INTRA-SESSION RELIABILITY METRICS FOR QUALITY ASSURANCE IN PRE-SURGICAL MAPPING WITH MAGNETOENCEPHALOGRAPHY

by

Mary Miedema

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2019

Does a poem enlarge the world,
or only our idea of the world?

How do you take one from the other,
I lied, or did not lie,
in answer.

— Jane Hirshfield, from "Mathematics," in *Given Sugar, Given Salt*

This thesis is dedicated to all those whose work and words have enlarged my idea of the world.

Table of Contents

List of Tables

List of Figures

Abstract

Magnetoencephalography (MEG) functional maps can localize brain activity for pre-surgical mapping, but their quality is difficult to quantify. Clinical standard metrics cannot be used when multiple sources of activity are distributed across the brain. This thesis validates the use of reliable fraction, a novel intra-session reliability metric, for focal maps. Scans were acquired in 'good' and 'poor' conditions, in which common MEG quality issues were simulated. Clinical standard methods and reliable fraction, along with two other possible metrics (the Dice and Pearson coefficients), were used to assess data quality. High quality data proved difficult to achieve, highlighting the need for robust quality assurance procedures. Reliable fraction was more sensitive to data quality issues than the Dice or Pearson coefficients. Comparison of reliable fraction with clinical standard metrics showed comparable sensitivity to changes in data quality and suggests reliable fraction may be a useful metric for cases of distributed brain activity.

List of Abbreviations and Symbols Used

A1	Primary auditory cortex
ANOVA	Analysis of variance
AUC	Area under a receiver-operator characteristic curve
BEM	Boundary element model
ECD	Equivalent current dipole
ECG	Electrocardiogram
EEG	Electroencephalography
EMG	Electromyogram
EOG	Electro-oculogram
ERB	Event-related beamformer
ERF	Event-related field
fMRI	Functional magnetic resonance imaging
FPR	False positive rate
Fr	Reliable Fraction
GFP	Global field power
GoF	Goodness of fit
HPI	Head position indicator
ICA	Independent component analysis
LCMV	Linearly-constrained minimum variance
M1	Primary motor cortex

MEG	Magnetoencephalography
MNS	Median nerve stimulation
MRI	Magnetic resonance imaging
QA	Quality assurance
ROC	Receiver-operator characteristic
ROC-r	Receiver-operator characteristic reliability
S1	Primary somatosensory cortex
SNR	Signal-to-noise ratio
SQUID	Superconducting quantum interference device
SSS	Signal space separation
TPR	True positive rate
tSSS	Temporal signal space separation
V1	Primary visual cortex

Acknowledgements

First, I'd like to thank my supervisor, Tim Bardouille, for his guidance and support over the last two years. Thank you for helping me to grow as a scientist while giving me the opportunities to explore my interests and develop new skills. I'd also like to thank my supervisor Steven Beyea and my committee members, Chris Bowen and Javeria Hashmi, for their feedback and enthusiasm over the course of this project.

Thank you also to Miriam, Jenny, Anna, and all of the other friends I've made through the medical physics program here at Dalhousie. You kept me grounded and smiling through the most challenging parts of the past two years.

Finally, I would like to thank my family and friends back home for supporting me from afar and always encouraging me. To my mom in particular: thank you for making it possible to finish my courses when I broke my ankle. I've never doubted your belief in me. To my sister: thank you for sharing the ups and downs and making me laugh through them all. To Rosanna and Christina, my sisters in all but biology: thank you for always being there for me no matter the distance between us. Lastly, to the member of my family who has been by my side through all the sleepless nights and research-fueled rambling: thank you, Brontë. You're a very good dog.

Chapter 1

Introduction

1.1 Pre-Surgical Mapping with Magnetoencephalography

Pre-surgical mapping of the specific functional anatomy of areas in the brain which may be affected by the removal of a lesion is important for planning and executing neurosurgery. In order to preserve function and more accurately predict surgical outcomes, it is crucial to localize areas of the cortex which correspond to abilities such as speech, movement, and touch [1]. While anatomical landmarks may be sufficient to localize these eloquent regions of the brain in healthy individuals, functional mapping is usually necessary for patients with brain tumours, who may experience unpredictable functional and structural reorganization as the result of tumour growth [2]. Pre-surgical mapping of the eloquent cortex in brain tumour patients facilitates the planning of a surgical approach which maximizes the extent of resection while reducing the time required for intraoperative mapping [1]. The pre-surgical functional map can be used as a starting point for intraoperative functional mapping via direct cortical stimulation during the resection surgery in order to ensure that function is preserved [3, 4].

Functional mapping may be performed using a number of techniques, but for the purposes of surgical planning non-invasive mapping with either functional magnetic resonance imaging (fMRI) or magnetoencephalography (MEG) is common. Abnormal blood flow near large gliomas has been shown to interfere with the interpretation of fMRI data [5]; localization of brain activity with MEG has been shown to be more accurate than fMRI in such cases [2, 6]. The use of MEG for pre-surgical mapping has been shown to correlate with favourable surgical outcomes after resection for a number of mapping paradigms [5, 7, 8].

1.1.1 Neurophysiology

Magnetoencephalography records magnetic fields generated by neural activity. Magnetic fields on the order of 50-500 fT can be detected outside the head using highly sensitive superconducting quantum interference devices (SQUIDs). These fields are generated by the cumulative postsynaptic activity of pyramidal cells in the cerebral cortex. The synchronous flow of ions along tens of thousands of parallel dendrites generates a perpendicular magnetic field which can be detected outside the head with temporal resolution better than a millisecond [9].

1.1.2 Signal Measurement and Mapping Paradigms

In order to isolate locations in the brain specific to particular functions, functional measurements are recorded in combination with the presentation of stimuli to evoke the relevant neural activity. Depending on the position of the lesion or tumour, mapping of the location and extent of somatosensory, motor, visual, auditory, and/or language-related regions of the brain may be required. Neuromagnetic deflections corresponding to known functional activity can be generated by a set of simple stimuli or performance of a well-defined task. For example, subjects may be exposed to physical or electrical stimulation to generate a response in the somatosensory cortex, asked to perform simple motions or press a button to generate a response in the motor cortex, passively watch changing geometric patterns or listen to auditory tones to generate a response in the primary visual or auditory cortex, or process and classify more complex verbal stimuli to generate a response in language-related areas [10]. The subject's neural response to such stimuli is collected over a number of trials within each scan, which can be averaged for more accurate localization. After the data has been collected, it is separated into epochs (segments corresponding to each stimulus). Averaging the data over all epochs is standard for analysis of an evoked response, since the signal stemming from a consistent response to the stimulus will be stable across epochs while uncorrelated brain activity will be suppressed.

1.1.3 Source Modelling Techniques and Interpretation

MEG sensor measurements can be used to estimate the location of the underlying generators within the brain. In general, the magnetic field generated by a current source in a conductive volume can be calculated using the Biot-Savart law:

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \mathbf{J}(\mathbf{r}_Q) \times \frac{\mathbf{r} - \mathbf{r}_Q}{|\mathbf{r} - \mathbf{r}_Q|^3} dv \tag{1.1}$$

where $\mathbf{B}(\mathbf{r})$ is the magnetic field outside the conductive volume at position $\mathbf{r}$ caused by a current source $J(\mathbf{r}_Q)$ at position $\mathbf{r}_Q$ within the conductive volume and μ_0 is the magnetic permeability of free space.

However, this inverse problem does not have a unique solution. Constraints must be applied to the problem in the form of a source model and a conductive model for the head volume. There are many methods of solving the inverse problem. Clinically, the most common approach is to specify a certain number of sources and fit each one using an equivalent current dipole (ECD), defined as a current flow over a distance approaching zero with a given strength, position, and orientation.

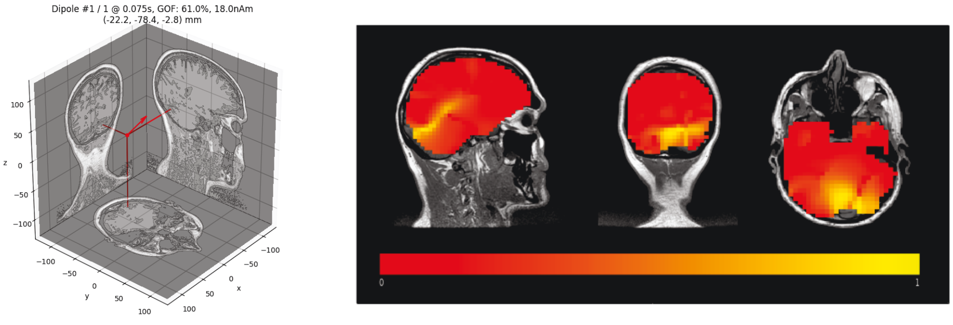
For the simplest case of a single equivalent current dipole (for which $\mathbf{J}(\mathbf{r}_Q) = \mathbf{Q}\delta(\mathbf{r} - \mathbf{r}_Q)$) in a spherical conductive volume, the forward solution may be calculated as follows:

$$B_r = -\frac{\mu_0}{4\pi} \frac{(\mathbf{Q} \times \mathbf{r}_Q) \cdot \mathbf{e}_r}{|\mathbf{r} - \mathbf{r}_Q|^3} \tag{1.2}$$

where B_r is the magnetic field perpendicular to the surface of the conductor, $\mathbf{e}_r$ is the radial unit vector, and all other variables are defined as in Equation 1.1 [11].

The most appropriate magnitude, orientation, and position (i.e. $\mathbf{Q}$ and $\mathbf{r}_Q$) for the ECD can be iteratively estimated by sampling the space of possible solutions until the minimum of a cost function is found. This cost function measures the difference between the measured field and the field predicted by the source model for a given ECD or set of ECDs (i.e. the forward solution). Alternatively, the difference between the measured and predicted fields can be quantified by the goodness of fit, as discussed in section 1.2.2, which is maximized as the cost function is minimized.

However, the ECD approach to source modelling is limited to a small number of sources (one or two), since it becomes significantly dependent on the *a priori* specification of the number and differing allowable properties of the ECD sources. A method of source localization better suited to the case of distributed source activity (i.e. more than one or two focal generators of neural activity located in separate regions of the brain) is the beamformer model. A beamformer can be thought of as a set of spatial filters, which are defined for each position on a pre-specified cortical grid. Each filter is selected to act on the sensor measurements by projecting maximum power to its corresponding source position and minimizing source activity elsewhere [9, 12]. Both methods require the model of the conductive properties of the head, which could range from a simple spherical approximation to a more sophisticated boundary element model constructed from anatomical data. For this reason, functional MEG mapping data is often combined with structural information acquired during magnetic resonance imaging (MRI) for more accurate source localization. Moreover, interpretation of functional MEG maps requires the location of source activity relative to the underlying anatomical structures, so MRI data is typically used for visualization. In Fig. 1.1, localization of a single focal source is shown for both methods and overlaid on anatomical data from an MRI scan. In this case, both methods localize activity to the similar regions of the right occipital lobe. However, neither method offers a clear interpretation of the spatial extent of the underlying source.



(a) Single equivalent current dipole fit

(b) Linearly constrained minimum variance beamformer (normalized to max. activity)

Figure 1.1: Mapping the right primary visual cortex with two methods of source localization. Both methods indicate a strong posterior source in the right occipital lobe approximately 70 ms after a visual stimulus was presented in the right half of this subject's field of view.

The resulting source localization may take the form of an activation map in the case of distributed source activity or one or several points of source activity in the case of ECD fitting. While fMRI activation maps may be thresholded to delineate a statistically significant cluster of source activity, the ill-posed nature of the MEG inverse problem means that the spatial extent of estimated source activity is more difficult to interpret. For this reason, the location(s) of peak activity of a beamformer activation map is typically interpreted as the location of source activity. For either method of source localization, the location of source activity is co-registered with anatomical data. It can then serve as a tool for pre-surgical planning – including risk assessment – and to interactively assist the surgeon using a image guidance system [3, 5].

1.2 Quality of Magnetoencephalography Data

The following sections will examine the causes of poor MEG data quality, standard quality assurance protocols, and potential approaches to quantitative quality assurance. In particular, I will discuss the need for well-defined and broadly applicable quality assurance measures during MEG data acquisition and processing. Our local imaging centre has experienced data losses on the order of 10% during MEG research studies due to quality issues. Extending this estimate to clinical cases is difficult due to low patient throughput, but our experience suggests that low quality mapping data could be acquired for a significant percentage of the patient population. In a clinical scenario, acquiring low quality mapping data could require a patient to return for a second imaging session since quality issues may not be discovered until after data has been processed. Moreover, these scans are typically rejected on the basis of operator inspection, rather than a quantitative assessment. This raises the problem of reproducible and consistent assessment of data quality.

1.2.1 Causes of Poor Data Quality and Compensatory Techniques

Collecting high quality MEG data is often challenging, in large part due to the minute scale of magnetic fields corresponding to neural activity. Although it is preferable to make good quality measurements during data collection, it is possible to compensate for many data quality issues during data analysis. This section will present an overview of common MEG data quality issues and discuss mitigating techniques.

Non-Physiological Artefacts

Environmental generators of magnetic field fluctuations (i.e. environmental noise) are a significant issue when measuring neuromagnetic activity; MEG signals are typically on the order of 10-100 fT and can be significantly obscured by competing magnetic signals. Although MEG data is acquired in a magnetically shielded room, shielding limitations along with sources of noise originating within the room can contaminate the signal arising from a subject's neural activity. The movement of nearby machinery, such as elevators, can also introduce significant signal contamination. Reference sensors far from the sensor helmet but still within the shielding can be used to suppress such sources of environmental noise. Effective noise reduction can be achieved by filtering signals to frequency bands of neurological interest using high-pass, low-pass, or band-pass filtering. Power-line noise is a common source of signal contamination which can be addressed with notch filters at the appropriate frequencies (50-60 Hz and the associated harmonics). High-pass filtering, as well as baseline correction, can be particularly useful to attenuate low-frequency drifts which may affect the MEG sensors throughout the scan [13]. During the scan, the operator should closely monitor sensor measurements to address any artefacts or noisy channels. Data segments containing overpowering artefactual signals or sensor channels containing significant noise can be rejected manually or using automatic detection [14] after the scan.

Magnetic artefacts resulting from magnetic materials moving within the room such as clothing, cosmetics, dental or medical implants can have a significant impact on MEG recordings. An initial artefact scan should be performed to detect potential magnetic artefacts, which typically appear as large low frequency changes in the data and can be associated with patient motion or blinking. In some cases the source of the artefact can be eliminated after detection (e.g. by removing cosmetics), but this is not always possible. The strong magnetic fields encountered during MR imaging can further exacerbate these sources of noise, but demagnetization can reduce this effect [15]. Sensors which are significantly impacted by these magnetic artefacts can be removed from data analysis. Another method of compensation for artefactual signals is signal space separation (SSS), which decomposes magnetic field readings into the combination of two linearly independent expansions of harmonic functions corresponding to signals generated either inside or outside the sensor array.

Signals from fields generated inside the sensor helmet can be retained, while signals originating from artefactual sources outside can be removed from the data. While SSS cannot typically suppress artefacts originating close to the sensor helmet [16], temporal signal space separation (tSSS) is well-suited to remove these sources of noise from the sensor data [17]. In these cases, complex artefactual signals which may 'leak' into both parts of the SSS model can be detected by searching for similar temporal patterns in both the internal and external set of fields. Since neurological signals should be able to be represented by fields originating inside the helmet without any such leakage, the temporal approach allows the artefactual signals to be removed. Reconstruction of the signals from missing or problematic sensors and compensation for head movement can also be incorporated into the tSSS procedure [17].

Physiological Artefacts

Aside from neural activity, many other biological processes are capable of generating strong magnetic fields. For example, electrical activity in the heart can generate a magnetic field more than ten times larger than neuromagnetic signals of interest [18]; contractions in muscles close to the MEG sensor array can also generate electrophysiological activity strong enough to contaminate neuromagnetic signals. The eyeball, itself an electrical dipole, can also be the source of large artefacts in MEG readings. Blinks can cause current flow along the inner eyelids and change the geometry of the surrounding conductive volume, while vertical and horizontal eye movements affect the signal differently based on the motion of the eyeball within the conductive volume [19]. Artefacts resulting from cardiac activity cannot be suppressed, but can be removed after data acquisition is completed [16]. For this reason, electrocardiogram (ECG) measurements should be acquired during the scan as a reference for artefact rejection. Likewise, electro-oculogram (EOG) measurements should be acquired for the removal of eye movement/blink artefacts, as well as electromyogram (EMG) measurements when appropriate [14]. These physiological signals tend to be systemic, affecting multiple sensors, and are best removed from the MEG data using spatiotemporal independent component analysis (ICA). This approach decomposes a signal into separate, linearly-mixed sources by maximizing statistical independence of the signal's components. Once separated, artefactual sources may be discarded. Classification of sources as artefactual may rely on expecting physiological sources (typically much stronger than neurological sources) to exceed a set threshold, looking for correlations with recorded ECG, EOG, and EMG data, or manual inspection [13].

Patient Behaviour

Even after the previously described noise reduction and artefact removal techniques, the signal-to-noise ratio (SNR) of the MEG signal is still very low. This can be improved by collecting more trials, since the signal-to-noise ratio of an evoked response will benefit from averaging over a greater number of epochs. However, there is a potential tradeoff: while a longer scan could improve SNR and facilitate rejection of trials containing artefacts, it could also increase patient movement or noncompliance.

Head movement is another prominent data quality issue, particularly in patient populations who may find prolonged stillness demanding. Since the magnetic field detected is proportional to the inverse square of the distance from the source, subject motion during the scan changes the sensitivity of the sensors to different regions within the brain. Further, systemic head movement over the course of a scan can significantly decrease the accuracy of co-registration between anatomical and functional data, resulting in inaccurate source localization. While head movement corrections can be applied in post-processing, head position indicator coils should be used to monitor head movement during the scan and reposition the patient if necessary [20]. As previously mentioned, issues with patient motion can be exacerbated by lengthy scans, as patients grow tired or restless.

Patient task performance is unique among data quality issues in that it directly affects the underlying neural activity, and cannot be corrected for after the scan. Passive stimuli, which activate regions of the brain without requiring a patient to actively perform a task, may still require a patient to remain still or limit eye movement. However, active tasks require a patient to respond to a stimulus in a well-defined and consistent way across trials. Patient attention to stimuli and focus on task performance can directly affect the quality of the evoked response. Even with high engagement, patient task performance may suffer due to confusion over instructions or physical or cognitive impairments. During the scan, the operator should monitor participant behaviour, including motion and task performance. Depending on the mapping paradigm, different measures to monitor task performance may be suitable beyond basic monitoring via camera and intercom. Performance could be integrated within the task itself, such as real-time performance logging for a button response task. For motor tasks, EMG data can be used to monitor the amplitude and reliability of the desired muscle activity. For passive stimuli or resting scans, EOG measurements can serve as an early indication of participant drowsiness.

1.2.2 Conventional Quality Assurance

In general, quality assurance (QA) procedures for MEG data acquisition and analysis require significant operator oversight. While qualitative protocols to detect and compensate for artefacts at scan time were addressed in the previous section, this section will discuss measures of data quality.

Data analysis typically begins with a visual assessment of data quality [16]. Evoked responses can be visualized by averaging data across trials and inspecting sensor topographies at periods of strong magnetic field deflections. The presence of artefacts, which could appear as noise masking expected peaks or as unexpected patterns of topography, may guide the operator's data processing choices, leading to the use of one or more of the compensatory techniques discussed in the previous section. As data analysis continues, several quantitative indications of data quality can be calculated. Peaks in the global field power (GFP) corresponding to the timing of expected responses can serve as an early indication of good data quality. However, it is also typical to report data quality following source localization with an ECD model, using dipole fitting metrics such as goodness of fit or the confidence volume.

Goodness of Fit

The quality of source localization performed with an ECD model is quantified by its goodness of fit, which can be thought of as the percentage of sensor variance predicted by that model. This is maximized during the fitting process described in section 1.1.3. Mathematically, goodness of fit GoF may be calculated by

$$GoF = \left[1 - \frac{\sum_{i=1}^{n}(s_i - \hat{s}_i)^2}{\sum_{i=1}^{n} s_i^2}\right] \times 100\% \tag{1.3}$$

where s_i represents the measurement made by the ith of n channels and $\hat{s}_i$ represents the corresponding measurement predicted by the ECD forward solution. Fig. 1.2 provides an example of the difference between predicted and measured field measurements for a simple ECD model.

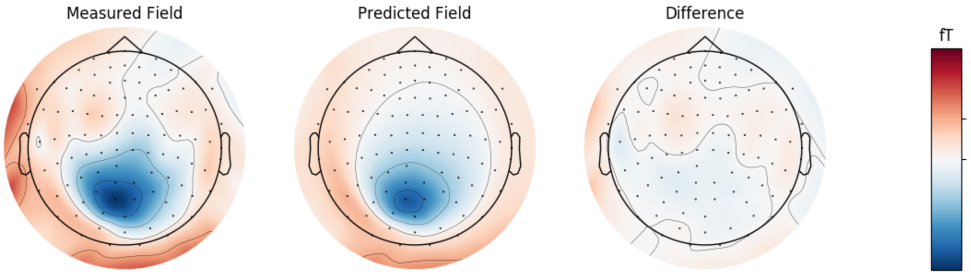


Figure 1.2: After localizing the right primary visual cortex with an ECD model, the difference between the measured and predicted field measurements can be quantified by goodness of fit. In this case, the dipolar source shown in Fig. 1.1 (a) accounts for 60% of sensor variance in the measured data.

For mapping paradigms that elicit only one or two focal dipolar topographies in the evoked field data, source activity can be well represented by one or two ECDs, so this metric acts as a useful surrogate for data quality. In this case, lower goodness of fit values indicate the presence of noise, unwanted sources of activation, or that the source of the activity does not fit the assumption of the ECD model (e.g. a spatially extended source). However, it is clear that this metric also depends on the appropriateness of the selected ECD model. The number of dipolar sources must be specified *a priori*, and allowed orientations or positions may also be defined by the MEG operator. Moreover, the goodness of fit (GoF) of a given ECD model can also depend on data acquisition parameters such as the number and type of channels (magnetometers vs. gradiometers), which can complicate the interpretation of the calculated goodness of fit value [16]. While most MEG reporting guidelines suggest reporting goodness of fit when an ECD model is used [14, 16], there is no universal threshold corresponding to ‘good’ data. Suggested goodness of fit criteria ranging from 70% to 90% have been reported in the literature [10, 21].

1.2.3 Intra-Session Reliability Measures

The previous sections have described the causes of poor data quality for task-based MEG scans and standard quality assurance practices to detect and address data quality issues. In large part, these standard quality assurance protocols are based on identifying well-defined effects of data quality issues which cause data to deviate from an expected result. This is largely performed by the operator, who is expected to recognize artefacts which may obscure the desired response and compensate accordingly during acquisition or processing. The main quantitative measure of data quality, dipole goodness of fit, is a further example of this approach to quality assurance. Goodness of fit is only an effective measure of data quality when an ECD model can be assumed to appropriately model data. Since it is based on assumptions about the number and properties of the underlying ECD sources, it relies on expert users and can be subject to inter-rater differences. In this section, I will describe a different approach for assessing data quality, which relies on fewer assumptions about the properties of 'good' MEG functional mapping data. This approach is rooted in a more general definition of data quality, expressed by two fundamental characteristics: accuracy and reliability.

In the context of source localization with MEG, accuracy is defined as the closeness of a measured source to the location of 'true' activation. Measuring accuracy therefore requires a comparison with a known ground truth. This poses a problem for non-invasive functional imaging techniques. The gold standard for the identification of brain function is direct electrical cortical stimulation, which is inherently invasive. While this rules out accuracy as a useful MEG quality metric, previous studies comparing MEG functional maps to results obtained during intra-operative electrical cortical stimulation have calculated MEG's accuracy to be on the order of millimeters for focal brain mapping in relevant clinical populations [3, 6, 22]. In contrast, reliability (sometimes characterized as repeatability or variability) is a more easily measured property of data. Inter-session reliability is a measure of the closeness of sources identified for the same paradigm in the same patient in different sessions. In general, repeated scans are not clinically feasible in the patient population, although at least one imaging centre has reported performing two replications of mapping paradigms in a single session for pediatric patients [23], with the overall effect of increasing the number of trials. Previous MEG studies have characterized an average inter-session reliability on the order of 3-8 mm for typical focal mapping paradigms [24, 25], though values up to 75 mm were measured for distributed sources relevant to language mapping [26]. It should be emphasized, however, that these estimates of accuracy and inter-session reliability were calculated in MEG data without any known quality issues, so they do not provide insight into what could be expected for a single scan of unknown quality. Given that most patients receive only a single scan, it would be useful to assess the reliability within the scan, rather than between multiple sessions.

