhinney SR, Debay D, Sadler RM, et al. Spatial MEG Laterality maps for language: Clinical applications in epilepsy. Hum Brain Mapp. 2013;34(8):1749–60.