In recent years, a novel approach has emerged to assess the intra-session reliability of task-based MEG data. A single scan, once segmented into epochs, can easily be divided into two split-half datasets. This technique was first used to estimate the reliability of independent components detected in EEG and MEG data [27], but can be used more generally to approximate two separate scans. The following sections describe three measures which quantify intra-session reliability using this split-half technique.

Receiver-Operator Characteristic Reliability Analysis: the Reliable Fraction

The receiver-operator characteristic reliability (ROC-r) framework was developed to evaluate the quality of fMRI activation maps and to provide an automated method of activation threshold selection by Stevens *et al.*, who later extended the ROC-r approach to MEG data [28, 29]. Since the framework is based on the comparison of two functional maps obtained from a test and retest dataset, a split-half approach was well-suited to single scan MEG data. The retest dataset is subjected to a receiver-operator characteristic (ROC) analysis, with the test dataset serving as a surrogate for true activation. In this pseudo-ROC framework, voxels active in both the test and retest maps are considered true positives while voxels active in only the test map are false negatives, allowing for the calculation of the true positive rate (TPR). Likewise, voxels inactive in both the test and retest maps are true negatives while voxels active in only the retest map are false positives, allowing for the calculation of the false positive rate (FPR). By calculating the sensitivity (TPR) and specificity

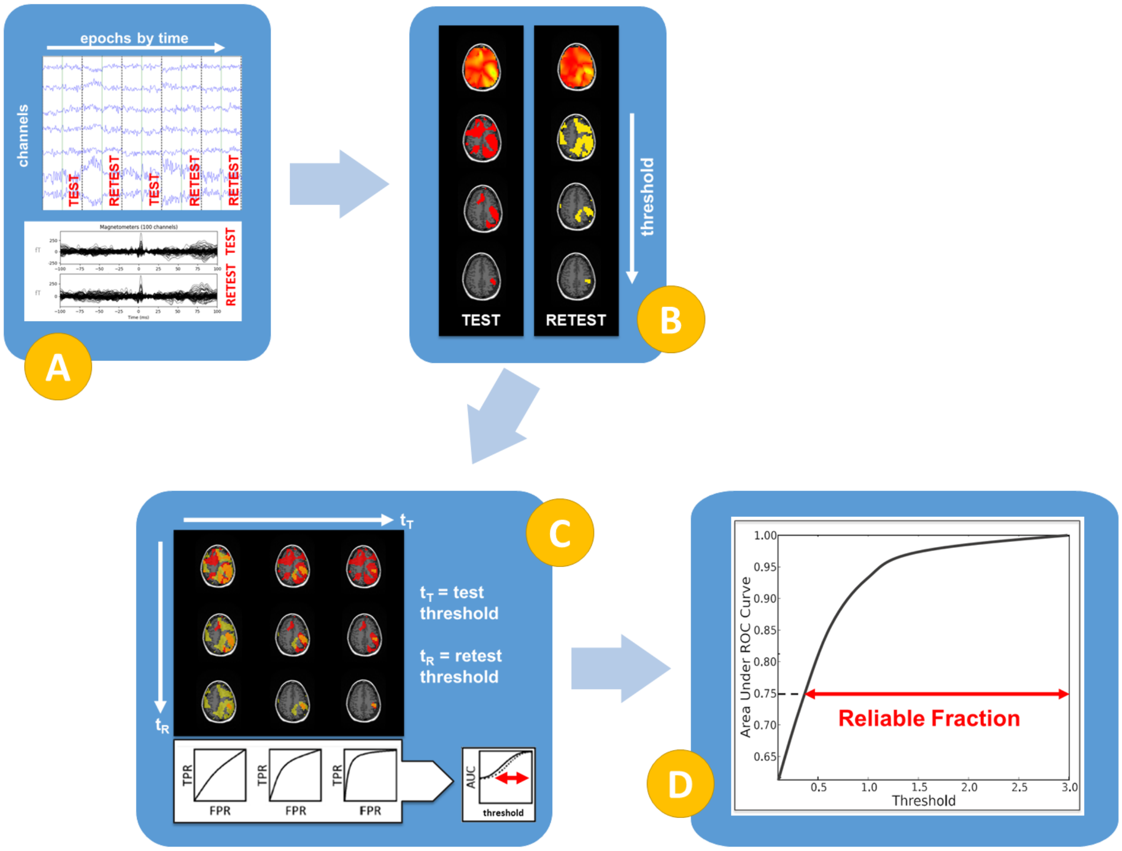


Figure 1.3: Overview of the receiver-operator characteristic reliability analysis framework:

- **(A)** After standard pre-processing, MEG data is segmented into epochs based on event timing. Epochs are randomly sorted into test and retest sets.
- **(B)** At a single time point, test and retest activation maps are created for equally spaced thresholds ranging from zero to the maximum of both maps.
- **(C)** Pseudo ROC curves are obtained from the overlap between test and retest maps. At a given threshold, the test map is considered to represent 'true' activation, while the classification sensitivity and specificity for the retest map is then calculated over all thresholds. This process is repeated for all test thresholds.
- **(D)** Area under each ROC curve (AUC) is plotted as a function of test threshold. The reliable fraction is defined as the fraction of test thresholds for which the AUC curve is greater than 0.75.

(1 - FPR) of a binary classifier corresponding to the retest dataset at a range of retest map thresholds, a pseudo-ROC curve is generated. The shape of this curve is highly dependent on the threshold chosen for the test dataset. As for traditional ROC analysis, a higher area under the ROC curve represents a better classifier, since there is less tradeoff between sensitivity and specificity. An area of one half would correspond to a completely arbitrary classification, while an area of one would correspond to a perfect classifier. In this case, the better the classifier, the closer the agreement between the test and retest datasets. If the test and retest datasets are highly similar, high-area ROC curves will be generated for a wide range of test dataset thresholds. This premise can be used to calculate a metric for intra-session reliability: the reliable fraction (Fr), defined as the fraction of test dataset thresholds for which the area under the ROC curve is greater than 0.75. Since the agreement between test and retest datasets is to some extent dependent on the original assignation of split-halves, the reliable fraction is typically averaged over a number of split-half divisions for greater stability. A schematic overview of this framework is presented in Fig. 1.3.

Stevens *et al.* validated the preliminary use of reliable fraction as a quality metric in a small MEG dataset mapping the somatosensory cortex in healthy subjects [29]. In particular, they found that reliable fraction correlated well across time with goodness of fit and proposed the use of reliable fraction as a promising quality metric for cases poorly suited to goodness of fit, such as distributed brain activity. A subsequent longitudinal MEG study examining the use of reliable fraction for automated processing pipeline selection found that choosing a processing pipeline which maximized the reliable fraction also minimized inter-session variability in localization [30]. Reliable fraction has emerged as a strong candidate for a broadly applicable MEG quality metric in these early studies, but it has not yet been validated across a wide range of pre-surgical mapping paradigms or in the case of known data quality issues.

Dice Similarity Coefficient

At its core, the receiver-operator characteristic reliability technique can be viewed as a framework for comparing two activation maps. It was created in part to address a significant shortcoming in overlap coefficients: the requirement for a specific threshold at which to make a comparison. Nonetheless, two overlap coefficients – the Dice and the Jaccard – have been frequently used by the functional neuroimaging community to assess agreement of test and retest activation maps. The two coefficients are closely related, although the Dice coefficient is more commonly used to assess reliability [31]. The Dice coefficient [32], first used to calculate fMRI test-retest reliability by Rombouts *et al.* [33], is defined as the number of active voxels which overlap for both datasets divided by the average number of active voxels across both datasets. It ranges from 0 (no agreement) to 1 (perfect agreement). The Dice coefficient has since been widely used to compare functional and anatomical images, from evaluating image segmentation quality [34, 35] to estimating the reliability of fMRI mapping paradigms [31, 36]. However, as for goodness of fit, there is no single established criterion for a 'good quality' overlap. Moreover the selection of varying thresholds at which to assess overlap has made meaningful comparison difficult across fMRI studies [31]. Although non-traditional in the fMRI community, it is possible to calculate the Dice similarity coefficient in its more general form to compare the agreement between two continuous vectors:

$$D = \frac{2|x \cdot y|}{|x|^2 + |y|^2} \tag{1.4}$$

In this thesis, I will test this threshold-independent version of the Dice similarity coefficient as a potential intra-session reliability metric.

Pearson Correlation Coefficient

The Pearson correlation coefficient, though rarely used to compare functional images directly, is one of the most popular statistics for measuring the association between two vectors. It can be interpreted as the covariance of two vectors divided by their standard deviations, and is calculated by

$$P = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{1.5}$$

Traditionally, fMRI studies have preferred the intra-class correlation coefficient [31], partially to allow the comparison of more than two images, but also because unlike intra-class correlation, Pearson correlation is mean-subtracted and measures correlation in the case of a linear relationship between vectors [37]. Since for the purposes of functional mapping, the locations of peaks within MEG beamformer maps are more relevant than the magnitude of estimated activity, this latter quality is advantageous for comparing MEG datasets for pre-surgical mapping. The Pearson correlation coefficient is thus a clear candidate for an intra-session reliability metric. Unlike the previously discussed metrics, the Pearson correlation coefficient ranges from -1 (perfect anti-correlation) to 1 (perfect correlation). However, I expect to calculate only positive values ranging from 0 (no correlation) to 1 for split-half data, because it is the extent (rather than the existence) of similarity between the two split-half maps that is in question.

1.2.4 Quality Assurance Measures: Summary and Context

Table 1.1 provides an overview of the four MEG quality metrics described in the preceding sections. Crucially, while goodness of fit (along with its ECD counterpart, confidence volume) is the only metric in clinical use, it is neither operator independent in the case of multiple sources or useful in the case of distributed brain activity. The purpose of this thesis is therefore to put forward reliable fraction as a novel, universally-applicable quality metric for task-based MEG data. Reliable fraction will be examined alongside the Dice and Pearson coefficients to provide context for other methods which may be suitable for judging similarity between split-half activation maps.

Metric	**Clinical standard**	**Intra-session reliability**		
	Goodness of fit	Reliable fraction	Dice coefficient	Pearson correlation
In clinical use?	Yes	No	No	No
Operator independent?	No	Yes	Yes	Yes
Works for distributed activity?	No	Yes	Yes	Yes

Table 1.1: A comparison of use cases and considerations for ECD goodness of fit and the three proposed intra-session reliability metrics.

1.3 Research Objectives

The primary objective of this thesis is to evaluate the performance of reliable fraction and other possible intra-session reliability metrics as a measure of data quality for pre-surgical maps acquired with MEG. To do so, I will compare the ability of goodness of fit, reliable fraction, and the Dice and Pearson coefficients to distinguish between 'good' and 'poor' MEG datasets. This thesis will particularly focus on the use of the reliable fraction and the ROC-r analysis framework for quality assurance in the case of focal sources in order to justify the future application of this approach to distributed sources. The Dice and Pearson coefficients will be tested alongside the reliable fraction as competing intra-session reliability measures, while goodness of fit will provide a reference as an established quality metric.

1.3.1 Data Quality Manipulation

MEG scans will be acquired using established pre-surgical mapping paradigms to localize the primary somatosensory, motor, visual, and auditory cortices in twelve healthy subjects. For each subject and paradigm, each MEG scan will be acquired twice: once in a standard 'good' condition, and once in a 'poor' condition where the data is manipulated by simulating quality issues such as those discussed in section 1.2.1.

I hypothesize that across all subjects and paradigms, all quality assurance metrics will measure higher values for scans acquired in the good condition than for those acquired in the poor condition at times when an evoked response is present in the good data. This hypothesis (hereon referred to as **Hypothesis 1**) relies on two separate assumptions:

1. The proposed manipulation of task performance and/or scan condition will significantly decrease the quality of source localization as compared to data acquired normally.

2. Both goodness of fit and the previously discussed intra-session reliability metrics will be sensitive to the expected changes in data quality.

1.3.2 Intra-Session Reliability Metric Performance

The most important component of this work will be to benchmark the performance of the three intra-session reliability metrics against the more established data quality measure: goodness of fit. This comparison is only possible in the case of focal sources where dipole fitting is an appropriate method of localization. This work therefore has the potential to inform the further application of intra-session reliability metrics in the case of distributed sources. As discussed in section 1.2.4, these metrics could fill a significant void in MEG quality assurance for complex source patterns. Thus, this thesis aims to assess the potential of the best-performing intra-session reliability metric as a universal MEG quality assurance measure.

Across all paradigms, I hypothesize that reliable fraction will measure the most significant separation between good and poor data (as compared to the three other metrics). This comparison will be performed on both a group level and for individual subjects. This hypothesis can therefore be separated into two complementary predictions (for each paradigm). **Hypothesis 2.1**: I predict that a significantly larger change in reliable fraction will be found between all good and poor datasets across the group, as compared to goodness of fit and the Dice and Pearson coefficients. **Hypothesis 2.2**: I predict that when examining the significance of the change in each metric at a single-subject level, reliable fraction will measure the largest difference between good and poor datasets across all subjects, as compared to the other metrics. These predictions are largely based on two assumptions:

1. Averaging intra-session reliability metrics across multiple split-half datasets will reduce variability in these metrics, resulting in a more significant difference between good and poor data than that measured with goodness of fit.

2. Reliable fraction will be more sensitive to changes in the higher activity regions of the data than the Dice or Pearson coefficients, corresponding to a greater separation between good and poor data.

Thus, this thesis will lay the groundwork for the implementation and validation of a universal MEG quality assurance measure based on intra-session reliability, with the eventual goal of ensuring high quality pre-surgical mapping data is acquired for each patient.

Chapter 2

Methods

2.1 Participants

This study was approved by the IWK Health Centre Research Ethics Board; all participants provided informed written consent. Eleven healthy adult participants underwent scans which were included in this study (6 female, ages 29.0 ± 10.3 years), including one participant who reported an ongoing mild psychiatric disorder during pre-screening which would not impact the responses of interest for this study. An additional two participants were involved in scans for this study, but one participant was found to be unsuitable at scan time due to the presence of a large dental implant, and one participant's data was excluded from all analysis post-collection due to an issue with equipment at scan time. Participants were each tested with the Edinburgh Handedness Inventory [38] and found to be right-handed for ten of the eleven participants and left-handed for one participant.

2.2 Experimental Paradigms

Data analyzed in this thesis were collected as part of a study examining the performance of the ROC-r analysis framework across somatosensory, motor, visual, and auditory pre-surgical mapping paradigms with manipulated data quality. Each participant underwent MEG mapping procedures based on clinical practice guidelines [10] to localize the primary somatosensory, motor, visual, and auditory cortices for the right side of the body, as well as a wakeful resting scan. Due to difficulties in consistently eliciting the desired motor and auditory responses across all subjects, this thesis compares the performance of metrics discussed in Chapter 1 for somatosensory and visual data only.

Data quality was manipulated differently for each mapping paradigm; the relevant details of data quality manipulation will be discussed in the following sections. For each paradigm, each participant was scanned once according to standard clinical procedures (from here on referred to as the 'good' condition) and scanned once following the same procedures with the addition of a specific action to deliberately decrease data quality (from here on referred to as the 'poor' condition). The twelve subjects scanned were divided into two groups of six. One group of six first underwent scans in the good condition followed by scans in the poor condition for each paradigm. For the second group, the order of the 'good' and 'poor' scans were reversed to mitigate the potential effects of repeated participant experience on data quality. The order of scans for each group is shown in Table 2.1.

Group 1	**Group 2**	**Scan Length**	**Number of Trials**
Somatosensory			
Good Condition	Poor Condition	1.5 minutes	400
Poor Condition	Good Condition	1.5 minutes	400
Motor			
Good Condition	Poor Condition	7 minutes	100
Poor Condition	Good Condition	7 minutes	100
Visual			
Good Condition	Poor Condition	1 minute	100
Poor Condition	Good Condition	1 minute	100
Auditory			
Good Condition	Poor Condition	3 minutes	100
Poor Condition	Good Condition	3 minutes	100
Rest		5 minutes	N/A

Table 2.1: Scan order and approximate duration for participants in study.

2.2.1 Somatosensory Paradigm

Localization of the primary somatosensory cortex (S1) was achieved by weak electrical stimulation of the participant's right median nerve. Using a DS7A Constant Current Stimulator (Digitimer, Hertfordshire, UK), a pulsed voltage was applied through two electrodes positioned on the surface of the participant's skin across the right median nerve. Prior to application, the electrodes were soaked in a saline solution to improve electrical conduction. Voltage, current, and the positioning of the two electrodes were adjusted to elicit a slight thumb twitch, a standard indication of successful median nerve stimulation (MNS). Stimulus timing was controlled by the Presentation software (Neurobehavioural Systems Inc., Berkeley, CA). The participant was instructed to look at a fixation cross presented in the participant's central field of view during the scan in order to limit eye and head movement. In total, 400 stimuli were presented at 217 ms intervals.

In the poor condition, the quality of somatosensory mapping data was manipulated by attaching a small magnetized piece of metal (1 cm section of paper clip) to the outside of each participant's lower left jaw with adhesive tape. This simulated a magnetic artefact similar to that which might result from a metallic dental implant, increasing noise levels in the recorded signal.

2.2.2 Motor Paradigm

Localization of the right primary motor cortex (M1) was achieved by visually-cued abduction of the participant's right index finger. The participant was instructed to look at a fixation cross presented in the participant's central field of view during the scan, and to quickly move their right index finger toward their thumb when the fixation cross turned yellow. The last three fingers of the participants hand were secured to each other with tape to restrict undesirable movement. The right index finger was attached to the right middle finger with elastic to facilitate a short movement during the abduction and quickly return the index finger to its default position. Electrodes were positioned across the right first dorsal interosseous muscle to monitor participant response. In total, 100 stimuli were presented at intervals ranging from 3.5-4.5 seconds.

In the poor condition, the quality of motor mapping data was manipulated by directing the participant to simulate non-compliance. The participant was verbally instructed to misperform the task by responding to the visual stimulus by selecting one of the following behaviours at random:

- No finger movement
- Delayed and/or prolonged abduction of the right index finger
- Abduction of the left index finger

- Abduction of both the right and left index fingers

As discussed in section 1.2.1, task non-compliance is a significant source of data quality loss in clinical populations, particularly for young patients or those with intellectual, physical, and/or neurological disabilities. The simulated poor task performance was expected to result in possible mapping of the left primary motor cortex and a more general loss of signal-to-noise in the evoked response.

2.2.3 Visual Paradigm

Localization of the right primary visual cortex (V1) was achieved using a hemifield checkerboard (twelve vertical and eighteen horizontal checks; 1.4 ± 0.1 cm side length per check) reversal pattern projected onto the right side of a viewing screen (at approximately 1 m distance from participant). To present a visual stimulus, the position of the black and white checks in the checkerboard was reversed at 500 ms intervals (i.e. checkerboard flickering at 2 Hz), with 100 stimuli presented in total. In the good condition, the participant was directed to look at a centrally positioned, stationary fixation cross during stimulus presentation to ensure maximal activation of the right visual field only.

To modulate data quality in the poor condition, the fixation cross moved around the viewing screen. After beginning in the centre of the viewing screen, the position of the fixation cross was shifted every five seconds in non-sequential order to one of seven different locations. These positions corresponded to the four corners of the viewing screen and the centres of the left, right, and bottom edges of the screen. Participants were instructed to follow the cross with their eyes, causing a disruption in the focal retinotopic mapping of the stimulus such as that which might occur due to a lack of participant engagement or confusion regarding the instructions. As discussed in section 1.2.1, we expected this manipulation to activate multiple regions of the primary visual cortex, resulting in a weaker, non-focal activation pattern in the evoked response. This extended activation would limit accurate localization of the visual response.

2.2.4 Auditory Paradigm

Localization of the primary auditory cortex (A1) was achieved by presenting a 500 ms pure 1000 Hz tone to the right ear at 80 dB SPL using EARTone 3A transducers connected to an audiometer. White noise was continually presented to the left ear at 55 db SPL. In the good condition, the participant was instructed to look at a fixation cross presented in the participant's central field of view during stimulus presentation to limit eye and head movement. In total, 100 stimuli were presented at one second intervals.

To modulate data quality in the poor condition, the fixation cross moved around the viewing screen to the same positions as in the poor condition for the visual paradigm, but at fifteen second intervals. Participants were instructed to follow the cross with their head, simulating head movement such as that which might occur due to participant restlessness or confusion regarding the instructions. We expected this manipulation to cause variability in the sensitivity of the sensors to the underlying activity over time, resulting in a non-focal activation pattern in the evoked response. This extended activation would limit accurate localization of the auditory response.

2.2.5 Rest Paradigm

Participants were instructed to remain still with their eyes open by focusing on a centrally-presented fixation cross while five minutes of resting-state data was acquired. This data was not used in this thesis.

2.3 Data Acquisition

Each participant in this study received an MEG scan and structural T1-weighted MRI scan. For all but two participants both scans were performed on the same day; the order in which the scans were acquired varied from subject to subject.

2.3.1 MRI Scan

Participants were scanned using a 3.0T GE MR750 system (GE Healthcare, Waukesha, WI) to obtain anatomical information for accurate source estimation and data overlay. For each participant, an initial three-plane localizer scan (25 s acquisition time) with ten slices (8 mm thickness) in each direction was acquired to plan subsequent scans. We then acquired a structural T1-weighted scan (7:07 min acquisition time) using a sagittal-oriented BRAVO sequence (TI=450 ms, TR=6.2 s, TE=2.3 ms, flip angle=12, NEX 2 with acceleration of 2 in phase direction) with 184 slices (1 mm isotropic resolution).

2.3.2 MEG Scan

Prior to participant preparation, an artefact scan was briefly acquired to ensure no artefactual signals would introduce noise to subsequent data collection. Following this, each participant underwent preparation for the MEG scan. To track head movement during the scan, four head position indicator (HPI) coils were attached to the head. Two were placed on the forehead near the hairline, and two on the mastoid behind the right and left ears. The location of these HPI coils were digitized using the Isotrak system (Polhemus, Colchester, VT). The nasion, the right and left preauricular points, and 200 additional points along the forehead, nose, and scalp were also digitized. Electrodes were positioned on both upper arms to record the electrocardiogram, and on the left collarbone to record a ground signal. Vertical eye movement was monitored with EOG electrodes above and below the left eye, while horizontal eye movement was monitored with EOG electrodes at the outer corner of both eyes. As previously discussed (section 2.2.2), EMG data was acquired during motor mapping to monitor subject task performance.

MEG data was acquired with a 306 channel MEG system (Elekta Neuromag Oy, Helsinki, Finland). MEG, electrophysiological data, and event markers corresponding to stimulus presentation were acquired continuously at 2500 Hz with an 833 Hz in-line lowpass filter and recorded for off-line analysis. Head movement was tracked continuously with HPI coils activated at frequencies greater than 250 Hz. As summarized in Table 2.1, each participant underwent a series of mapping paradigms followed by a 5 minute eyes-open wakeful resting scan. This was followed by a 5 minute empty room scan to account for environmental noise.

2.4 MRI Data Processing

The structural T1-weighted MRI data was reconstructed and segmented using the open source FreeSurfer image analysis package (Martinos Center for Biomedical Imaging, Charlestown, MA; [39, 40, 41, 42, 43, 44, 45, 46]). This data was co-registered with the head digitization data acquired prior to the MEG scan using the vendor-supplied MRILab graphical user interface (Elekta Neuromag Oy, Helsinki, Finland). In particular, the nasion and preauricular points were visually identified on the reconstructed anatomical image and used as landmarks for initial co-registration. The co-registration was then adjusted to best align the remaining digitization points along the surface of the head.

As discussed in the previous chapter, subject-specific anatomical data can be used to improve accuracy in source localization. To do so, using the FreeSurfer analysis package, a boundary element model (BEM) was created for each participant for accurate source estimation. Additionally, a volume source space consisting of approximately 12000 voxels on a 5 x 5 x 5 mm grid at which to calculate beamformer solutions was defined within the brain.

2.5 MEG Data Processing

The following section describes the MEG processing pipeline used for noise reduction and source localization for each mapping paradigm. In general, MEG data for each subject was processed according to accepted clinical guidelines [10] to create beamformer maps of activation for each paradigm and performance. With the exception of the inital tSSS, all analysis was conducted with MNE Python (version 0.16.2).

Parameter	Somatosensory	Motor	Visual	Auditory
High-pass filter (Hz)	None	1	1	1
Low-pass filter (Hz)	330	40	40	40
Power-line notch filter	60 Hz intervals	None	None	None
Epoch interval (ms)	(-100, 100)	(-600, 500)	(-200, 300)	(-200, 300)
Baseline correction interval (ms)	(7, 14)	(-500, -400)	(-50, 0)	(-100, -50)
Beamformer active interval (ms)	(0, 100)	(-200, 200)	(0, 200)	(0, 200)
Beamformer noise interval (ms)	(-100, 0)	(-600, -200)	(-200, 0)	(-200, 0)

Table 2.2: Parameters used for MEG data processing for each pre-surgical mapping paradigm. Intervals are defined with respect to stimulus presentation.

2.5.1 Initial Pre-Processing

Temporal signal space separation was applied to all MEG data in order to reduce environmental noise. This was carried out using the vendor-supplied software MaxFilter (version 2.2, Elekta Neuromag Oy, Helsinki, Finland). Additionally, frequency filtering was applied to restrict data for each paradigm to the frequency bands of interest. Filter parameters are detailed in Table 2.2. Data were down-sampled to 1000 Hz to reduce processing time.

2.5.2 Data Epoching

In this study, stimulus timing was controlled by the Presentation software (Neurobehavioural Systems Inc., Berkeley, CA); however, the projection system exhibited a delayed response to the electronic stimulus timing reported by the Presentation software. A correction for visually-cued stimulus timing was therefore found by measuring the voltage induced in a photodiode attached to a checkerboard square projected onto the participant viewing screen. Prior to participant scans, a test of the visual paradigm design found a delay of 34.2 ± 0.3 ms between the actual stimulus times and the times reported by the Presentation software. This delay was therefore accounted for by adding a constant offset of 34 ms to the reported stimulus time. Likewise, a delay in median nerve stimulation of 2 ms was accounted for by adding an offset to the reported time. Stimulus timing for motor data was determined using the EMG channel to find the onset of muscle activity corresponding to task performance in the good condition. For all paradigms, the recorded or calculated stimulus timing was used to segment the data into epochs of time surrounding each trial, using the epoch interval times shown in Table 2.2.

2.5.3 Data Averaging

For this study, we used the fastICA algorithm [47] implemented in MNE-Python to remove spatiotemporal patterns in the MEG data likely to be artefacts. Independent sources were excluded from analysis if the magnitude of the component exceeded a threshold of 4 pT for magnetometers or 400 pT/cm for gradiometers, or if the time course of the component was highly correlated with the recorded EOG or ECG signals. The MEG data were reconstructed without the excluded components to generate 'clean' MEG epoch data. The epochs were then averaged over all trials to generate the MEG event-related field (ERF) data. A baseline correction using paradigm-specific intervals (see Table 2.2) was applied to both the epochs and evoked (inter-trial average) data. Notably, an interval ranging from 7-14 ms post-stimulus was chosen for the somatosensory data. While it is more typical to choose a pre-stimulus interval for the baseline correction, we chose a relatively short post-stimulus interval to correct for changes in the MEG sensors induced by the MNS electrical stimulation. This has been shown to reduce distortion of the N20m evoked field caused by the artefact during electrical stimulation [48], and we found it was more effective than a pre-stimulus baseline interval at uncovering the expected N20m peak across subjects.

2.5.4 Source Localization

A distributed source solution was calculated for each dataset using the linearly-constrained minimum variance (LCMV) beamformer implemented in MNE-Python [49]. Covariance estimates were calculated from the clean epoched data for the active interval and noise interval relevant to each paradigm (see Table 2.2). With the BEM model obtained from the MRI data, we then calculated the LCMV beamformer spatial filter, as discussed in the previous chapter. The beamformer spatial filter was then applied to the downsampled ERF data to create an event-related beamformer (ERB) map with estimates of source activity at each point in the volume source space for each time sample in the ERF data. The absolute value of estimated activity was taken to eliminate ambiguity in the dipole orientation and moment.

2.6 Calculating Data Quality Metrics

This section describes the process by which each data quality metric (goodness of fit, reliable fraction, and the Dice and Pearson coefficients) were obtained for each subject, paradigm (auditory data excepted), condition, and time point. As discussed in detail in the following chapter (see in particular sections 4.2.2 and 4.2.4), the quality of motor and auditory data acquired in the good condition for this study was deemed insufficient for a fruitful comparison of metric performance. In the case of motor data, metric time series were calculated for a preliminary examination of data quality. For auditory data, no quality metrics were calculated.

2.6.1 Goodness of Fit

To calculate goodness of fit, each dataset was fit with an equivalent current dipole model. MNE-Python's *mne.fit_dipole()* was used to fit a single dipole at each time point in the active interval, using the same boundary element model and covariance estimates as used for the beamformer solution. For each participant, paradigm, and condition, the dipole position, orientation, magnitude, and goodness of fit was therefore obtained a function of time post-stimulus.

2.6.2 Intra-Session Reliablity Metrics

Each of the proposed intra-session reliability metrics (reliable fraction, Dice and Pearson coefficients) are calculated by comparing the similarity of two activity maps. For a single task-based MEG scan, these two maps may be generated by randomly dividing the collected epochs into two split-half datasets (the 'test' and 'retest' datasets). Since the value of the calculated metric generally depends on the initial assignation of epochs into the test and retest datasets, the metric calculation is repeated and averaged across a number of randomly assigned split-half datasets. After a preliminary investigation, it was determined that in general averaging each metric over 30 pairs of split-half datasets was sufficient for metric convergence; this number was used for the calculation of metrics for comparison at the group level as described in section 2.7.1. A different approach was used to select a number of split-half datasets for the bootstrap investigation at the single subject level as described in section 2.7.2. The following section outlines the process by which split-half datasets were generated and used to calculate the reliable fraction, Dice, and Pearson coefficients for each time point in the active interval for each participant, paradigm, and condition. This analysis was implemented in Python 2.7.15.

Having processed the MEG data as described in sections 2.5.1-2.5.3, each set of epochs was randomly divided into two split-half datasets, which were averaged to create two evoked datasets. The appropriate baseline for each paradigm was then applied to the split-half evoked datasets and the covariance computed using the split-half epochs as described in sections 2.5.3-2.5.4. Likewise, a distributed source solution was calculated for each split-half dataset using the LCMV beamformer. For each time point within the active baseline interval, there then existed two split-half event-related beamformer maps describing the estimated source activity in each voxel in the volume source space as a function of time.

The ROC-r analysis framework, as discussed in the previous chapter, assesses the internal reliability of a dataset by calculating the reliable fraction. This metric is calculated by first thresholding the ERB maps to classify each voxel as active or inactive. For each pair of split-half datasets, twenty equally-spaced thresholds were chosen ranging from zero to the maximum activity present in each dataset. One of the two split-half ERB maps was then randomly defined to be the test dataset, while the other was designated the retest dataset. As described in Fig. 1.3, these two datasets and thresholds were used to generate a set of twenty receiver-operator characteristic curves, each consisting of twenty data points. The area under each such curve (AUC) was estimated using rectangular integration. Thus, for each set of split-half datasets, a set of twenty AUC values varying with the test dataset threshold was plotted as an AUC vs. threshold curve. As previously defined (Fig. 1.3), the reliable fraction is the proportion of thresholds for which the AUC is greater than 0.75. This value was calculated by interpolating the AUC vs. threshold curve to determine the threshold at which the AUC initially surpassed 0.75 for each pair of split-half datasets.

It is notable that in past works using the ROC-r analysis framework [29, 30], after an AUC vs. threshold curve had been created, the assignation of the test and retest datasets was reversed and the process then repeated to create a second AUC vs. threshold curve. The two curves were then averaged, giving a single AUC vs. threshold curve for each randomized set of split-half datasets. This thesis takes a slightly different approach. Since the initial assignment of epochs to the test and retest datasets is random and reliability results are averaged across multiple split-half datasets, it was judged unnecessary to repeat the ROC-r analysis with the test and retest assignations reversed. This halves the computational time necessary for ROC-r analysis without changing the reported reliability as long as results are averaged across a sufficient number of split-half datasets (i.e. until the average reliable fraction converges across datasets).

The Dice and Pearson coefficients were more straightforward to calculate, and were obtained at each time point by flattening each three-dimensional ERB map into a one-dimensional vector. The similarity of the two vectors was then calculated with Eq. 1.4 to calculate the Dice coefficent and Eq. 1.5 to calculate the Pearson coefficent.

At the end of the described analysis, the value of the goodness of fit, reliable fraction, Dice coefficient, and Pearson coefficient had been calculated at each time point for MEG mapping data corresponding to each subject, paradigm, and condition.

2.7 Statistical Analysis

2.7.1 Group Level Comparison

After obtaining goodness of fit vs. time and intra-session reliability metric vs. time curves (hereafter generalized as quality metric time curves) for each subject, paradigm, and condition, results were averaged across subjects for qualitative comparison. However, in general, the value of each quality metric vs. time is of less clinical utility than the value of each quality metric at the expected response peak. That is, high data quality should be measured when a focal source is evoked. This is true for both goodness of fit, which is high when a strong dipolar source is present, and for intra-session reliability metrics, as such a source should present a highly reliable activation pattern across epochs. Thus, to compare each metric's ability to distinguish between good and poor data quality, I conducted a paired t-test to evaluate the difference in each metric between the good and poor conditions at the response peak. This was done separately for both V1 and S1 data.

The general latency of the response peak was determined from the post-stimulus event-related field data for each subject in the good condition. The average source pattern was visually inspected for the presence of a strong dipolar source in short time windows (5-10 ms) with significant ERF activity, particularly those corresponding to known response peaks (N20m and N75m for the somatosensory and visual responses, respectively). The precise latency assigned to the response peak was then determined from the corresponding peak in global field power (GFP), a metric derived from the spatial standard deviation across sensors at each time point, reflecting the overall magnetic field activity at that time. When multiple dipolar sources were present in the ERF data, I chose the response peak occurring closer to the N20m and N75m responses which would be used clinically, if present. Latencies found for each subject and paradigm in the good condition are presented in Table 3.2.

For each subject and paradigm, I extracted the maximum values of each metric (in both the good and poor condition) within a window centred at the pre-determined latency. The width of the window was chosen to be approximately 25% of the full width at half maximum of the GFP peak across all subjects; within these windows the largest 3 and 8 metric values were extracted for the somatosensory and visual time curves, respectively. The average peak value for each metric and each condition were plotted for all subjects for each paradigm. For each metric, the difference between the average peak value of the metric in the poor condition and in the good condition was then used for a paired t-test across subjects. This was conducted with the null hypothesis that any observable differences in the metric between 'good' and 'poor' data were the result of random variation and the alternative hypothesis that the metric value in the good condition was significantly higher than the metric value in the poor condition across subjects.

2.7.2 Single Subject Comparison

To determine the statistical power of each metric to distinguish between good and poor data at the single subject level, I implemented bootstrapping to estimate the stability of each metric. A similar use of the nonparametric bootstrap was used by Darvas *et al.* to examine variability in equivalent current dipole localization [50]. Since task-related MEG data is composed of a number of epochs which are then averaged to create a single dataset, it is straightforward to use random resampling with replacement to generate bootstrap datasets. For any measure calculated on the original dataset, a bootstrap distribution can be generated by repeatedly calculating the measure for a number of bootstrap datasets, providing an estimate of the measure's variability [51].

Since this process is time consuming, I chose a single time point at which to perform bootstrapping for each metric. This was selected as the time at which each metric reached its greatest value (in the original metric vs. time series) in the good condition within the response peak window defined in the previous section. This time point, while potentially varying from metric to metric, was chosen in order to compare metrics at a point of optimal separation between good and poor data. The selected time points are provided in Table 3.5.

In practice, bootstrapping is often performed until the distribution of interest converges [52]. After an initial investigation into metric convergence across bootstrap datasets (see Fig. 3.12 (a)) indicated that variance in some metric distributions could increase indefinitely as bootstrapping continued, analysis was limited to 120 bootstrap datasets for each subject in the interest of time.

For each bootstrap dataset, dipole fitting was performed and split-half datasets were generated as described in section 2.6. Intra-session reliability metrics were calculated as previously described, with one notable change. Instead of averaging each metric across 30 split-half datasets, each metric was calculated for as many split-half datasets produced a significant improvement in metric convergence, with a minimum of 12 split-half datasets. After each metric had been calculated for a split-half dataset, the variance in that metric's value was calculated across all split-half datasets thus far. The mean variance in that metric was then calculated across the most recent three split-half datasets and compared to the mean variance across the fourth, fifth, and sixth most recent split-half datasets. This provided a measure of the change in variance in two windows of three split-half datasets. Calculations across split-half datasets continued for that metric until variance decreased by less than 50% in two such adjacent windows. The number of split-half datasets required for convergence for each intra-session reliability metric was recorded for each bootstrap dataset. Once each metric had converged, its average value was calculated over all split-half datasets for a single bootstrap dataset.

Finally, for each subject, after goodness of fit had been calculated and each intra-session reliability metric had converged for each of 120 bootstrap datasets, the difference between good and poor data was calculated for each metric using a one-sided Student's t-test. This was conducted with the null hypothesis that any observable differences in the metric between 'good' and 'poor' data were the result of random variation and the alternative hypothesis that the metric value in the good condition was significantly higher than the metric value in the poor condition. The performance of each metric across subjects was then compared using a one-way repeated measures analysis of variance (ANOVA). Minitab (Minitab 19 Statistical Software, Minitab Inc., USA) was used to perform the one-way ANOVA; a Ryan-Jones normality test and Levene's test for homogeneity of variances were first performed to ensure the appropriateness of one-way ANOVA for this dataset.

Chapter 3

Results

3.1 Overview of Pre-Surgical Mapping Data Acquired

Prior to a comparison of quality metric performance across good and poor data, this section qualitatively examines the effect of data quality manipulation on the acquired data.

3.1.1 Evoked Fields

Figure 3.1 shows event-related field magnetometer data in grand average for all paradigms in both conditions. Global field power and topographies of interest selected based on peaks in the GFP are shown for each paradigm and condition.

Somatosensory event-related fields in the good and poor conditions exhibit large deflections in the magnetic field both at the time of the stimulus itself and at later peaks approximately 20, 35, 50, and 70 ms post-stimulus. The peak at stimulus onset is an artefact caused by the electrical stimulus used to evoke a somatosensory response. Subsequent peaks correspond to the N20m peak, commonly used to localize the primary somatosensory cortex [6], and later peaks also localizing to the contralateral somatosensory cortex [24]. These peaks are present in both good and poor somatosensory data, despite the presence of noise in poor data resulting from the magnetic artefact.

Motor event-related fields in the good condition indicate a peak deflection at approximately 60 ms post-stimulus. The sensor topography at this peak appears to show a dipolar source in the left frontal region, but the spreading components of this source could correspond to more than one active source contributing to this pattern. A later peak at approximately 180 ms in the good condition shows an even more diffuse dipolar pattern in the same sensor region, likely corresponding to a number of sources active in the motor cortex. Neither peak is present in the GFP for the poor condition, although the sensor topography at 60 ms post-stimulus instead shows a weak dipolar source in the central occipital region, likely the response to the visual cue for task performance. A similar occipital response can be seen at approximately 50 ms pre-stimulus in both good and poor conditions.

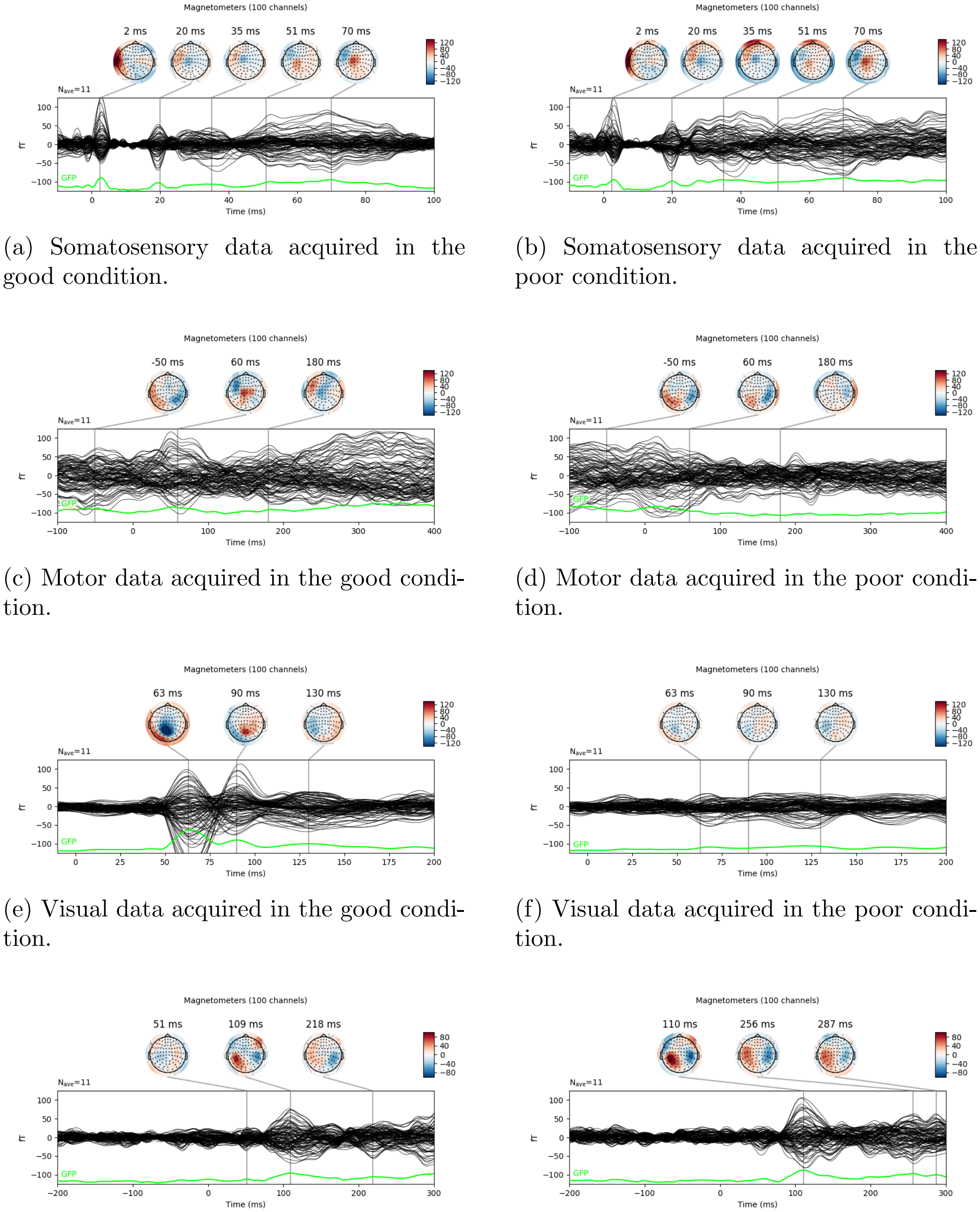
Visual event-related fields in the good condition indicate a prominent deflection in the magnetic field at approximately 63 ms post-stimulus. This latency corresponds to the N75m response peak, known to localize to the primary visual cortex on the contralateral side of a hemi-field checkerboard reversal stimulus [53]. At the N75m peak, the sensor topography shows a dipolar pattern in the sensors located near the left occipital region of the brain, which is expected for a checkerboard reversal presented to the right field of view. Two subsequent peaks at 90 and 130 ms post-stimulus can also be seen. The former peak, which has a sensor topography likewise localizing to the right occipital sensors, possibly corresponds to the P100m visual physiological response. The latter peak shows diffuse activity of unclear origin. In the poor condition, only small deflections are present at these peak latencies, demonstrating a significant inhibition of the visual response as the result of the data quality modulation.

Auditory event-related fields in both the good and poor conditions show a peak at approximately 110 ms post-stimulus. The topography at this latency appears to be the result of two dipolar sources in the temporal regions, which would correspond to the expected bilateral response localizing the primary auditory cortex [54]. This response does not appear significantly different in the poor condition, indicating that head movement does not necessarily modulate data quality at the sensor level.

3.1.2 Event-Related Beamformer Maps

Fig. 3.2 shows event-related activity localized with a beamformer for all paradigms and conditions in grand average, at the response peaks present in the evoked data (Fig. 3.1).

For the somatosensory paradigm, the response at 50 ms post-stimulus localized to similar regions of the left parietal lobe for both good and poor data, as expected for right median nerve stimulation [24]. This is the usual location of the right primary somatosensory cortex in healthy subjects. Importantly, data quality modulation



(a) Somatosensory data acquired in the good condition.

(b) Somatosensory data acquired in the poor condition.

(c) Motor data acquired in the good condition.

(d) Motor data acquired in the poor condition.

(e) Visual data acquired in the good condition.

(f) Visual data acquired in the poor condition.

(g) Auditory data acquired in the good condition.

(h) Auditory data acquired in the poor condition.

Figure 3.1: Event-related field measurements (magnetometers only) in grand average, with global field power (green) and relevant sensor topographies.

manifesting in noise in the poor ERF data did not prevent the localization of this response in the event-related beamformer (ERB) map.

For the motor paradigm, data in the good condition was localized to the primary motor cortex at approximately 180 ms post-stimulus, as expected from the ERF data. However, as predicted by the sensor topography at this latency, the activity in the motor ERB map is widely distributed. This suggests that even in the good condition, the motor mapping paradigm may not have achieved focal activation of the primary motor cortex, but may have instead activated several sources in the surrounding regions. In the poor condition, no corresponding peak in activation near the right posterior frontal lobe can be seen, indicating unsucessful localization of the primary motor cortex.

For the visual paradigm, data in the good condition at the N75m response peak (shown at 62 ms post-stimulus in Fig 3.2) localized to a focal peak in the right occipital cortex, suggesting successful localization of the primary visual cortex. At the same latency in the poor condition, no strong activation was observed, suggesting that the data quality modulation successfully inhibited localization of the primary visual cortex.

For the auditory paradigm, similar localization was achieved for both good and poor data at 107 ms post-stimulus. Activation was most prominent in the left hemisphere of the brain, while the bilateral response seen in the sensor data (Fig. 3.1) was not present. This is likely an artefact introduced by the beamformer spatial filter, which attenuates strongly temporally correlated sources [55]. This suppression of activation in the right hemisphere was present for both good and poor data. It is difficult to judge whether the head movement introduced during data acquisition significantly affected localization in the left hemisphere; the location of the ERB peak in the poor condition appears to be slightly more lateral than in the good condition.

3.1.3 Data Quality Manipulation Outcomes

To provide additional clarity on the quality of data obtained for each subject, paradigm, and performance, event-related field data was examined for each individual subject. Specifically, the average sensor topography was assessed for the presence of strong dipolar sources near the expected brain regions at peaks in the GFP for each subject. Table 3.1 indicates whether the expected response was achieved in each case. Although results in grand average (see previous section) indicate successful data

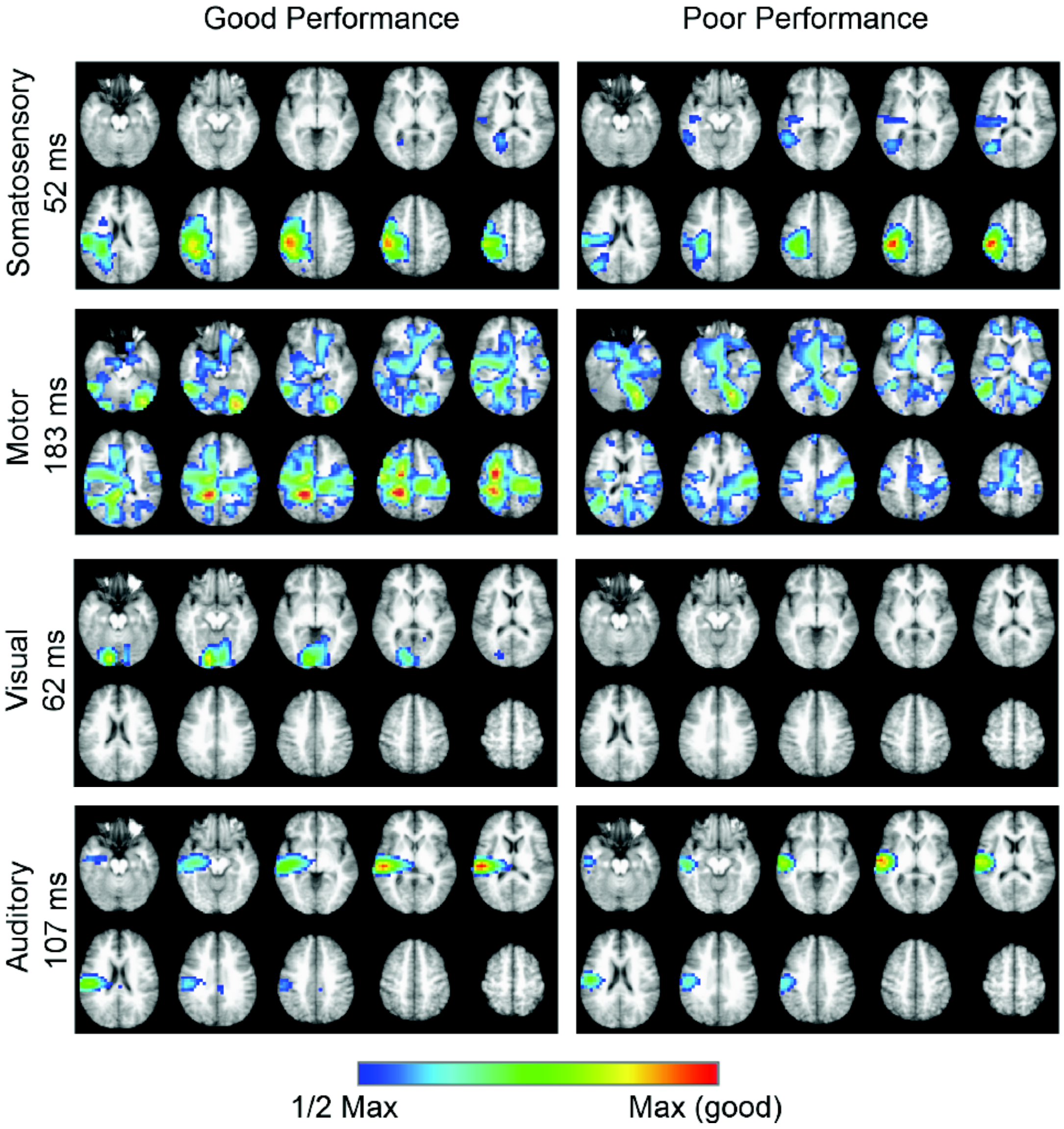


Figure 3.2: Event-related beamformer maps in grand average, thresholded to half the maximum in the good condition for each paradigm.

quality modulation for somatosensory, visual, and motor data, significant variability in successfully eliciting the desired response existed between paradigms. In the good condition, a somatosensory response was present in seven out of eleven subjects, each of whom also exhibited an observable response in the poor condition. For the motor paradigm, eight out of eleven subjects exhibited a motor response, while no such response was present in the poor condition. In the visual data, all subjects exhibited a visual response in the good condition, while only one subject also exhibited a visual response in the poor condition. The auditory data was of notably poor quality: only three subjects exhibited an auditory response in the good condition, all of whom also exhibited a response in the poor condition.

	Somatosensory		**Motor**		**Visual**		**Auditory**	
Subject	Good	Poor	Good	Poor	Good	Poor	Good	Poor
sub01	Y	Y	Y	N	Y	N	N	Y
sub02	Y	Y	-	-	Y	N	N	N
sub03	Y	Y	Y	N	Y	N	Y	Y
sub04	Y	Y	Y	N	Y	N	Y	Y
sub05	-	N	-	-	Y	Y	N	N
sub06	Y	Y	Y	N	Y	N	N	N
sub07	N	N	Y	N	Y	N	-	-
sub08	Y	Y	Y	N	Y	N	Y	Y
sub09	N	N	Y	N	Y	N	N	N
sub10	N	N	Y	N	Y	N	N	N
sub11	Y	Y	-	-	Y	N	N	N

Table 3.1: Indication of whether expected response was present (as evaluated by visual inspection) in sensor-level data for each paradigm, subject, and performance. Y: response present, N: response absent, -: results unclear.

In order to compare metric performance, somatosensory and visual data were selected for further analysis based on the relative homogeneity of subject response. After initial calculation of quality metrics as a function of time, motor data was excluded from statistical comparison of metric performance because although on average the expected response was observed in good data, it was difficult to identify distinct peaks corresponding to an evoked response at single time points (see Fig. A.1 and further discussion in section 4.2.2). Auditory data was excluded from metric calculation entirely due to the small number of subjects in which the desired response was evoked.

3.2 Group Quality Metric Comparison

3.2.1 Quality Metric Time Series

Metric time courses across the group for somatosensory, motor, and visual data are presented in Figs. 3.3, 3.4, and 3.5, respectively. Single subject curves are provided in Supplementary Materials (Appendix A, Figs. A.3 and A.4. In both somatosensory and visual paradigms, intra-session reliability metrics show a strong correlation with goodness of fit. For these datasets, all four metric time curves reach their lowest values during the baseline interval, and peak at times corresponding to large field deflections in the evoked data (Fig. 3.1). For the somatosensory paradigm in both the good and poor condition, the largest and most consistent peak across metrics corresponds to the N20m response, although a less prominent peak near 35 ms can also be seen. In the good visual data, three prominent metric peaks can be observed at approximately 63 ms, 90 ms, and 110-115 ms, which agree well with peaks in the good visual ERF data. The first peak corresponds to the N75m response, also seen in the ERF data at this same latency. The second peak also occurs at the same time in the ERF data, while the third metric peak occurs earlier than its ostensible GFP counterpart at 130 ms. This peak could correspond to the positive magnetic field deflections reaching a local maximum at approximately 112 ms in the visual ERF data. Later shallow peaks observed in the GFP are not clearly seen in the metric time courses, perhaps due to less consistent timing of responses across individual subjects. Moreover, the Dice and Pearson coefficients exhibit notably higher baseline values relative to goodness of fit and reliable fraction. For motor data, while the intra-session reliability metrics for the good condition reach a peak at approximately 260 ms post-stimulus, goodness of fit remains low. This likely indicates the presence of extended regions of activity which, while somewhat reliable (peaking at a lower reliability score than achieved by either somatosensory or visual data), is poorly modelled by a single ECD. In the somatosensory data, an early peak corresponding to a reliable artefact from the electrical stimulus applied across the participant's median nerve can be seen for all three intra-session reliability metrics. There is no clear separation between somatosensory data acquired in the good and poor conditions, as expected from the group beamformer activity maps (Fig. 3.2), further indicating that the quality of activation maps was not significantly changed by the introduction of a metallic artefact. This contradicts Hypothesis 1 (a measurable change in data quality in all metrics) for somatosensory data. In contrast, metric time curves corresponding to good and poor visual data show a significant separation between conditions for all four metrics. This is a further confirmation of successful data quality modulation (Hypothesis 1) for visual data.

3.2.2 Quality Metric Values at Response Peaks

Latencies found for the response peak for each subject and paradigm in the good condition are presented in Table 3.2. The distribution of metric values at these response peaks for each subject is shown in Fig. 3.6 and Fig. 3.7 for somatosensory and visual data, respectively. For both mapping paradigms, a high degree of variability in data quality metrics can be seen across subjects. This is particularly true for visual data in the good condition, where the clinical standard metric, goodness of fit, ranges from approximately 20-70% between subjects. The results of a statistical comparison of metric values calculated for the good and poor condition at the N20m somatosensory response peak and the N75m visual response peak are given in Table 3.3 and Table 3.4, respectively. For somatosensory data, no significant increase in any data quality metric was found for data acquired in the good condition relative to data acquired in the poor condition. For visual data, all metrics measured a significant increase in data quality for data acquired in the good condition relative to the poor condition. Out of all four metrics, goodness of fit measured the largest statistical increase in data quality across the group ($p = 5.2 \times 10^{-5}$), followed by reliable fraction ($p = 1.4 \times 10^{-4}$), the Pearson correlation ($p = 1.0 \times 10^{-3}$), and the Dice coefficient ($p = 1.3 \times 10^{-3}$). This is only a partial confirmation of Hypothesis 2.1: although reliable fraction outperformed the other two intra-session reliability metrics, goodness of fit remained the most sensitive to changes in data quality across the group.

3.2.3 Intra-Session Reliability Metric Convergence

Running the ROC-r analysis in a preliminary test of convergence demonstrated that the reliable fraction reached a variance of less than 0.001 after averaging across 30

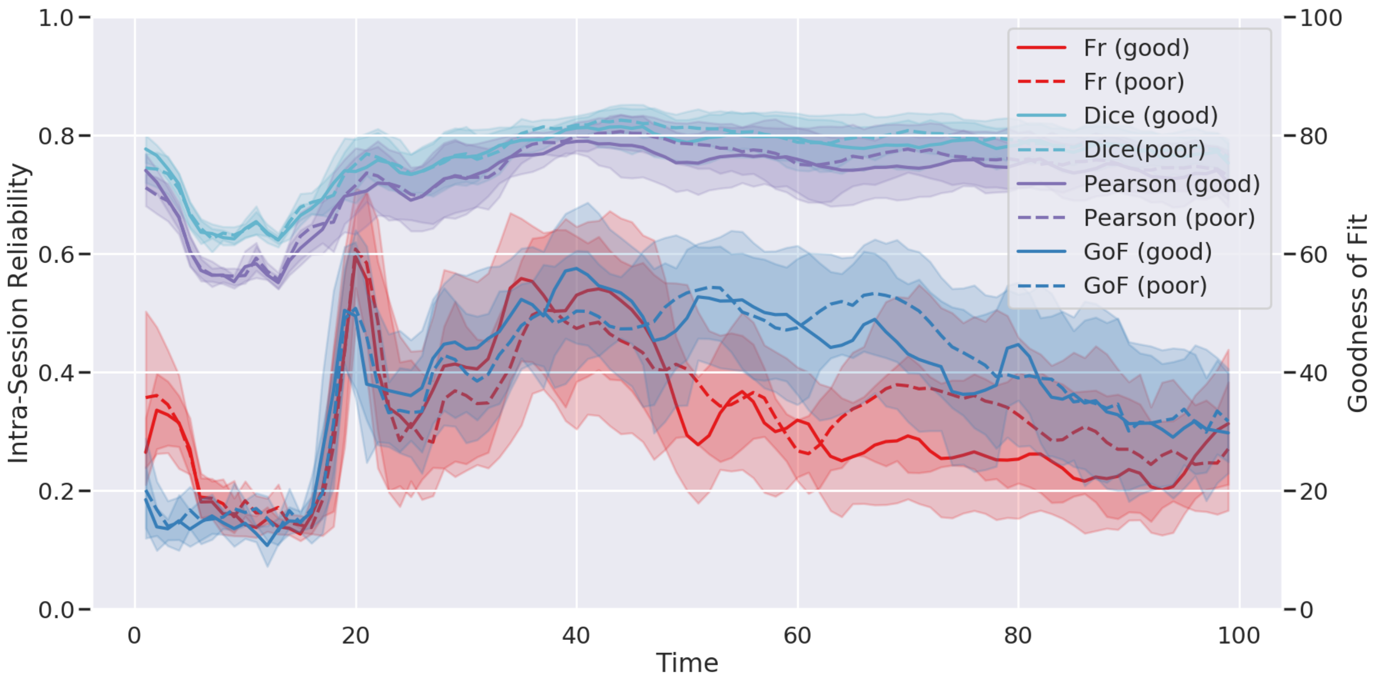


Figure 3.3: Quality metric time courses in grand average peak at N20m response for somatosensory data.

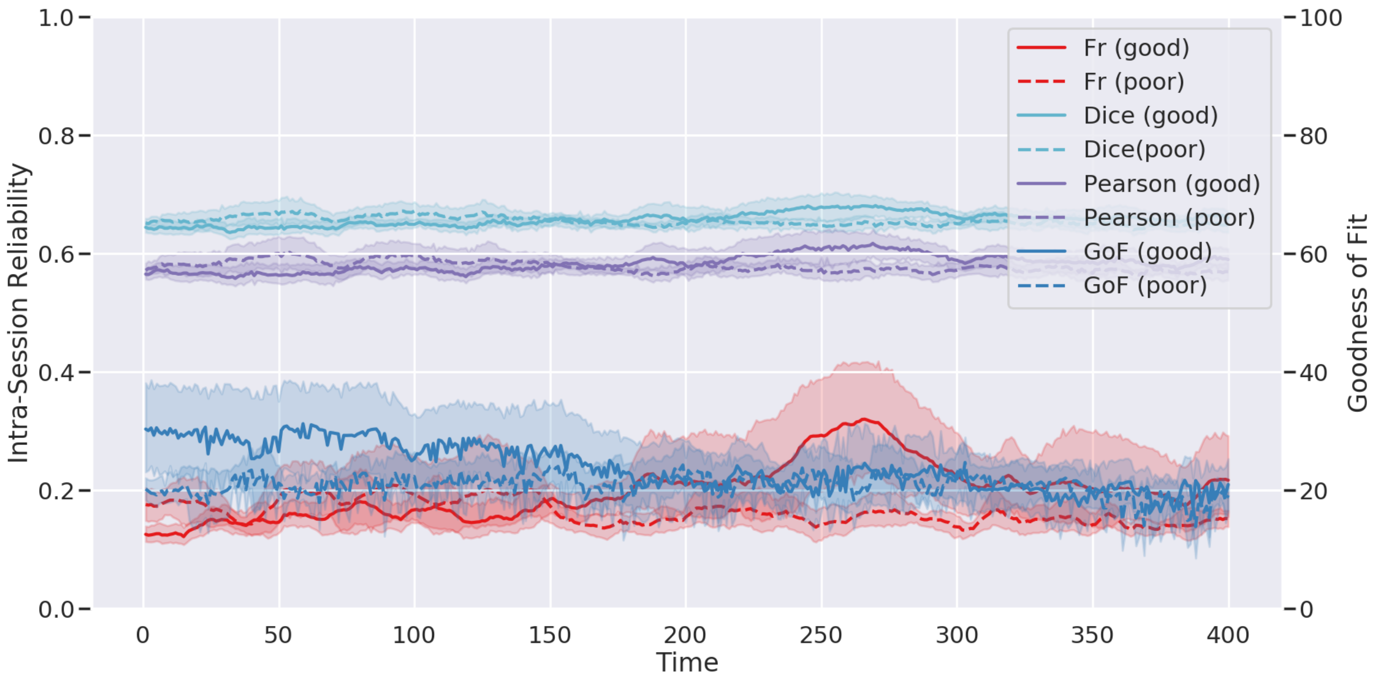


Figure 3.4: Quality metric time courses in grand average for motor data.

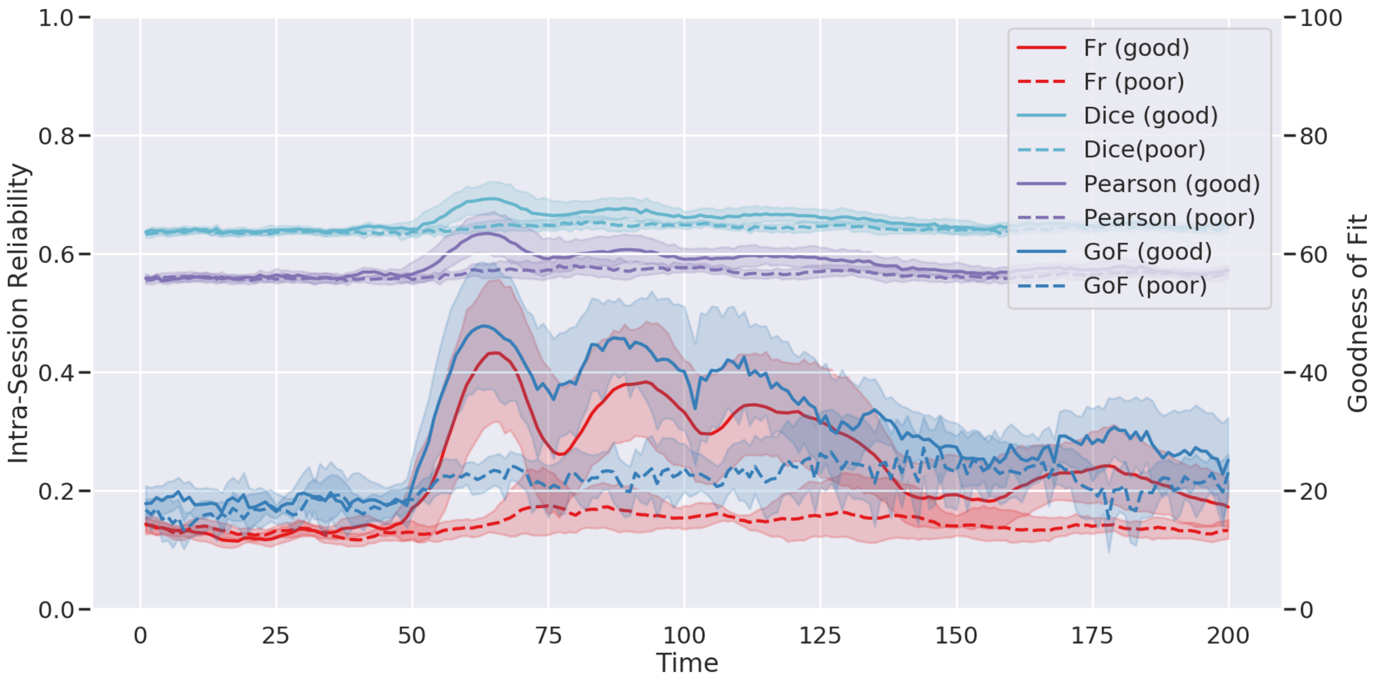


Figure 3.5: Quality metric time courses in grand average peak at N75m response for visual data acquired in the good condition.

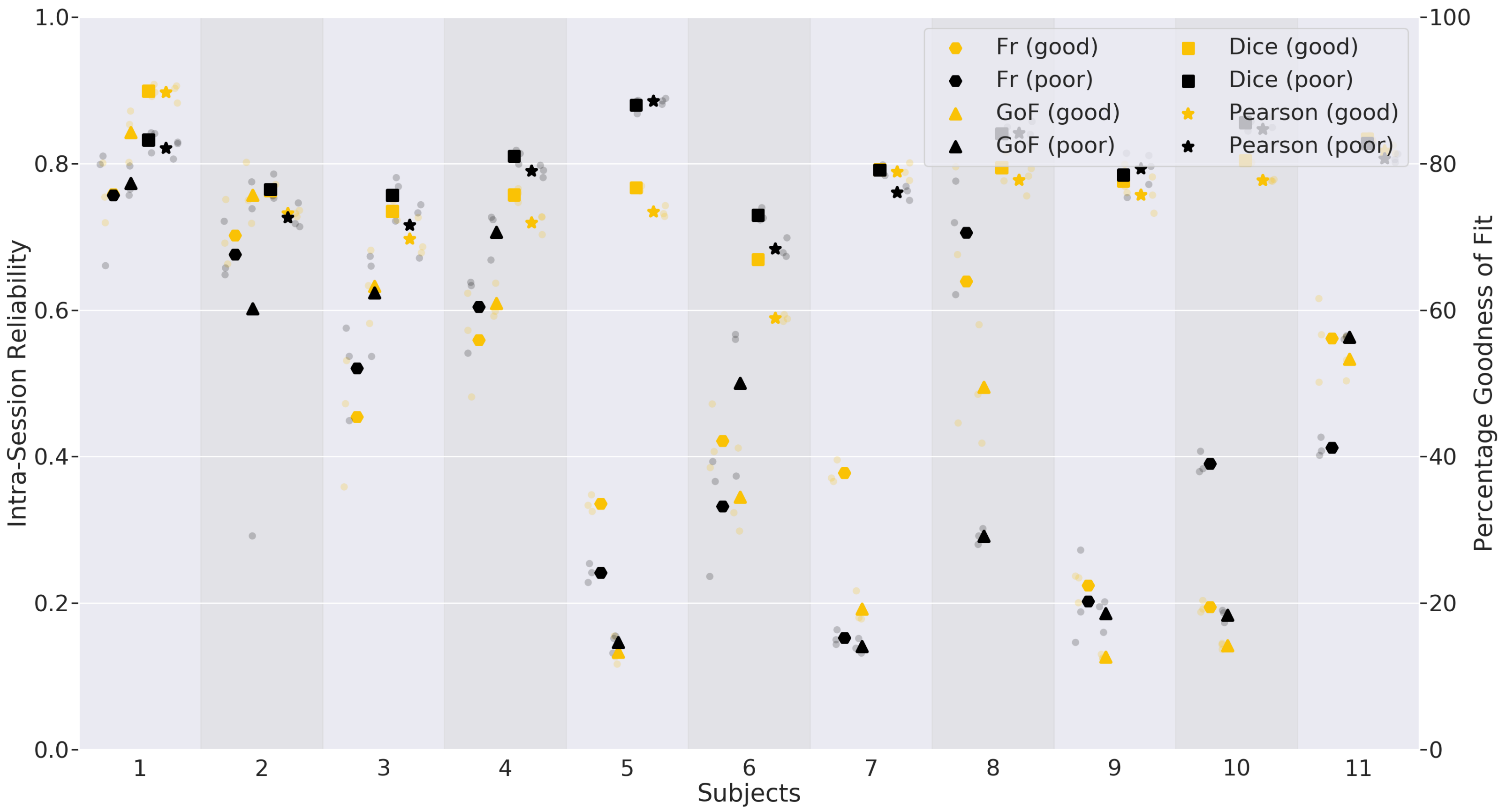


Figure 3.6: Average quality metric scores (opaque) at N20m somatosensory response peak. Semi-opaque markers reflect the spread of points within the response peak window.

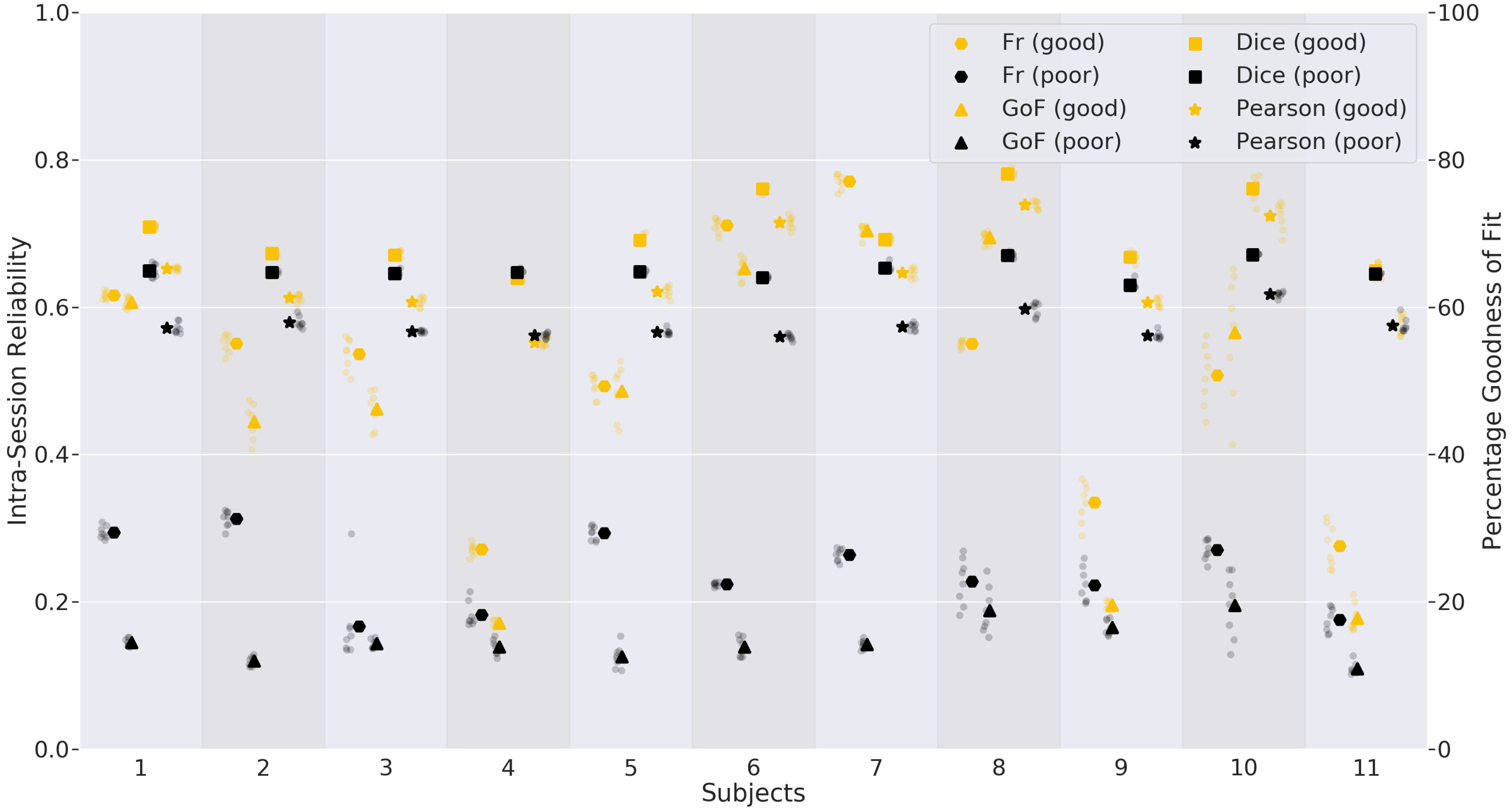


Figure 3.7: Average quality metric scores (opaque) at N75m visual response peak. Semi-opaque markers reflect the spread of points within the response peak window.

Subject	**Somatosensory Peak Latency (ms)**	**Visual Peak Latency (ms)**
sub01	21	73
sub02	19	67
sub03	20	84
sub04	21	66
sub05	40	62
sub06	19	68
sub07	74	63
sub08	38	58
sub09	19	61
sub10	25	64
sub11	24	74

Table 3.2: Latencies of response peaks (in milliseconds post-stimulus) found from the event-related field data in the good condition for each subject.

	t-value	**p-value**
Goodness of Fit	0.606	0.279
Reliable Fraction	0.258	0.401
Pearson Correlation	-1.813	0.950
Dice Coefficient	-1.834	0.952

Table 3.3: Statistical significance of difference between good and poor somatosensory data at the N20m response peak measured by each metric across all subjects.

	t-value	**p-value**
Goodness of Fit	6.18	5.2×10^{-5}
Reliable Fraction	5.45	1.4×10^{-4}
Pearson Correlation	4.12	1.0×10^{-3}
Dice Coefficient	3.96	1.3×10^{-3}

Table 3.4: Statistical significance of difference between good and poor visual data at the N75m response peak measured by each metric across all subjects.

randomly selected split-half data sets. These results are shown in Fig. 3.8. As such, all values for reliable fraction presented at the group comparison level were averaged across 30 split-half datasets.

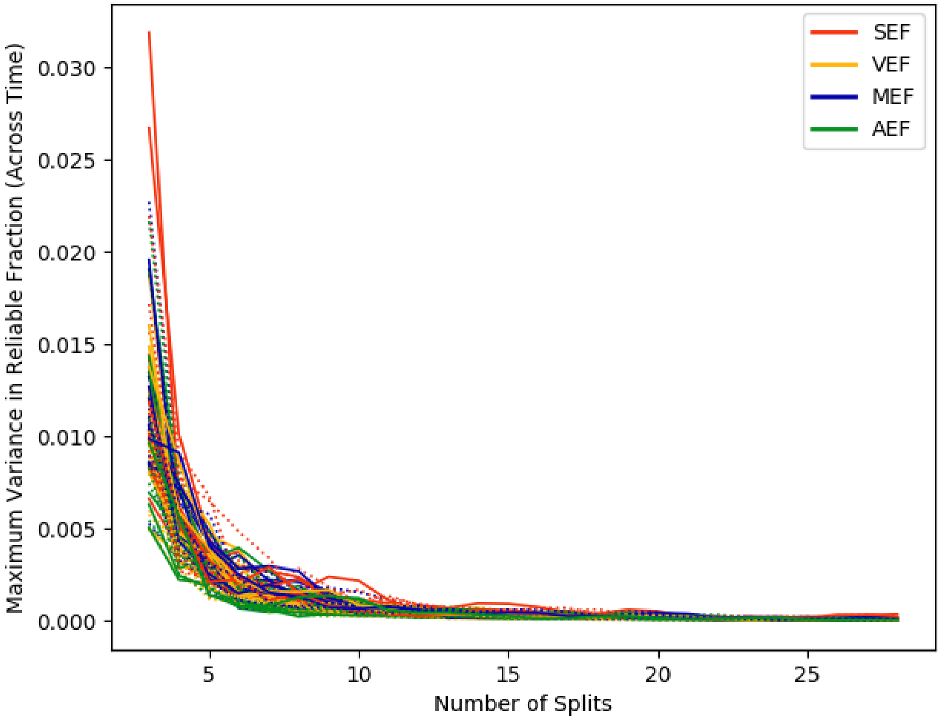


Figure 3.8: Maximum variance in the reliable fraction across all time points and paradigms when averaged across varying numbers of split-half datasets for subject 1. Solid lines correspond to data acquired in the good condition; dashed lines correspond to data acquired in the poor condition.

3.3 Single Subject Quality Metric Comparison

A comparison of metric performance at the single-subject level was made for visual data only, since this dataset provided consistent activation in the good condition and a successful modulation of activation map quality. A nonparametric bootstrap was used to generate a distribution of metric values for data in each condition for each subject.

3.3.1 Response Peak Distributions

Each bootstrap distribution was obtained at a single time point for each metric (and subject). This time point was selected by choosing the time corresponding to the maximum value reached by each metric near the response peaks selected for the previous group comparison. These time points are presented in Table 3.5. The difference between the time points at which each metric reached its peak and the response peak latency previously selected from the peak in the global field power (Table 3.2) is given in Table 3.6. No significant difference exists between the mean times at which each metric peaked across all subjects. However, it was slightly more common for metrics to peak at a time later than the global field power (true for all metrics for 6/11 subjects).

For each subject, a distribution of quality score values is plotted for each metric and performance condition in Fig. 3.9 over 120 bootstrap datasets. For most subjects (8/11), all metrics show a large separation (falling outside the interquartile range of the opposite condition's distribution) between the mean values for good and poor data. However, the good and poor distributions generally overlap for all subjects and metrics, with the only exception being goodness of fit and reliable fraction for subjects 6 & 7. The Dice and Pearson metric values exhibit a much smaller range (approximately 0.5-0.8) than goodness of fit and reliable fraction (0-0.8). This is to be expected from the group quality metric curves (Fig. 3.5), where the Dice and Pearson coefficient curves have a higher baseline value and a smaller separation between good and poor curves than goodness of fit and the reliable fraction. Three subjects (4, 9, & 11) appear to have a much smaller separation between good and poor data due to reduced quality scores in the good condition across all metrics.

3.3.2 Statistical Comparison

For each subject and metric, a one-sided t-test was used to measure the statistical significance of the separation between good and poor data. The results of these statistical tests are shown in Table 3.7. All metrics measured a significant ($p<0.001$) increase in data quality values for the good condition, as compared to the poor condition, for all subjects, with the exception of the Dice and Pearson coefficients for subject 11, which measured $p \approx 0.03$. This is a further validation of a measurable change in data quality across all metrics as the result of our manipulation (Hypothesis 1) for the visual dataset. The t-statistic measured for each subject by each metric is plotted in Fig. 3.10. For eight out of the eleven subjects, goodness of fit and reliable fraction measured a more significant separation than the Dice or Pearson coefficients. For the

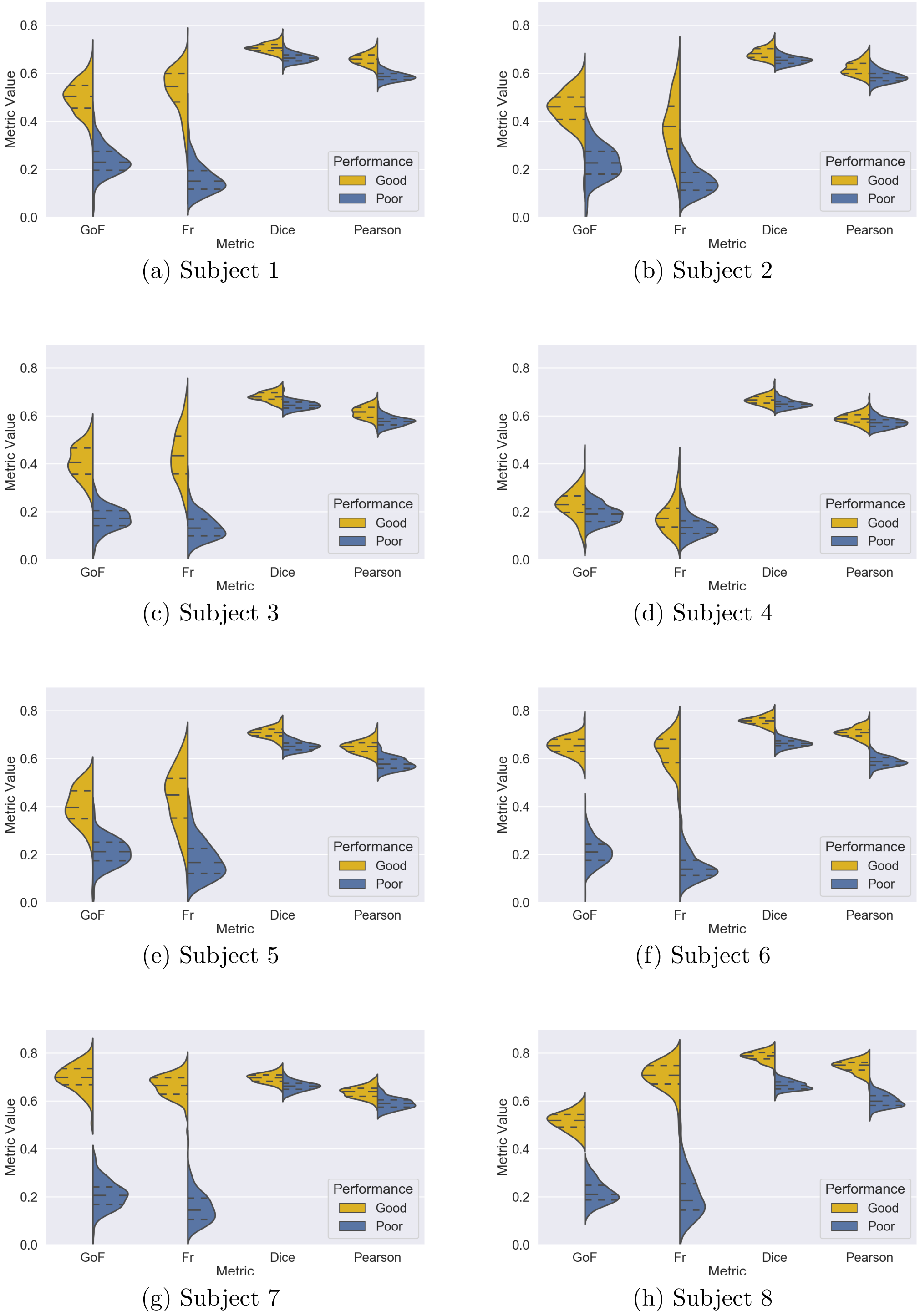
Subject	Goodness of Fit	Reliable Fraction	Dice	Pearson
sub01	80	70	80	66
sub02	65	65	63	64
sub03	95	94	94	98
sub04	72	61	73	73
sub05	61	64	60	62
sub06	65	66	65	65
sub07	63	63	63	63
sub08	61	59	60	61
sub09	68	68	64	64
sub10	71	71	71	71
sub11	81	81	81	81

Table 3.5: Latencies (in milliseconds post-stimulus) at which each metric reached its maximum value in a window centred around response peak latencies (Table 3.2) for visual data in the good condition for each subject.

Subject	Goodness of Fit	Reliable Fraction	Dice	Pearson
sub01	7	-3	7	-7
sub02	-2	-2	-4	-3
sub03	11	10	10	14
sub04	6	-5	7	7
sub05	-1	2	-2	0
sub06	-3	-2	-3	-3
sub07	0	0	0	0
sub08	3	1	2	3
sub09	7	7	3	3
sub10	7	7	7	7
sub11	7	7	7	7
Average ± St. Dev.	3.8±4.6	2.0±5.0	3.1±4.8	2.5±6.0

Table 3.6: Difference between latency at which each metric reached its peak (Table 3.5) and latency of response peak in event-related field data (Table 3.2) for visual data in the good condition for each subject.

remaining three subjects (4, 5, & 9), the Dice and Pearson coefficients outperformed reliable fraction in all cases and outperformed goodness of fit in two cases. All metrics



(a) Subject 1

(b) Subject 2

(c) Subject 3

(d) Subject 4

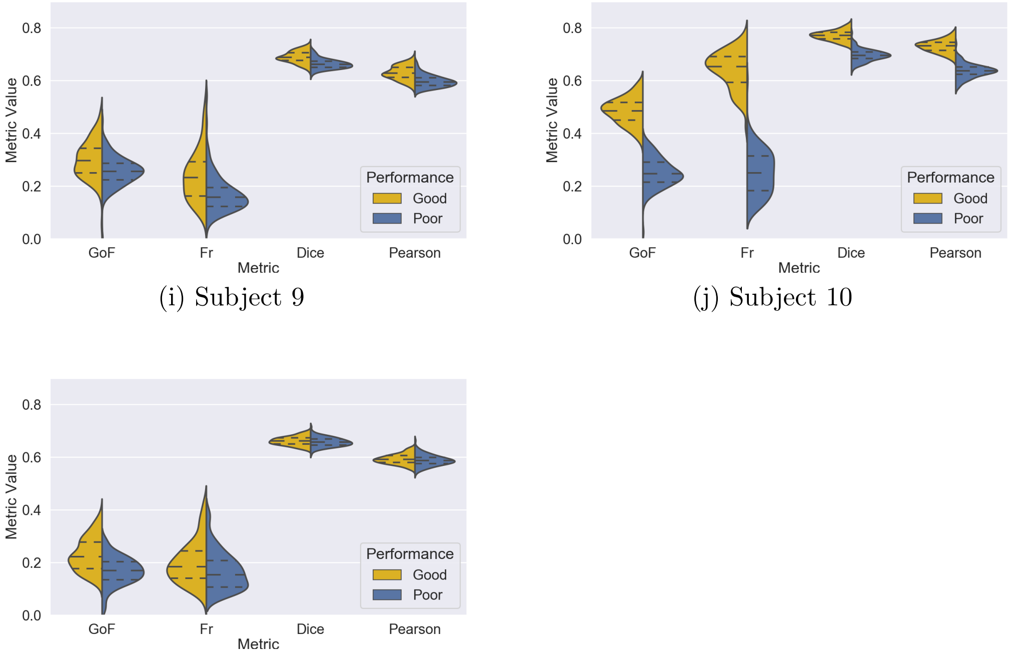
(e) Subject 5

(f) Subject 6

(g) Subject 7

(h) Subject 8

Figure 3.9: Distributions of metric values estimated at each subject's visual response peak over 120 bootstrap datasets.



(i) Subject 9

(j) Subject 10

(k) Subject 11

Figure 3.9: Distributions of metric values estimated at each subject's visual response peak over 120 bootstrap datasets.

measured relatively small increases in quality for good data compared to poor data in these three subjects. However, in the case of subject 11, which had a similarly small change in data quality, goodness of fit and reliable fraction outperformed the Dice and Pearson coefficients. As shown in Table 3.7, goodness of fit tended to measure the greatest separation between good and poor data (6/11 subjects), followed by reliable fraction (3/11) and the Dice coefficient (2/11). Goodness of fit had the largest mean t-statistic (32.00), followed by reliable fraction (30.98), the Pearson coefficient (19.56), and the Dice coefficient (19.20). Group means with their corresponding 95% confidence intervals are reported in Table 3.7, and visually represented in Fig. 3.11. However, there were no statistically significant differences between the group mean t-stat for each metric as determined by one-way ANOVA (F(3,40) = 1.51, p = 0.228). The group means cannot therefore be interpreted as a validation of Hypothesis 2.2, as reliable fraction did not measure a statistically greater change in data quality for all subjects. However, further ranking the metrics by the significance of the change in data quality measured for each subject from largest (1) to smallest (4) and summing these rank scores for all subjects, I found that goodness of fit ranked best across all subjects (rank 19), followed by reliable fraction (rank 24), the Dice coefficient (rank 33), and the Pearson coefficient (rank 34). This suggests that the larger group mean t-statistics for goodness of fit and reliable fraction are not the result of statistical variation, but reflect a greater sensitivity to changes in data quality.

Subject	**Goodness of Fit**	**Reliable Fraction**	**Dice**	**Pearson**
sub01	30.02 (2)	43.87 (1)	18.76 (4)	26.81 (3)
sub02	23.15 (1)	18.59 (2)	11.17 (3)	10.59 (4)
sub03	31.16 (1)	27.81 (2)	14.93 (3)	12.49 (4)
sub04	6.027 (3)	5.42 (4)	7.13 (1)	6.06 (2)
sub05	21.20 (1)	19.26 (4)	20.86 (2)	20.58 (3)
sub06	72.97 (1)	57.26 (2)	39.48 (4)	42.11 (3)
sub07	66.85 (2)	70.27 (1)	13.53 (4)	15.39 (3)
sub08	52.82 (1)	50.60 (2)	45.96 (3)	40.56 (4)
sub09	5.29 (4)	6.16 (3)	9.86 (1)	9.82 (2)
sub10	34.34 (2)	38.30 (1)	27.62 (4)	28.89 (3)
sub11	8.25 (1)	3.22 (2)	1.89 (3)	1.86 (4)
Average t-stat	32.00±15.73	30.98±15.29	19.20±9.13	19.56±9.08
Overall ranking	19	24	33	34

Table 3.7: Statistical difference between good and poor data for each subject measured by each metric using a one-sided t-test. For each subject, bracketed numbers show the metrics ranked in order of the magnitude of the change in data quality measured (1 being the greatest).

3.3.3 Convergence Across Bootstrap Datasets

The variance in each metric was initially tracked over the course of 1000 bootstrap datasets for subject 1 to investigate the progression of the convergence of bootstrap datasets. As shown in Fig. 3.12 (a), variance in reliable fraction in the good condition increased as the number of bootstrap datasets increased from approximately 100 to 1000. It is notable that the variance in reliable fraction in the good condition increased sharply during the first 50 bootstrap resamples. This behaviour, along with any later increase, was not observed for the other metrics and for reliable fraction in

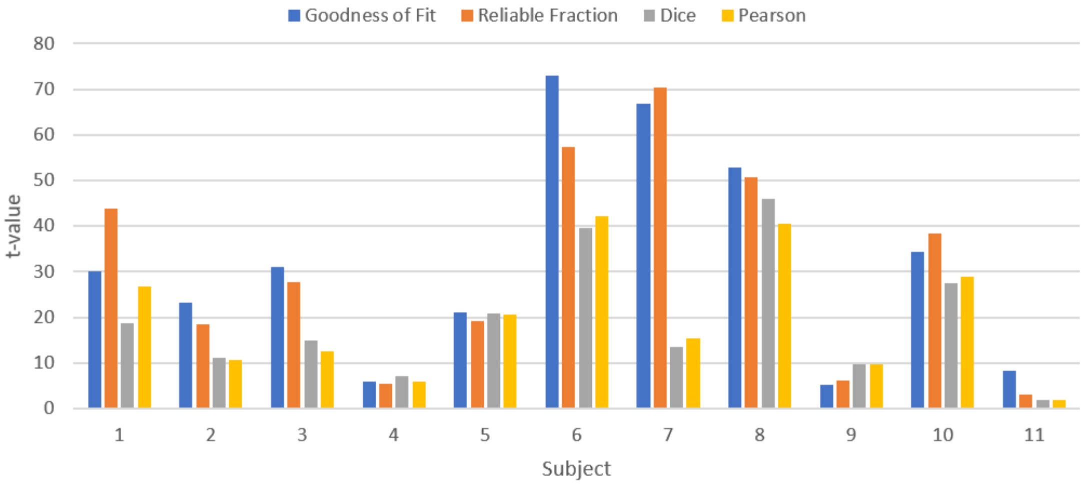


Figure 3.10: Statistical difference between good and poor data for each subject measured by each metric using a one-sided t-test.

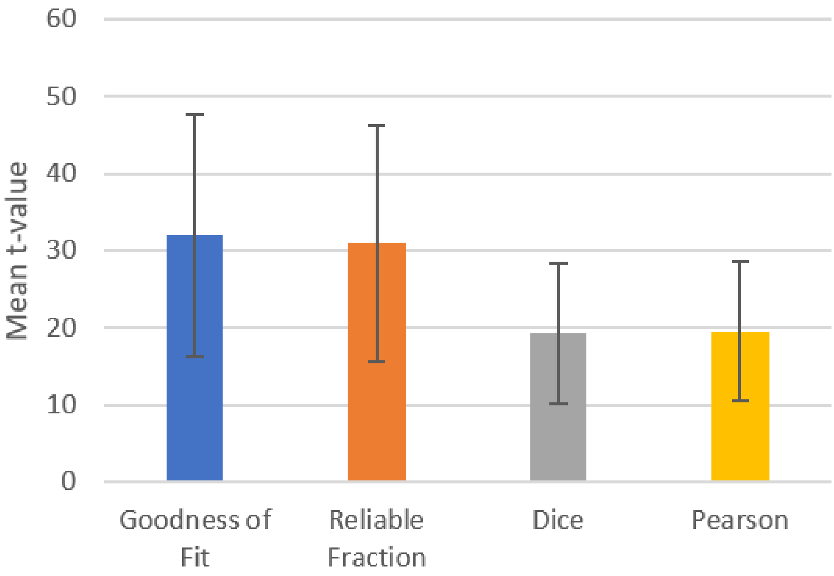
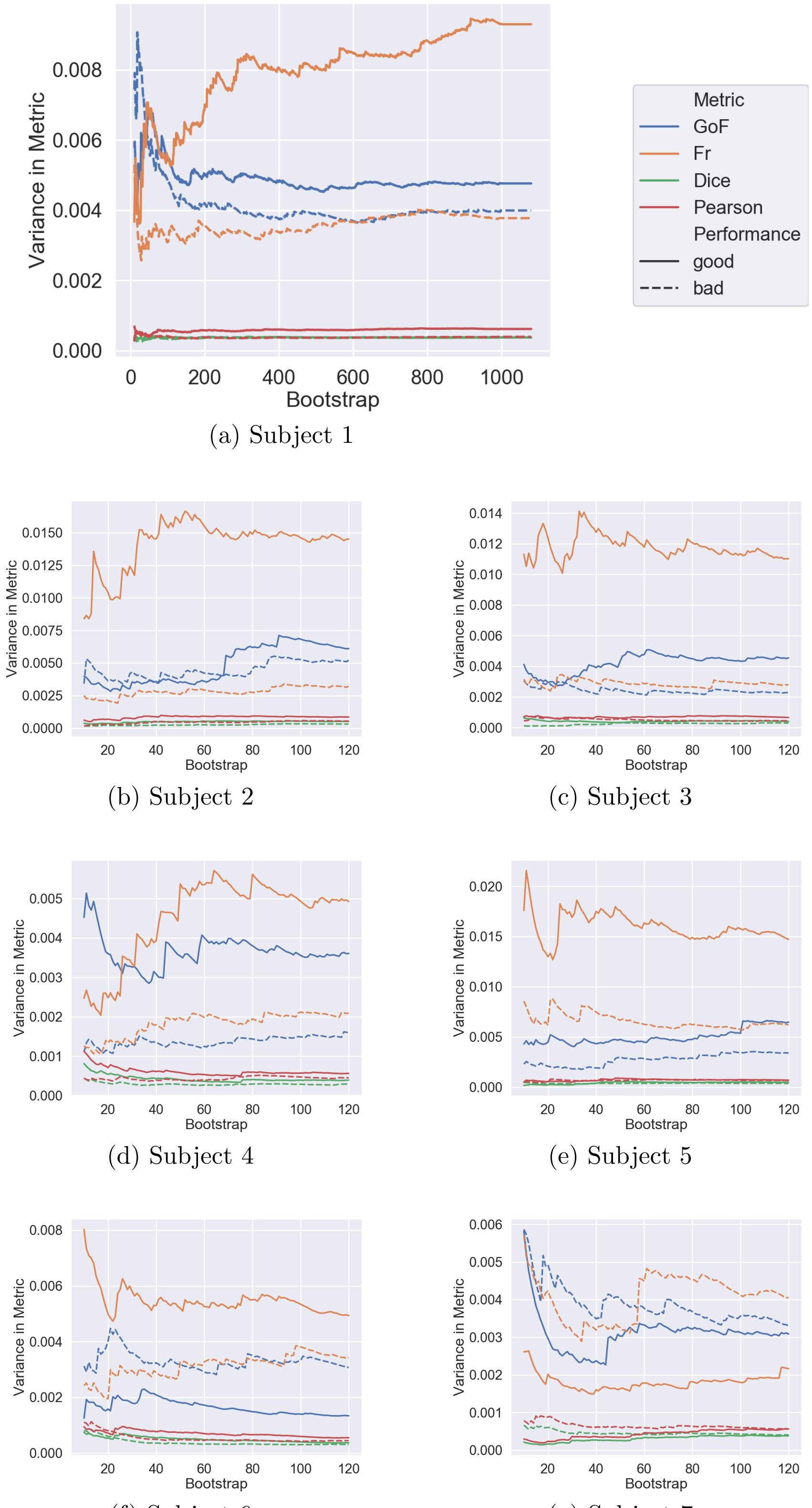


Figure 3.11: Mean statistical difference between good and poor data for each subject measured by each metric using a one-sided t-test. Error bars indicate 95% confidence intervals.

the poor condition, which converged to a stable variance within 100-150 bootstrap resamples. It is unclear whether a similar increase in variance would manifest in the other subjects, which were run for only 120 bootstrap resamples due to limitations on processing time. A similar initial increase in variance for reliable fraction in the good condition may be seen in subjects 2, 4 & 11, while reliable fraction in the poor condition exhibits a very large increase ($> 200\%$) near bootstrap 30 in subject 9. In my opinion, even without complete convergence, the metric distributions estimated over 120 bootstraps should be sufficient for the metric comparison in the previous section. The large increase in variance in cases such as subject 9 suggests that significant outliers can result from metric calculations on a single bootstrap dataset. Randomly resampling with replacement from a relatively small number of epochs (100) could be expected to occasionally generate a dataset composed of a large number of repeated epochs, which might unduly bias the estimation of reliability-based metrics.



(a) Subject 1

(b) Subject 2

(c) Subject 3

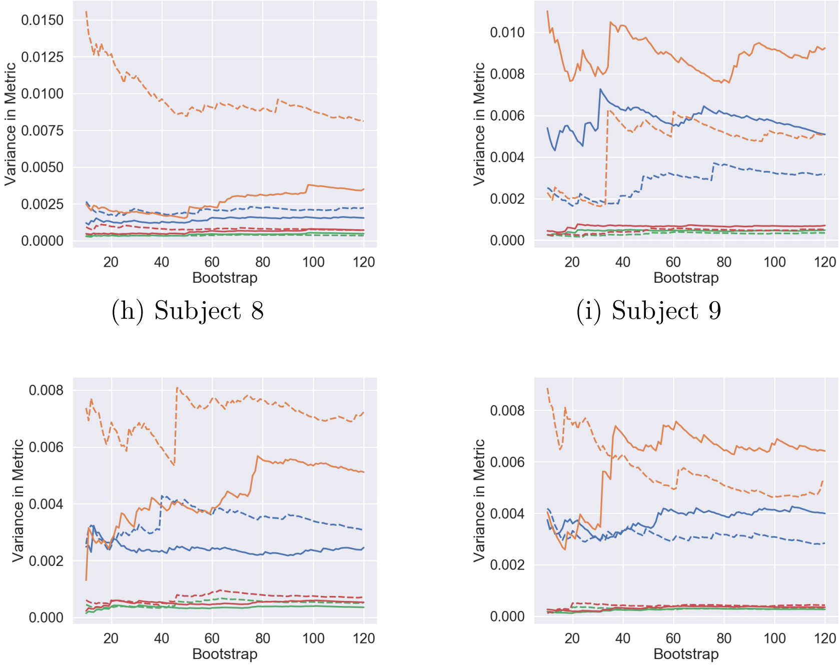
(d) Subject 4

(e) Subject 5

(f) Subject 6

(g) Subject 7

Figure 3.12: Variance in each metric as the number of bootstrap datasets increases.



(h) Subject 8

(i) Subject 9

(j) Subject 10

(k) Subject 11

Figure 3.12: Variance in each metric as the number of bootstrap datasets increases.

Chapter 4

Discussion

This thesis has compared the performance of three intra-session reliability metrics in distinguishing MEG functional data acquired for focal mapping paradigms in good and poor conditions. This comparison was complicated by acquiring high quality data with varying levels of success for different paradigms in the absence of a 'true' measure for data quality. Nonetheless, in the two suitable mapping paradigms, intra-session reliability metrics demonstrated comparable sensitivity to data quality relative to the clinical standard metric, ECD goodness of fit. Moreover, reliable fraction displayed greater sensitivity to changes in data quality than either the Dice or Pearson coefficients. In this chapter, I will discuss this comparison and its confounds in more detail while drawing out differences between the three intra-session reliability metrics.

4.1 Summary of Main Findings

I acquired MEG data using pre-surgical mapping paradigms to localize the primary somatosensory, motor, visual, and auditory cortices in twelve healthy subjects. For each subject and paradigm, I acquired data in both a standard 'good' condition and in a 'poor' condition where data quality issues were simulated in unique ways for each paradigm. My hypotheses for the outcome of data manipulation and the performance of metrics across subjects and paradigms were as follows:

> **Hypothesis 1**: Across all subjects and paradigms, I hypothesized that all quality assurance metrics would measure higher values for scans acquired in the good condition than for those acquired in the poor condition at times when an evoked response was present in the good data.

> **Hypothesis 2.1**: I predicted that a significantly larger change in reliable fraction would be found between all good and poor datasets across the group, as compared to goodness of fit and the Dice and Pearson coefficients.

Hypothesis 2.2: I predicted that at the single-subject level, reliable fraction would measure the largest difference between good and poor datasets across all subjects, as compared to the other metrics.

4.1.1 Outcomes for Hypothesis 1 and 2.1

Significant variation in data quality across mapping paradigms was observed. While the expected responses were evoked in somatosensory and visual data, it was difficult to elicit a consistent motor response. Moreover, no auditory response was observed in 8/11 subjects.

Metric performance was evaluated across the group for both visual and somatosensory data. No significant difference between good and poor somatosensory data was found by any metric at the group-level (Table 3.3). This was expected based on the MEG data processing pipeline's ability to mitigate sources of noise such as the simulated dental implant. While data quality was low prior to processing, the overall quality of activation maps achieved in the good condition was high. Thus, while Hypothesis 1 was not observed in somatosensory data, this can be viewed as a positive effect for all data quality metrics. The most useful comparison of metric performance across the group was shown for the visual mapping paradigm, since qualitative examination of visual data showed significant differences between data acquired in the good and poor conditions (Figs. 3.1 and 3.2). All metrics measured a significant change in data quality across the group (Table 3.4), confirming Hypothesis 1 for this data. However, goodness of fit measured the largest change in quality, contradicting Hypothesis 2.1. Notably, reliable fraction measured a larger change than the Dice or Pearson coefficients, suggesting greater sensitivity to changes in data quality.

4.1.2 Outcome for Hypothesis 2.2

Using a nonparametric bootstrap to estimate metric distributions for each subject's good and poor visual data, I found that goodness of fit measured a larger change in data quality in the majority of subjects (6/11; see Fig. 3.10). Reliable fraction measured the largest change in data quality in 3 subjects, while the Dice coefficient measured the largest change in data quality in 2 subjects. One-way ANOVA across the group found no significant difference in metric performance at the group level. However, ranking metric performance for each subject, reliable fraction (rank 24) consistently outperformed the Dice (rank 33) and Pearson (rank 34) coefficients (Table 3.7) for individual subjects. Thus, while Hypothesis 2.2 was not confirmed across the group, reliable fraction demonstrated a considerable advantage over the Dice and Pearson coefficients, exhibiting similar behaviour to goodness of fit (rank 19). For cases of distributed activity where goodness of fit is not a viable quality metric, reliable fraction should be considered as a promising measure of data quality.

4.2 Manipulation of Data Quality and Metric Performance

In this section, I discuss the overall success of acquiring data in two distinct 'good' and 'poor' conditions for each mapping paradigm. For paradigms in which a useful comparison could be made between good and poor quality data, I discuss the observed metric behaviour. This section also serves as justification for the exclusion of the motor and auditory datasets from the metric comparison aspect of this thesis. In these cases, I discuss the difficulties in eliciting a robust response and suggest potential solutions for future data acquisitions.

4.2.1 Somatosensory Data

Median nerve stimulation is known to be a highly robust method of localizing the primary somatosensory cortex. However, no response was observed in 3/11 subjects. These cases are the result of subject movement displacing the electrodes during the scan. The electrodes could have been especially susceptible to displacement due to operator inexperience with the mapping paradigm. It should be noted here that many issues with collecting high quality data can at some level be attributed to the MEG operator(s). That is not to assign blame but to point out that following established procedures and qualitatively inspecting the data during the scan to ensure the expected response is present is one of the mainstays of MEG quality assurance. Data for this study was collected by an operator new to this particular MEG imaging centre and not involved in the study design, and was assisted by myself (with no previous MEG-related experience), which perhaps contributed to substandard adherence to standard mapping procedures.

However, in the cases where a somatosensory response was successfully evoked, it was straightforward to acquire both good data and data simulating the presence of a dental implant. While the effects of this manipulation were observed as noise in the sensor level data (Fig. 3.1), a difference between good and poor activation maps across the group was difficult to distinguish, both qualitatively (Fig. 3.2) and quantitatively (Fig. 3.3 and Table 3.3). This is a useful demonstration of the MEG processing pipeline's ability to compensate for artefactual sources of noise. In particular, tSSS is well-equipped to deal with typical dental artefacts [56]. While the beamformer's natural suppression of noise could have further contributed to the removal of the magnetic artefact [12] and has been shown to compensate for similar artefacts [17], the high quality scores measured by goodness of fit in the poor condition suggest that the artefact was suppressed prior to source localization. The sensitivity of all four metrics to the quality of source localization rather than to artefacts in the raw data demonstrates their utility as quality assurance metrics for pre-surgical mapping.

4.2.2 Motor Data

While a motor response was elicited in most subjects in the good condition (Table 1.1) and localization in the area of the motor cortex was observed in grand average ERB data (Fig. 3.2), the overall quality of purportedly good motor data was poor. This can be seen from a qualitative inspection of the ERF data acquired in the good condition for each subject (Fig. A.1). While sensor-level responses were observed at the single-subject level, these were found by manually selecting potential peaks in the GFP and averaging the sensor topography over several milliseconds. Here, averaging did not amplify response peaks for each subject, suggesting that a focal motor response was not consistently evoked. It is probable that this was caused by improper task performance. Most guidelines recommend training subjects in the performance of motor tasks and providing opportunities to practice prior to the scan [15]; more comprehensive training by an experienced technologist might have improved data quality. Monitoring the EMG response during the scan showed consistent muscle activity following the visual cue in the good condition, but drawn out movement or improper movement of other muscles would not have been obvious to detect. Due to the difficulty in identifying response peaks in single-subject data and the overall low data quality measured in the good condition (Fig. 3.4), this paradigm was deemed unsuitable for a further comparison of metric performance.

4.2.3 Visual Data

It is well known that the visual evoked response to shifts in stimulus pattern is suppressed during eye movement [57]. Beyond direct suppression as the result of eye movement in the poor acquisition condition, the moving fixation cross caused a translation of the checkerboard across the retina. Since changing images in the visual field activate specific regions of the primary visual cortex corresponding to their retinotopic location [58], controlled eye movement disrupted focal retinotopic mapping in the poor visual data. If several regions of the visual cortex were activated during the short time periods in which the fixation cross was stationary, there may have been too few trials to distinguish the corresponding signals from the background noise. Alternatively, once the data had been epoched and averaged, cancellation of competing signals from different regions of the visual cortex could have masked a distributed response [59]. Ultimately, data quality manipulation was very successful for this paradigm. The N75m visual response peak was strongly present in all subjects' good visual data, and in only one instance for poor visual data (Table 1.1). This paradigm was therefore a good candidate to test the ability of each metric to distinguish good and poor data quality. All metrics measured a significant difference between good and poor data across subjects at the visual response peak, as expected. However, variability in the quality of good data across subjects made a statistical comparison of metric performance difficult. This limitation and possible solutions are further addressed in sections 4.5 and 4.6.2.

4.2.4 Auditory Data

Although an auditory response could be observed in group-level data (Figs. 3.1 and 3.2), the corresponding response was observed in only 3/11 subjects (Table 1.1). This was caused by sound levels too low to evoke an auditory response in all subjects, most likely due to improper transducer positioning within the ear. Although participants were asked to confirm that they were hearing a series of tones in one ear and white noise in the other after a brief audio test, it could have been more informative to ask subjects to clearly describe the sounds heard and to qualitatively characterize their loudness. Due to the small number of useful auditory datasets, this paradigm was excluded from further analysis.

4.3 Advantages of Reliability Analysis at the Source Level

Topographical analysis at the sensor level can often provide insight into MEG data. It is worthwhile to consider the potential applications for intra-session reliability applied to sensor measurements, rather than estimations of source activity. On one hand, this approach would allow for much faster calculation of intra-session reliability metrics. Not only would time-consuming source estimation be omitted, but the dimensionality of each split-half dataset would be greatly reduced. For example, assessing reliability using magnetometers would require calculations to be performed on only 100 channels, rather than on 12000 voxels. This reduction in processing time is most significant for reliable fraction. Sensor-level analysis could potentially lead to a real-time analysis framework, in which epochs are rapidly sorted into test and retest sets during acquisition, averaged, and compared immediately to guide the operator in collecting a sufficient amount of good quality data.

However, the shortcomings of this approach can be seen from the intra-session reliability results obtained for somatosensory data. As previously discussed, noise and other artefacts in raw data can be effectively mitigated by the MEG processing pipeline, meaning that a low quality score at the sensor level might not correspond to a low quality activation map. Moreover, it could be more difficult to distinguish any reliable artefacts from the desired response in the raw data without the additional context of source estimation. For this reason, although sensor-level reliability could be a useful tool for the operator to assess the prominence of artefacts in raw data or as a general measure of task performance, it is not well-suited to measure the quality of data for the purpose of pre-surgical mapping.

4.4 Comparing Intra-Session Reliability Metrics

In general, the three intra-session reliability metrics behaved very similarly, increasing in value at similar times in response to large magnetic field deflections for good data across subjects and paradigms. However, the Dice and Pearson coefficients measured a high baseline reliability score (approximately 0.6; see Figs. 3.3 and 3.5) relative to reliable fraction (approximately 0.15). As demonstrated in Fig. A.2, baseline subtracting each metric and then scaling to the variance in the baseline further emphasizes the similarities between all three intra-session reliability metrics. In particular, these three metrics have significantly less variance in the baseline interval relative to goodness of fit, perhaps the result of averaging over a number of split-half datasets.

However, the Dice and Pearson coefficients' high baseline values are noteworthy in that this suggests a significant proportion of these metrics' value is not sensitive to the presence of reliable brain activity. In the case of a series of two randomly generated vectors with values ranging from 0 to 1, the average Dice coefficient will converge to 0.75, indicating a large overlap between the vectors not caused by any structural similarities in the two images. It should further be noted that this value decreases if the scaling of one of the two vectors changes relative to the other. It may be that variation in the Dice coefficient is therefore more sensitive to similarities in the ERB scaling than to similarities in the location of ERB peaks. In contrast, the Pearson coefficient is mean-subtracted, and so is not sensitive to changes in the mean ERB activity which might exist between the test and retest dataset. Unlike the Dice coefficient, the average Pearson correlation for a series of randomly generated vectors will converge to 0. Thus, the high baseline value of the Pearson coefficient does not result from noise, instead indicating that some structural similarities exist between the flattened test and retest maps even in the absence of strong brain responses. Further investigation is required to identify the source of these structures, which could perhaps relate to the smoothness of the beamformer solution.

In contrast, the ROC-r comparison of test and retest datasets through successive thresholding ensures that the reliable fraction is strongly dependent on the relative location of ERB peaks. The sensitivity of these metrics to the properties of peaks in activation maps could be further tested by restricting the region of comparison to exclude regions of the brain exhibiting little significant activity. In this case, I would expect reliable fraction to increase, as it did in a previous study when ROC-r analysis was performed only on the hemisphere in which the mapping paradigm evoked a response [30]. In contrast, I would expect the Dice and Pearson coefficients to take on a more variable range of values and potentially decrease, depending on the strength of the evoked response.

On the group level, all three metrics reach their peaks at nearly the same time at the N20m somatosensory response and at the same time at the N75m visual response. However, on the individual level, the three metrics were less consistent. Table 3.6 shows that across all subjects, each metric reached its peak for a fairly wide range of latencies relative to the peak in global field power. It is particularly noteworthy that the peaks of the three reliability metrics differed significantly from each other in timing for several subjects, and were generally not located at the peak in the global field power. Peaks in global field power have been associated with high similarity in field topography across nearby latencies for EEG measurements [60], so it is surprising that this measure does not correspond more directly with agreement between epochs at a given time point. Assuming that similar topographies across several latencies may correspond to temporal variations in the same response, it is possible that a split-half comparison could lead to greater sensitivity to differences in response amplitude indicating a suboptimal alignment of responses across epochs. However, since in several cases all metrics reach a maximum for different latencies, a more thorough investigation of the features classified as reliable by each metric is warranted. In the future, it would be informative to compare topographies at which each metric is maximized at both the source and sensor levels. Subtraction with topographies at the GFP peak could highlight characteristics of reliable data prioritized by each metric.

4.5 Variance in Data Quality Between Subjects

Section 4.2 discussed the general success of each mapping paradigm in acquiring good and bad data. However, even for the paradigms where data manipulation was generally successful, significant variability in data quality between subjects can be seen. This is particularly true for visual data, which even in the good condition exhibits a wide range in goodness of fit (approximately 18-70%; see Fig. 3.7). While upon first inspection there is a range of quality scores for good somatosensory data (Fig. 3.6), a qualitative inspection of subjects with low goodness of fit in the good condition shows that no somatosensory response was evoked in these cases (subjects 5, 7, 9, and 10; see Table 1.1). The goodness of fit ranges from approximately 50-90% for the remaining good somatosensory data. In constrast, a visual response was observed for all subjects in the good condition, even for datasets with low goodness of fit. Thus, it seems that the variability in somatosensory data quality is in fact smaller than the variability in visual data quality. Median nerve stimulation is known to produce a robust somatosensory response, but the extent of this disparity suggests that data quality issues affected the visual data in a unique way. The visual mapping paradigm was performed 17 minutes into the scan (not including set up time), so it is possible participant fatigue played a role. Although neither paradigm required active involvement from the participant, eye movement or drowsiness during the good visual scan could have particularly affected the evoked response (in the same way that was intended during acquisition in the poor condition). The extent of this effect would likely vary significantly from subject to subject. While the visual response was still generally observed in this data, sporadic eye movement in some subjects could have caused a broader region of activation than would be well explained by a single dipolar source.

4.5.1 Insight from Single-Subject vs. Group Comparisons

While it is fairly clear from metric scores across the group (along with the complete absence of expected responses in some subjects) that there is significant variability in 'good' data quality between subjects, the exact distribution of this variability is unknown. This particularly limits the interpretation of the separation between good and poor data on a group level (Tables 3.3 and 3.4). For example, consider a group of subjects with high variability in the change in data quality between good and poor data. A metric which measured a similar change in data quality for each subject would have a highly significant separation between good and poor data across the group while poorly reflecting the underlying quality distribution. In contrast, a metric which measured a highly variable change in data quality across the group would have a much less significant separation, but better correspond to the underlying quality distribution. In the absence of a quantification of the underlying variability, it is impossible to tell which metric has actually measured data quality more appropriately.

In contrast, a single-subject comparison is less affected by this variability. For each single subject, it is more justifiable to assume that a greater separation between good and poor metric scores corresponds to a greater sensitivity to MEG data quality. However, a comparison of metric performance (in the form of significance measured by each metric) across the group is similarly stymied by the great variance in quality between subjects. Thus, even though goodness of fit and reliable fraction measured a larger separation between good and poor subjects on average, their performance was statistically equivalent to the Dice and Pearson coefficients (Fig. 3.11).

4.6 Limitations

4.6.1 Absence of a "Gold Standard" Quality Metric

The gold standard for accurate functional mapping is intraoperative cortical stimulation. The quality of a non-invasive pre-surgical map can be most directly assessed by measuring the difference between its source localization and the location of the relevant functional activity found during the resection surgery. In the absence of such a gold standard measure, I compared intra-session reliability metrics to goodness of fit, but as previously discussed, goodness of fit is a clinically used surrogate for data quality based on the agreement of the data with an ECD model rather than a true measure of the accuracy of source localization. In another sense, I evaluated the performance of all four metrics on their ability to measure a difference between data to which I had assigned a label of 'good' or 'poor' quality. My qualitative evaluation of whether or not the expected response was present in this data was justification for this label (Table 3.3), but the strength of the expected response – corresponding to the magnitude of the separation between good and poor data – was not quantified. In this sense, it was unclear how significant the separation between good and poor data should be for each subject. The implications of this for a group comparison of metric quality are discussed in section 4.5.1.

Going forward, it might be fruitful to quantify data quality during data acquisition, depending on the expected manipulation. For instance, data from head position indicator coils could be used to quantify head movement at each time point or eye movement could be tracked via EOG readings. However, the problem of relating these physiological readings to the quality of activation maps remains. As shown in the somatosensory data with a simulated dental implant, a measurement of the intended data quality manipulation (noise resulting from a magnetic artefact) could reflect a low quality score, when the artefact did not in fact affect the quality of localization. Such a comparison would instead be most useful when quantifying a quality issue which could not be mitigated by the processing pipeline, such as poor task performance. However, this would still not account for variability in data quality resulting from unmonitored (unintentional) quality issues.

Moreover, since the ultimate goal is to measure data quality in order to improve surgical outcomes, the most informative evaluation of metric performance would be based on the clinical utility of a given activation map. In healthy subjects, anatomical landmarks can usually be used for localization [61]. The quality of functional localization could be rated by a neurologist based on agreement with anatomical data for a study comprised of healthy subjects. More ideally, a study could be designed to assess the utility of a quantitative 'quality score' in guiding a MEG operator in deciding whether to accept or reject a functional map in the presence of known quality issues. If available, combining quality scores and MEG operator ratings with the location of functional areas determined during intra-operative cortical stimulation could examine the performance of intra-session reliability metrics as a surrogate for accuracy.

4.6.2 Variability in Data Quality Across Subjects

The statistical power of the one-way ANOVA comparison of metric performance across subjects was significantly limited by the high variability in metric data quality across subjects. This can be clearly seen in Figure 3.11. Although the mean t-statistic measured by goodness of fit and reliable fraction was higher than that measured by the Dice and Pearson coefficients, there is no statistically significant difference between the four metrics due to the great disparity in 'good' data quality across the group (Fig. 3.7). However, ranking all four metrics by their performance in each subject further suggested that goodness of fit and reliable fraction were more sensitive to changes in data quality. It seems likely that the overlapping confidence intervals in the mean t-statistic are the result of variability across subjects rather than variability in metric performance. This could be overcome with a larger number of subjects to facilitate outlier detection and removal. As discussed in the previous section, a measure of quantification of data quality manipulation collected during data acquisition could also be helpful to account for variability across subjects. This could even be used to normalize data quality across subjects. For example, eye monitoring could be used to quantify eye movement during each epoch. For each subject, a dataset could be constructed to have a similar eye movement score across epochs, or accounted for as a covariate during statistical analysis. This would allow a more rigorous comparison of metric performance across subjects.

4.6.3 Variability in Source Localization Techniques

In this thesis, I have examined the performance of a dipole fitting metric and the performance of intra-session reliability metrics using split-half activation maps generated with a spatial beamformer. It is important to consider that the reliability metrics measured the agreement between activation maps, which could be strongly affected by the method of source localization. For example, auditory data was excluded from analysis due to difficulty eliciting the desired response. However, if auditory data had been included, I might have expected low reliability scores, even in the good data. The beamformer method of source localization is known to suppress strongly correlated sources [55], so reliability metrics measured on beamformer activation maps may perform poorly in the case of bilateral activation. The same is not necessarily true of other methods of localization, such as minimum-norm estimation. Furthermore, the spatial extent of the peaks localized by distributed source solutions is strongly dependent not only on the data itself, but also on the method of localization [62]. Methods with greater source leakage might have more overlap between peaks on test and retest maps, resulting in a higher reliability score.

4.6.4 Reliable Artefacts

The operator-independent interpretation of intra-session reliability metrics is thus far limited by the ability to distinguish reliable brain signals from reliable artefacts. The main assumption of a split-half reliability analysis is that noisy signals will average out, while the evoked response of interest will not. This is partially true, as we see high reliability at latencies corresponding to expected response peaks (Figs. 3.3 and

3.5). However, we also see high reliability in the presence of a strong time-locked artefact from the electrical stimulus generating the evoked somatosensory response. While the difference between this artefactual signal and the expected response peaks are clear (with an understanding of the data and typical somatosensory evoked fields) at the group level, greater variability on the single subject level could make similar artefacts difficult or impossible to distinguish. Artefacts such as undesirable visual or muscular responses could be reliable when generated by the stimulus itself or following task performance. For example, a subject could consistently perform finger abductions with the wrong hand or combine correct abductions with other movement and still receive a high reliability score. Visual cues could evoke a reliable visual response, while blinking can likewise be time-locked to cued task performance. The difficulty of identifying these causes of poor data quality emphasizes the need for proper training in task performance and the continued use of in-scan monitoring systems, such as EMG, EOG, or eye-tracking using high-speed cameras to monitor task performance and facilitate artefact removal.

4.6.5 Metric Convergence for Bootstrap Datasets

As discussed briefly in section 3.3.3, variance in reliable fraction and goodness of fit did not always converge to a stable value as the number of bootstrap datasets increased. It is particularly notable that this behaviour was most commonly observed for reliable fraction in the good condition. It is further interesting that most metrics appear to have a qualitatively symmetric distribution (Fig. 3.9). Suppose that a good dataset may generally be described as a large number of good epochs mixed with a small number of bad epochs contaminated with artefacts. If variance in a metric was primarily caused by outlying bootstrap datasets with a disproportionate number of bad epochs, the metric distributions should be asymmetric, weighted toward poor quality scores. This does not appear to be the case. A closer examination of the composition of bootstrap datasets which produced significantly outlying quality scores could provide insight into how reliability was calculated for these datasets, especially for reliable fraction compared to the Dice and Pearson coefficients. It should be noted, however, that for subject 1, the relative significance of the change in quality measured by each metric did not change when the number of bootstrap datasets was increased from 120 to 1000. Nonetheless, an even greater number of bootstrap datasets could potentially affect the metric distributions found for reliable fraction or goodness of fit. It is also noteworthy that for all metrics, the median metric value obtained by bootstrapping (Fig. 3.9) was generally lower than the metric value calculated for the entire dataset at the same time point (Fig. 3.7). A more reliable estimate of metric confidence intervals could be obtained by recording data for more than one hundred trials to reduce the effect of inadvertently amplifying artefactual signals in bootstrap datasets as the result of random resampling with replacement.

4.7 Future Directions

4.7.1 Clinical Implementation

While this thesis has compared the performance of the discussed quality metrics, it has not explicitly compared their relative ease of implementation. Since goodness of fit is determined as part of the dipole fitting procedure, its calculation cannot be viewed as separate from source localization. However, the three intra-session reliability metrics rely on data-splitting and repeated source estimation, and must therefore be calculated separately from source localizations using the entire dataset. Ideally, an intra-session QA procedure would be simple to implement. The most basic requirement is an MEG dataset which has been separated into epochs, although anatomical information for source localization should be provided if available. In this thesis, intra-session reliability calculations were readily implemented alongside standard MEG data processing using the open-access MNE-Python framework. In the future, I envision an extension of this functionality where an operator-independent QA module could be run as part of the larger data processing pipeline. Although a single quality score for the entire dataset would be easiest to interpret, we have seen that metric scores vary significantly with time (Figs. 3.3, 3.5, and 3.4), and that it is difficult to pre-select a time of interest without visual inspection (Table 3.6). An intra-session reliablity QA framework could be most useful by returning a metric time series, indicating high-quality time points at which to generate activation maps. As a tool to score MEG data scans, the highest value or average value near a paradigm-specific peak could be returned as a surrogate for overall data quality. However, operators might prefer being able to specify a single time point of interest, which would reduce computation time. Integration with a standard processing pipeline and fast metric computation could lead to the adoption of these metrics as a real-time analysis tool to prevent clinical data losses, but that is beyond the scope of this thesis.

More importantly, before these metrics should be adopted for clinical QA it will be necessary to determine a better system of assigning meaning to metric scores. As previously discussed in section 1.2.2, there is no single goodness of fit threshold used to distinguish good and poor data. If intra-session reliability metrics were to be widely adopted, operators may likewise come to associate a range of scores with good data quality based on their own underlying understanding of their datasets. However, the use of these metrics for rigorous QA should be founded in a thorough study of the performance of these metrics relative to performance in the clinical workflow. At the very least, use of a standard QA processing pipeline and better reporting of quality scores in the MEG literature could eventually lead to a community-wide consensus on the interpretation of these metrics.

4.7.2 Computational Efficiency

Closely related to ease of implementation is the processing time required for calculation of each metric. I will discuss only the efficiency of calculating the intra-session reliability metrics, since calculation of goodness of fit is typically performed within third party optimization routines and does not require processing time separate from source localization. Of the three reliability metrics, reliable fraction requires the most time to calculate. Since the ROC-r analysis framework is subject to ongoing development, calculation of the reliable fraction currently relies on uncompiled code. The largest improvement in efficiency will likely come from a transition to compiled code. However, this discussion will focus on possible improvements in efficiency which could be achieved from methodological changes in calculating the reliable fraction. While all of the proposed intra-session reliablity metrics are repeatedly calculated and averaged over multiple split-half datasets, the Dice and Pearson coefficients are simple functions (Eq. 1.4 and Eq. 1.5), obtained quickly and easily from the test and retest activation maps. In contrast, reliable fraction requires the comparison of the test and retest activation maps over a number of thresholds to generate different

ROC curves for each test threshold. Interpolation is then required to estimate the threshold at which the area under the ROC curve would first be equal to 0.75 ($t_{0.75}$).

In order to make these metrics clinically appealing, it may be necessary to further optimize the required processing time. In my single-subject metric comparison, I switched from calculating each metric over a fixed number of thirty split-half datasets to calculating each metric over as many split-half datasets as significantly reduced variance in the values obtained. For each subject, time point, and bootstrap dataset, my approach found no significant improvement in metric convergence after twelve split-half datasets, a 60% reduction in processing time. However, it is worthwhile to consider whether further improvements in computational efficiency may be made in this area. A more rigorous analysis of the spread in metric values over different split-half dataset configurations could prove instructive. For example, for quality assurance, the most conservative estimate for the metric value might be most desirable. If datasets were split in half chronologically rather than randomly, this could provide a measure more sensitive to degradations in quality resulting from subject fatigue over the course of the scan. Quality issues (such as a lack of alertness or engagement with task performance) affecting the data as a whole might be more concentrated in the latter half of the scan, resulting in a greater difference between test and retest datasets and lower intra-session reliability. A future investigation could examine the correlation between metric values and the chronological ordering of epochs during split-half divisions. It would also be helpful to more formally quantify the expected confidence limits for each metric. Combined with a pre-determined threshold for ‘good’ data, a scan scoring well above or below this threshold on its first split-half dataset might not require further averaging. In contrast, a score close to the threshold could be flagged for further analysis.

For reliable fraction in particular, the process by which $t_{0.75}$ is obtained for each threshold could be further streamlined. The number of thresholds required for estimation of $t_{0.75}$ could potentially be reduced without comprimising the accuracy of interpolation. Alternatively, an iterative search method could be used to calculate $t_{0.75}$ while comparing the test and retest datasets at thresholds more relevant to reliable fraction. Calculating $t_{0.75}$ more quickly and accurately could help to reduce the variability in reliable fraction between split-half datasets, further improving computational efficiency. These possibilities should be explored before putting reliable fraction (or the Dice or Pearson coefficients) forward for clinical implementation.

4.7.3 Distributed Activity

This thesis has been motivated in large part by the prospect of an MEG data quality metric suitable for any evoked response – focal or distributed – so the natural extension of this work is to assess the performance of intra-session reliability for more complex mapping paradigms. For pre-surgical mapping, language mapping can often benefit from localization with distributed sources, particularly when assessing hemispheric dominance [5, 63]. Mapping of individual language areas, particularly for later evoked responses, can also result in complex field patterns difficult to fit with pre-specified sources [64].

However, interrogating the performance of intra-session reliability metrics in cases of distributed activity will require a different approach than this thesis. Dipole fitting for more than one or two focal sources is strongly dependent on the *a priori* specification of the number of ECDs and their allowable configurations, so goodness of fit will not be a robust quality measure for comparison. Without an obvious surrogate for data quality to compare to, it may be possible to manipulate data quality in a physiologically quantifiable way, as discussed in section 4.6.1. However, a stronger argument for the use of these metrics to measure data quality could be made by demonstrating a correlation between intra-session reliability and inter-session reliability. For instance, it would be informative to acquire language mapping data over several sessions in an attempt to replicate a previous study which showed that processing pipeline optimization maximizing intra-session reliability metrics also minimized inter-session variability in source localization [30].

Thus far the discussion of intra-session reliability metrics has been limited to the case of task-based activity. It is worthwhile to mention the case of resting-state scans. Although this is another instance of distributed activation, there is no natural extension of a similar intra-session reliability approach to quality analysis. Resting-state data have no events corresponding to stimuli for epoching (although are sometimes arbitrarily divided for a sliding window analysis based on the frequencies of interest), so there is no natural division into split-half datasets. Moreover, even arbitrarily dividing resting-state data into a test and retest dataset, there is no reason to expect a reliable response to be present in both activation maps. Intra-session reliability is therefore better suited to future examination in cases of evoked (rather than spontaneous) activity.

4.7.4 Other Reliability Metrics

As previously discussed (section 4.4), the Dice and Pearson coefficients are only partially dependent on peaks in the activation map. Since the reliability of source localization (corresponding to the peaks) is more clinically relevant than the reliability of low-activity structures in the beamformer activation map, an ideal quality assurance metric would be more sensitive to peak location. Previous pattern recognition algorithms have adapted the Dice coefficient to be more sensitive to certain image pixels by calculating a weighted Dice coefficient [65, 66]. I propose that a similar approach be investigated for the Dice and Pearson coefficients as intra-session reliability measures. Activation maps could be thresholded to some percentage of maximum activation before calculating the Dice and Pearson coefficients, essentially weighting low-activity voxels as zero. Although these weighted coefficients would once again be sensitive to the user's choice of threshold, their implementation as a new measure could be accompanied by an initial investigation as to the effect of the threshold chosen to provide a recommendation for future use. It might also be possible to adopt a similar approach to the calculation of the reliable fraction, by plotting the Dice or Pearson coefficient as a function of threshold and defining a new metric again based on the fraction of thresholds for which a sufficiently high coefficient was calculated.

Furthermore, other possible approaches to data quality could be adapted from the functional imaging literature. One notable analysis framework has used both reliability and prediction accuracy as a surrogate for fMRI and PET data quality [67]. Strother *et al.*'s nonparametric prediction, activation, influence, and reproducibility resampling (NPAIRS) framework uses repeated split-half resampling across subjects to generate test and retest datasets on the group level. In this framework, reliability is calculated using a similarity measure such as the Pearson correlation, and represented by the histogram of that measure across all split-half datasets. This is an approach analogous to that followed in this thesis, although MEG epochs allow our analysis to be performed on the subject rather than group level. However, NPAIRS further uses the retest datasets as a training set for the prediction of experimental parameters. Validation of the trained classifier on the test datasets allows the measurement of prediction accuracy, which Strother *et al.* put forward as a data quality optimization metric.

Inspiration for new reliability metrics could also be drawn from image comparison techniques in other medical fields. For example, the gamma index is a radiotherapy quality assurance measure used to evaluate the difference between a calculated and delivered dose distribution [68]. A functional neuroimaging analogue to the gamma index would compare the activity in each voxel in the test image to the activity in the same voxel in the retest image, but would also measure the distance to the nearest voxel containing the same activity. The gamma index at a voxel is then a function of the difference in activity and this distance to agreement, and indicates higher agreement between datasets for lower gamma. A similar criterion for reliability might be a better reflection of localization similarity than the Dice or Pearson coefficients for a normalized test and retest dataset. Unlike the Dice or Pearson coefficients, which flatten the activation map into a vector form for comparison, the gamma index is better suited to compare spatial similarities between images. For example, a small translation of the same beamformer map might significantly decrease the Dice or Pearson coefficients since both metrics essentially only compare agreement between each voxel in the test image to the same voxel in the retest image.

4.7.5 Reliability Mapping

As the previous section begins to suggest, the main drawback to a simple reliability score is that it is difficult to be sure that the reliability of source localization is well-represented by the reliability of the data as a whole. Moreover reliable artefacts are indistinguishable from reliable activation without additional context. Although a single number quantifying reliability is easy to use and report for quality assurance, it provides little guidance for the interpretation of questionable data. However, approaches such as gamma analysis generate a reliability score on a voxel by voxel basis. One such approach uses replicated fMRI datasets to model an underlying reliability distribution [69]. The parameters defining this distribution can be estimated on a voxel by voxel basis to generate reliability maps [70]. If possible, adapting a similar voxel-based reliability measure could provide insight into where reliable areas of activation are located and identify unreliable activation corresponding to actefactual signals.

4.7.6 Beyond Magnetoencephalography

Electroencephalography (EEG) is notably similar to MEG, measuring the electrical currents rather than magnetic fields generated by neural activity. Although distortion of electrical head currents tends to reduce the spatial resolution of EEG [9], paradigms to evoke a response are largely the same and most analysis methods are applicable to either imaging technique. There is no reason why intra-session reliability measures would not be applicable for EEG data. Since EEG systems are not generally used for pre-surgical mapping, the quality of activation maps might be of less interest than the consistent presence of the expected response. A sensor-level implementation of intra-session reliability analysis could potentially provide such a measure while monitoring EEG data quality in real-time.

Chapter 5

Conclusion

This thesis set out to validate the performance of the reliable fraction, a novel intra-session reliability metric, to measure data quality for pre-surgical maps acquired with MEG. I discussed the need for quantitative quality assurance measures in MEG and the limitations of the current clinical standard metric, goodness of fit. In particular, goodness of fit is not an effective measure of data quality when the equivalent current dipole model is not an appropriate method of source localization. This is true for cases where several regions of the brain may be activated, such as language mapping paradigms. ROC-r, my group's framework for reliability analysis, splits task-based data in half in order to compare the agreement between activation maps within a single MEG scan. This intra-session reliability is quantified by the reliable fraction, which we put forward as a quality metric suitable for all activation maps. I also drew attention to two other possible intra-session reliability measures, the Dice and Pearson coefficients, which could be calculated using an analogous split-half approach. I hypothesized that reliable fraction would outperform both the clinical standard, goodness of fit, and the Dice and Pearson coefficients in measuring the change in data quality for MEG activation maps acquired in 'good' and 'poor' conditions.

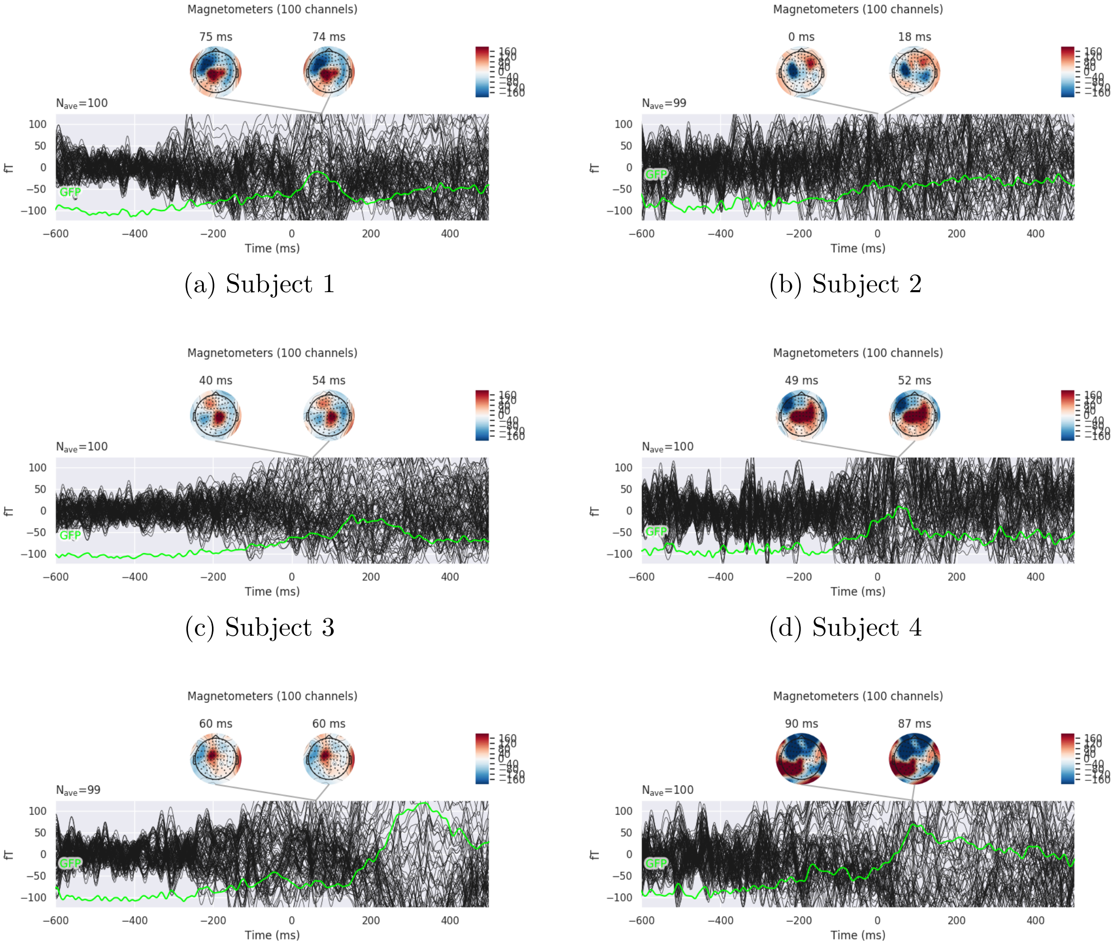
I acquired data using pre-surgical mapping paradigms to localize the primary somatosensory, motor, visual, and auditory cortices in twelve healthy subjects. For each subject and paradigm, I acquired scans in both a standard 'good' condition and a 'poor' condition simulating common MEG data quality issues. This diverse dataset highlighted the need for rigorous quality assurance for MEG data. I observed significant variations in data quality across mapping paradigms and found that even in the good condition, the expected motor and auditory responses were inconsistently elicited across all subjects. While high quality somatosensory and visual data was successfully acquired in the good condition, different effects of data quality manipulation were observed for both paradigms. The MEG processing pipeline effectively mitigated the introduction of a magnetic artefact in the poor quality somatosensory data, and no significant difference between good and poor somatosensory data was found by any metric at the group-level (Table 3.3). This demonstrated the suitability of the proposed intra-session reliability metrics for measuring the quality of activation maps rather than the quality of raw data, but did not provide an opportunity to compare metric sensitivity to changes in data quality. This comparison was performed for the visual mapping data, where only data acquired in the good condition successfully localized the primary visual cortex. At the group-level, goodness of fit measured the largest change in data quality at the visual response peak ($p = 5.2 \times 10^{-5}$), followed by the reliable fraction ($p = 1.4 \times 10^{-4}$), the Pearson correlation ($p = 1.0 \times 10^{-3}$), and the Dice coefficient ($p = 1.3 \times 10^{-3}$). However, variability across subjects in the quality of 'good' visual data necessitated a comparison of metric performance at the single-subject level using a nonparametric bootstrap. Although a one-way ANOVA across the group found no significant difference in the mean change in data quality measured by each metric, ranking the metrics by the significance of the change in data quality measured for each subject, I found that goodness of fit ranked best across all subjects (rank 19), followed by reliable fraction (rank 24), the Dice coefficient (rank 33), and the Pearson coefficient (rank 34).

Overall, my results validate the performance of the reliable fraction for use as a quality assurance metric for MEG data. Of the three potential intra-session reliability metrics, the reliable fraction appears most sensitive to changes in MEG activation map quality, and performs well in comparison with goodness of fit. My thesis justifies the further exploration of the reliable fraction as a quality assurance metric for cases of distributed brain activity. Moreover, my thesis provides much-needed context for future studies of quality metric performance in MEG data. I have identified the shortcomings of data acquisition specific to this study as well as the limitations of interpreting data quality scores in the absence of a gold standard metric. I have also suggested several areas for future study in order to better understand the behaviour of reliable fraction as well as other intra-session reliability metrics going forward.

Beyond their immediate implications, I believe that my results reveal the need for a better understanding of the typical data quality achieved in MEG scans. Not only is there no widely established threshold for rejecting data on the basis of poor quality for any of the quantitative metrics I have discussed here, but there is also little emphasis on quantitative, operator-independent quality assurance measures in the MEG literature. Going forward, establishing a rigorous basis for data quality reporting and comparison across MEG studies and imaging centres will be important for the transparency and reproducibility of MEG research as a whole. In the clinical context, such measures will be essential to ensure high quality pre-surgical mapping data is acquired for each patient. My research supports the use of the reliable fraction as one such metric, and provides a frame of reference for its application across a number of mapping paradigms.

Appendix A

Supplementary Figures



(a) Subject 1

(b) Subject 2

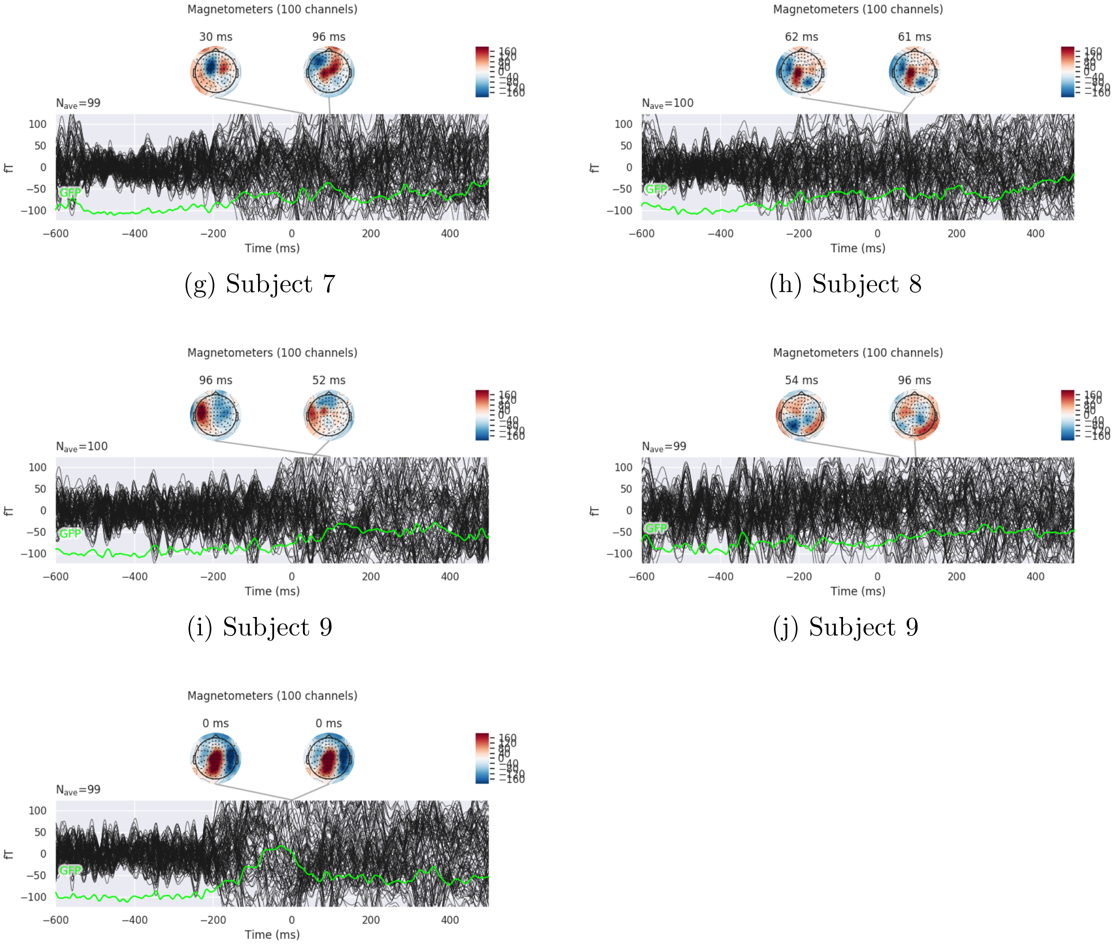
(c) Subject 3

(d) Subject 4

(e) Subject 5

(f) Subject 6

Figure A.1: Event-related fields measured for each subject in the good condition for motor data. Global field power and topographies at automatically selected peaks are also shown.



(g) Subject 7

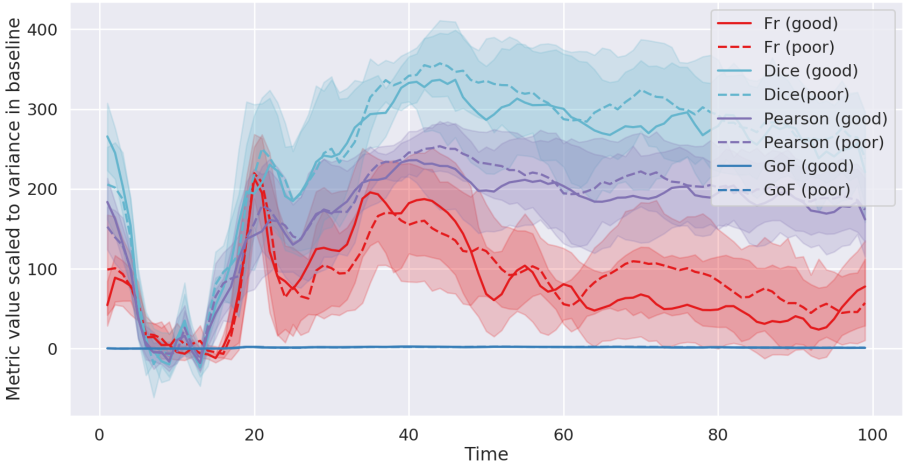
(h) Subject 8

(i) Subject 9

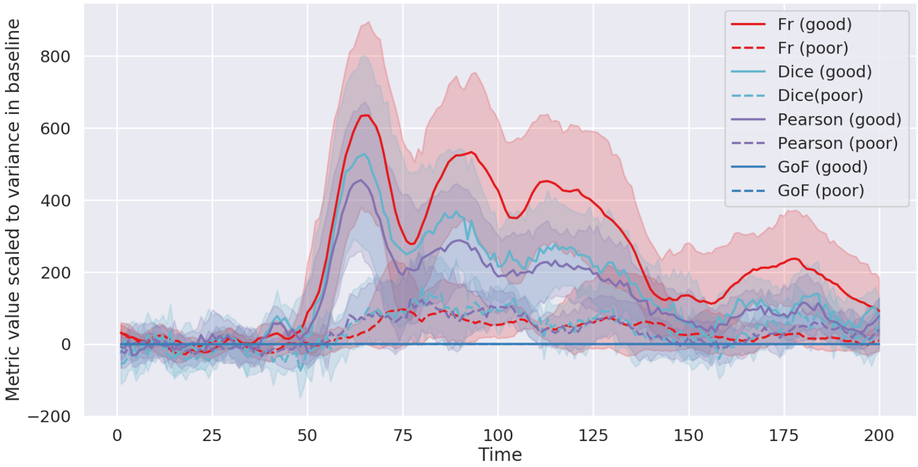
(j) Subject 9

(k) Subject 11

Figure A.1: Event-related fields measured for each subject in the good condition for motor data. Global field power and topographies at automatically selected peaks are also shown.

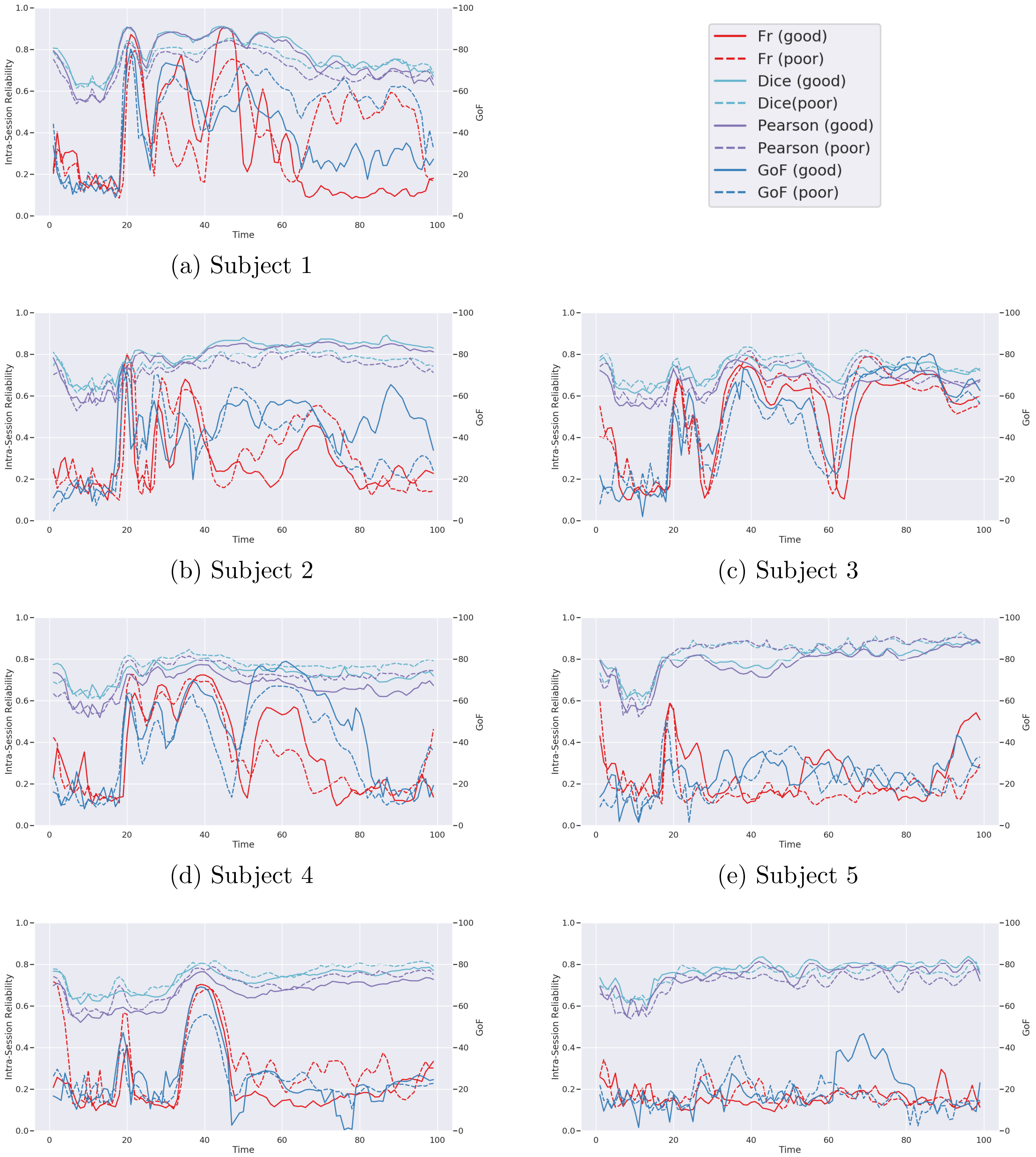


(a) Somatosensory quality scores, scaled



(b) Visual quality scores, scaled

Figure A.2: Metric time series in grand average, baseline subtracted and scaled by the variance in the baseline interval (Table 2.2).



(a) Subject 1

(b) Subject 2

(c) Subject 3

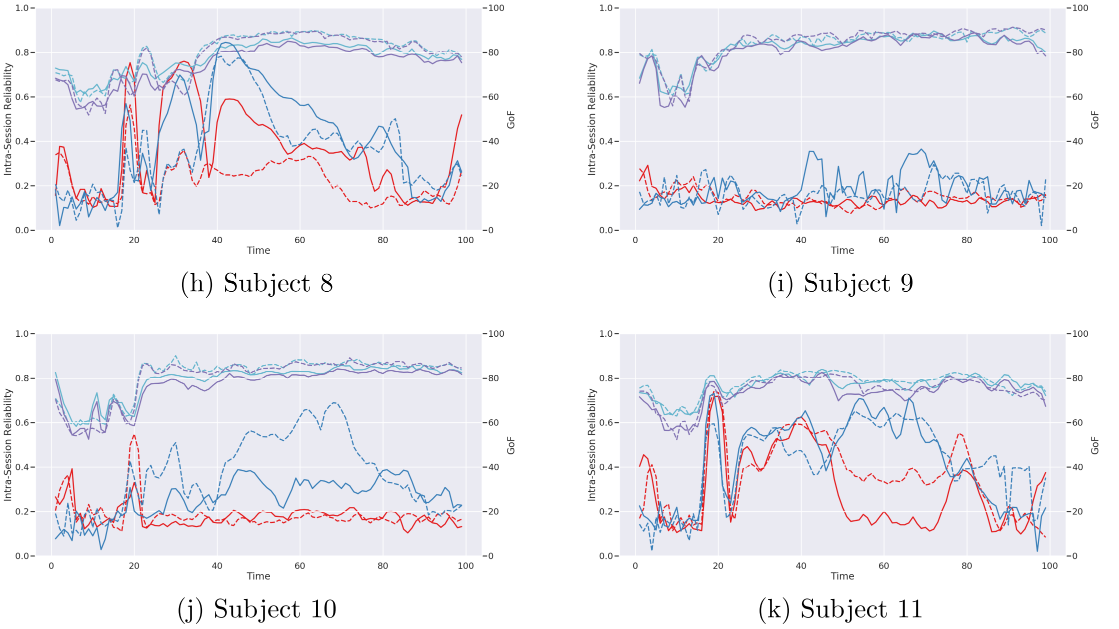
(d) Subject 4

(e) Subject 5

(f) Subject 6

(g) Subject 7

Figure A.3: Metric time series measuring somatosensory data quality for each subject.



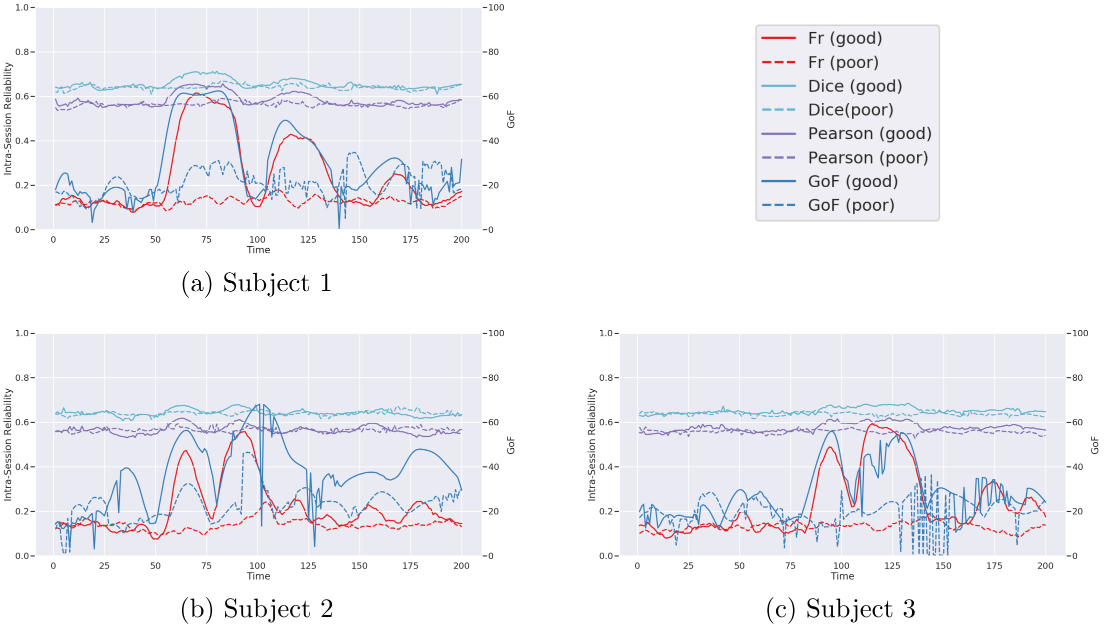
(h) Subject 8

(i) Subject 9

(j) Subject 10

(k) Subject 11

Figure A.3: Metric time series measuring somatosensory data quality for each subject.

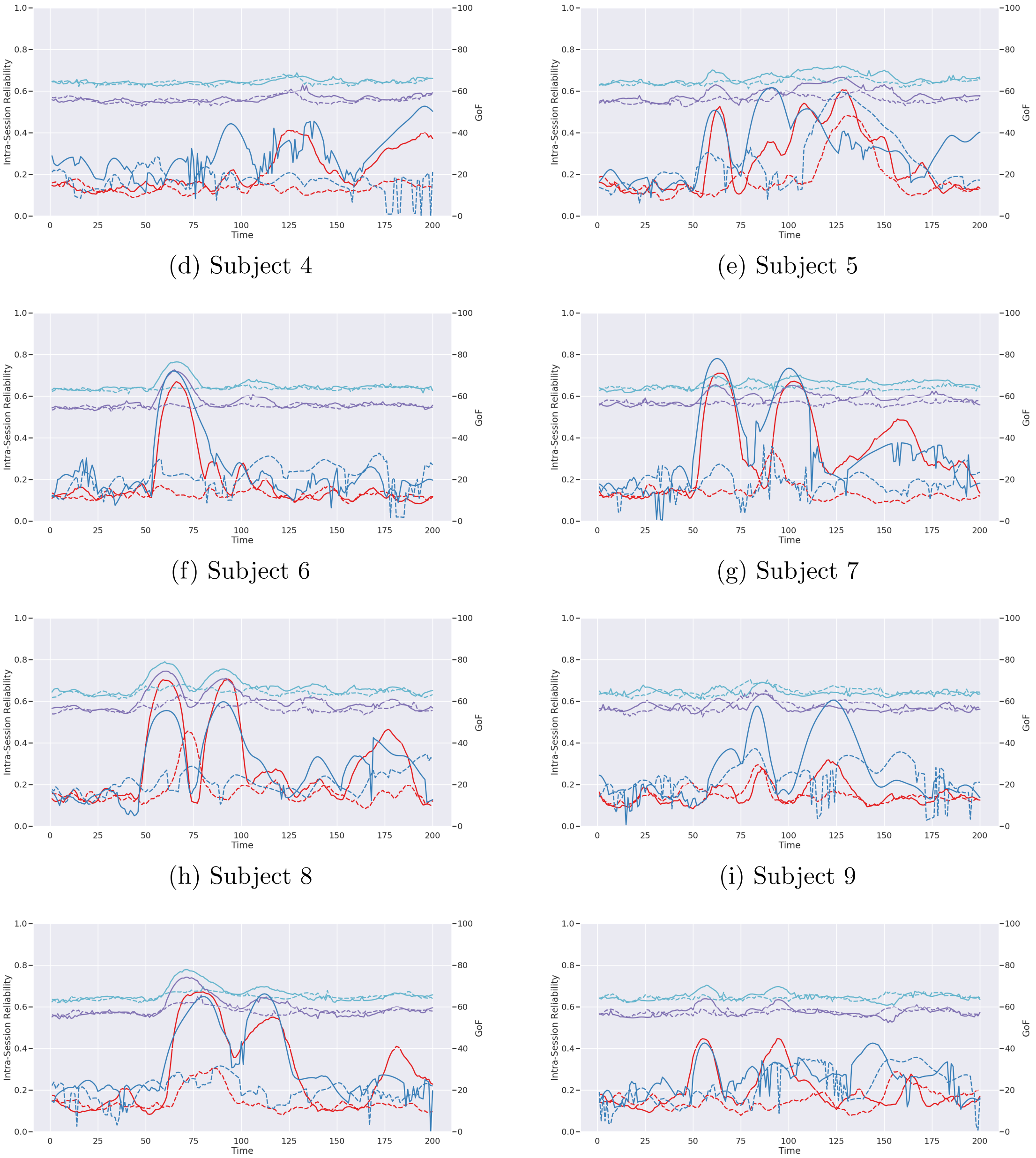


(a) Subject 1

(b) Subject 2

(c) Subject 3

Figure A.4: Metric time series measuring visual data quality for each subject.



(d) Subject 4

(e) Subject 5

(f) Subject 6

(g) Subject 7

(h) Subject 8

(i) Subject 9

(j) Subject 10

(k) Subject 11

Figure A.4: Metric time series measuring visual data quality for each subject.

Bibliography

[1] S. Sagar *et al.* Functional brain mapping: overview of techniques and their application to neurosurgery. *Neurosurg Rev.*, pages 1–9, 2018.

[2] T. Inoue *et al.* Accuracy and Limitation of Functional Magnetic Resonance Imaging for Identification of the Central Sulcus: Comparison with Magnetoencephalography in Patients with Brain Tumors. *Neuroimage*, 10:738–748, 1999.

[3] H. Schiffbauer *et al.* Preoperative magnetic source imaging for brain tumor surgery: a quantitative comparison with intraoperative sensory and motor mapping. *J Neurosurg.*, 97(6):1333–1342, 2002.

[4] W. Gaetz *et al.* Presurgical localization of primary motor cortex in pediatric patients with brain lesions by the use of spatially filtered magnetoencephalography. *Neurosurgery*, 64(3 Suppl):ons177–185, 2009.

[5] P. Grummich *et al.* Combining fMRI and MEG increases the reliability of presurgical language localization: A clinical study on the difference between and congruence of both modalities. *Neuroimage*, 32:1793–1803, 2006.

[6] A. Korvenoja *et al.* Sensorimotor cortex localization: comparison of magnetoencephalography, functional MR imaging, and intraoperative cortical mapping. *Radiology*, 241(1):213–222, 2006.

[7] O. Ganslandt *et al.* Functional neuronavigation with magnetoencephalography: outcome in 50 patients with lesions around the motor cortex. *J Neurosurg.*, 91:73–79, 1999.

[8] A. Niranjan *et al.* Preoperative Magnetoencephalographic Sensory Cortex Mapping. *Stereotact Funct Neurosurg.*, 91:314–322, 2013.

[9] F. Darvas, D. Pantazis, E. Kucukaltun-Yildirim, and R.M. Leahy. Mapping human brain function with MEG and EEG: methods and validation. *Neuroimage*, 23:S289–S299, 2004.

[10] R.C. Burgess *et al.* American Clinical Magnetoencephalography Society Clinical Practice Guideline 2: Presurgical Functional Brain Mapping Using Magnetic Evoked Fields. *J Clin Neurophysiol.*, 28(4):355–361, 2011.

[11] M. Hämäläinen *et al.* Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev Mod Phys.*, 65(2):413–497, 1993.

[12] D. Cheyne, A.C. Bostan, W. Gaetz, and E.W. Pang. Event-related beamforming: A robust method for presurgical functional mapping using MEG. *Clin Neurophysiol.*, 118:1691–1704, 2007.

[13] M. Jas *et al.* A Reproducible MEG/EEG Group Study With the MNE Software: Recommendations, Quality Assessments, and Good Practices. *Front Neurosci.*, 12:530, 2018.

[14] J. Gross *et al.* Good practice for conducting and reporting MEG research. *Neuroimage*, 65:349–363, 2013.

[15] A. Puce and M.S. Hämäläinen. A Review of Issues Related to Data Acquisition and Analysis in EEG/MEG Studies. *Brain Sci.*, 7(6):E58, 2017.

[16] R. Hari *et al.* IFCN-endorsed practical guidelines for clinical magnetoencephalography (MEG). *Clin Neurophysiol.*, 129(8):1720–1747, 2018.

[17] S. Taulu and R. Hari. Removal of Magnetoencephalographic Artifacts With Temporal Signal-Space Separation: Demonstration With Single-Trial Auditory-Evoked Responses. *Hum Brain Mapp.*, 30:1524–1534, 2009.

[18] V. Jousmäki and R. Hari. Cardiac artifacts in magnetoencephalogram. *J Clin Neurophysiol.*, 13(2):172–176, 1996.

[19] A. Antervo *et al.* Magnetic fields produced by eye blinking. *Electroencephalogr Clin Neurophysiol.*, 61(4):247–253, 1985.

[20] A. Stolk, A.Todorovic, J.M. Schoffelen, and R. Oostenveld. Online and offline tools for head movement compensation in MEG. *Neuroimage*, 68:39–48, 2013.

[21] J. Velmurugan, S. Sinha, and P. Satishchandra. Magnetoencephalography recording and analysis. *Ann Indian Acad Neurol.*, 17(Suppl 1):S113–S119, 2014.

[22] E.W. Pang *et al.* Intraoperative Confirmation of Hand Motor Area Identified Preoperatively by Magnetoencephalography. *Pediatr Neurosurg.*, 44:313–317, 2008.

[23] R. Sharma *et al.* Magnetoencephalography in children: Routine clinical protocol for intractable epilepsy at the hospital for sick children. *Int Conf Ser.*, 1300:685–688, 2010.

[24] E.M. Castillo *et al.* Integrating sensory and motor mapping in a comprehensive MEG protocol: Clinical validity and replicability. *Neuroimage*, 21:973–983, 2004.

[25] J. Solomon, S. Boe, and T. Bardouille. Reliability for non-invasive somatosensory cortex localization: Implications for pre-surgical mapping. *Clin Neurol Neurosurg.*, 139:224–229, 2015.

[26] D. Lee *et al.* Reliability of language mapping with magnetic source imaging in epilepsy surgery candidates. *Epilepsy Behav.*, 8:742–749, 2006.

[27] D.M. Groppe, S. Makeig, and M. Kutas. Identifying reliable independent components via split-half comparisons. *Neuroimage*, 45(4):1199–1211, 2009.

[28] M.T.R. Stevens *et al.* Thresholds in fMRI studies: Reliable for single subjects? *J Neurosci Methods*, 219(2):312–323, 2013.

[29] T. Stevens *et al.* Fully automated quality assurance and localization of volumetric MEG for single-subject mapping. *J Neurosci Methods*, 266:21–31, 2016.

[30] S. McLeod. Investigating reliability as a tool for patient-specific pipeline selection. Master's thesis, Dalhousie University, Halifax, Nova Scotia, 2017.

[31] C.M. Bennett and M.B. Miller. How reliable are the results from functional magnetic resonance imaging? *Ann N Y Acad Sci.*, 1191(1):133–155, 2010.

[32] L.R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26:297–302, 1945.

[33] S.A. Rombouts *et al.* Test-retest analysis with functional MR of the activated area in the human visual cortex. *AJNR Am J Neuroradiol.*, 18:1317–1322, 1997.

[34] K.H. Zhou *et al.* Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. *Acad Radiol*, 11(2):178–189, 2004.

[35] P. Aljabar *et al.* Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *Neuroimage*, 46(3):726–738, 2009.

[36] S.M. Wilson *et al.* Validity and reliability of four language mapping paradigms. *Neuroimage Clin.*, 16:399–408, 2016.

[37] J. Liu *et al.* Correlation and agreement: overview and clarification of competing concepts and measures. *Shanghai Arch Psychiatry*, 28(2):115–120, 2016.

[38] R.C. Oldfield. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, 9(1):97–113, 1971.

[39] A.M. Dale, B. Fischl, and M.I. Sereno. Cortical surface-based analysis - I. Segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194, 1999.

[40] B. Fischl and A.M. Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20):11050–11055, 2000.

[41] B. Fischl, A. Liu, and A.M. Dale. Automated manifold surgery: Constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Trans Med Img.*, 20(1):70–80, 2001.

[42] B. Fischl, M.I. Sereno, and A.M. Dale. Cortical surface-based analysis - II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9(2):195–207, 1999.

[43] B. Fischl *et al.* High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum Brain Mapp.*, 8(4):272–284, 1999.

[44] B. Fischl *et al.* Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

[45] B. Fischl *et al.* Automatically parcellating the human cerebral cortex. *Cereb Cortex*, 14(1):11–22, 2004.

[46] B. Fischl. FreeSurfer. *Neuroimage*, 62(1):774–781, 2012.

[47] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.*, 10:626–634, 1999.

[48] T. Bardouille *et al.* Variability and bias between magnetoencephalography systems in non-invasive localization of the primary somatosensory cortex. *Clin Neurol Neurosurg.*, 171:63–69, 2018.

[49] B. van Veen, W. van Drongelen, M. Yuchtman, and A. Suzuki. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Trans Biomed Eng.*, 44:867–880, 1997.

[50] F. Darvas *et al.* Investigations of dipole localization accuracy in MEG using the bootstrap. *Neuroimage*, 25(2):355–368, 2005.

[51] B. Efron. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, 1981.

[52] R. Beran. Diagnosing Bootstrap Success. *Ann Inst Stat Math.*, 49(1):1–24, 1997.

[53] H. Shigeto *et al.* Visual evoked cortical magnetic responses to checkerboard pattern reversal stimulation: a study on the neural generators of N75, P100 and N145. *J Neurol Sci.*, 156(2):186–194, 1998.

[54] N. Nakasato *et al.* Functional localization of bilateral auditory cortices using an MRI-linked whole head magnetoencephalography (MEG) system. *Clin Neurophysiol.*, 94(3):183–190, 1995.

[55] M.J. Brookes *et al.* Beamformer reconstruction of correlated sources using a modified source model. *Neuroimage*, 34(4):1454–1465, 2007.

[56] A. Hillebrand, P. Fazio, J.C. de Munck, and B.W. van Dijk. Feasibility of clinical Magnetoencephalography (MEG) functional mapping in the presence of dental artefacts. *Clin Neurophysiol.*, 124:107–113, 2013.

[57] E.G. Gross, H.G. Vaughan Jr., and E. Valenstein. Inhibition of visual evoked responses to patterned stimuli during voluntary eye movements. *Electroencephalogr Clin Neurophysiol.*, 22(3):204–209, 1967.

[58] B.A. Wandell, S.O. Dumoulin, and A.A. Brewer. Visual Field Maps in Human Cortex. *Neuron*, 56(2):366–383, 2007.

[59] G. Perry *et al.* Retinotopic mapping of the primary visual cortex - a challenge for MEG imaging of the human cortex. *Eur J Neurosci.*, 34:652–661,, 2011.

[60] W. Skrandies. Global Field Power and Topographic Similarity. *Brain Topogr.*, 3(1):137–141, 1990.

[61] D.F. Sobel *et al.* Locating the central sulcus: comparison of MR anatomic and magnetoencephalographic functional methods. *AJNR Am J Neuroradiol.*, 14(4):915–925, 1993.

[62] T. Hedrich *et al.* Comparison of the spatial resolution of source imaging techniques in high-density EEG and MEG. *Neuroimage*, 157:531–544, 2017.

[63] M. Hirata *et al.* Language dominance and mapping based on neuromagnetic oscillatory changes: comparison with invasive procedures. *J Neurosurg.*, 112(3):528–538, 2010.

[64] S.M. Stufflebeam. Clinical Magnetoencephalography for Neurosurgery. *Neurosurg Clin N Am.*, 22:153–167, 2011.

[65] S.-H. Cha, C. Tappert, and S. Yoon. Enhancing Binary Feature Vector Similarity Measures. *"J Pattern Recognit Res.*, 1:63–77, 2006.

[66] G. Wei and H. Gao. The Generalized Dice Similarity Measures for Picture Fuzzy Sets and Their Applications. *Informatica*, 29(1):107–124, 2018.

[67] S.C. Strother *et al.* The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *Neuroimage*, 15(4):747–771, 2002.

[68] D.A. Low and J.F. Dempsey. Evaluation of the gamma dose distribution comparison method. *Med Phys.*, 30(9):2455–2464, 2003.

[69] C.R. Genovese, D.C. Noll, and W.F. Eddy. Estimating test-retest reliability in functional MR imaging. I: Statistical methodology. *Magn Reson Med.*, 38(3):497–507, 1997.

[70] R. Maitra, S.R. Roys, and R.P. Gullapalli. Test-retest reliability estimation of functional MRI data. *Magn Reson Med.*, 48(1):62–70, 2002.