**Experts’ accuracy in identification of the language network from resting state fMRI in patients with brain tumors.**

**Humans are inaccurate in identifying the language network from resting state fMRI in patients with brain tumors**

**Can fMRI experts accurately identify the language network from resting state fMRI in patients with brain tumors?**

**Identification of the language network from resting state fMRI in patients with brain tumors: how accurate are experts?**

Sachin K. Gujar MBBS1 , Jistupa Wongsripuemtet MD1, 2Thammasin Ingviya MD, Daniel Ryan MD. Shruti Agarwal PhD1, Martin Lindquist PhD3, Brian Caffo PhD3, Jay J. Pillai MD1, Haris I. Sair MD1\*

1 Division of Neuroradiology, The Russell H. Morgan Department of Radiology and Radiological Science

2 Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

2 Department of Family Medicine and Preventive Medicine, Prince of Songkla University, Thailand

3 Department of Biostatistics, Johns Hopkins University, Baltimore MD.

\*corresponding author:

Haris I. Sair MD

Johns Hopkins University School of Medicine

600 North Wolfe Street, Phipps B-112A

Baltimore, MD 21287

[hsair1@jhmi.edu](mailto:hsair1@jhmi.edu)

T: 443-287-0386/410-955-2353

F: 410-614-1213

Funding: Grant ID: RSCH1420: Grant Program: RSNA Research & Education Foundation Carestream Health/RSNA Research Scholar Grant

**INTRODUCTION**

Resting state functional magnetic resonance imaging (rs-fMRI) has emerged as a novel tool to analyze brain function. In contrast to traditional task-fMRI, no explicit task is required in rs-fMRI while Blood Oxygen Level Dependent (BOLD) images are acquired. Owing to an assortment of naturally occurring fluctuations of BOLD activity in various regions of the brain, a set of intrinsic brain networks can be identified by examining spatially distinct however temporally synchronous BOLD signals at rest. The exact number of intrinsic brain networks is not determined, however a relatively consistent set of networks has been reliably demonstrated in many studies (e.g. see Barkhof F, Haller S, Rombouts Serge Radiology July 2014 for a review).

Although clinical applications of rs-fMRI are currently limited partly due to the relatively high variability of brain networks when examined at the single subject level (Airan et al), considerable efforts have been made to translate this technique into clinical practice. In this realm, the most promising utilization of rs-fMRI currently seems to be in the domain of presurgical brain mapping (Lee Megan et al Topics in Magnetic Resonance Imaging 25(1) Feb 2016 11-18). While task-fMRI has been successfully used to identify critical brain regions for preoperative planning, task-fMRI requires compliance in performing behavioral paradigms necessary for determining brain activation. This compliance may be limited or absent in many cases, for example in the pediatric population or in the elderly, in patients with language barriers, or in those with physical debilitation limiting movement. Given lack of a task, therefore, the use of rs-fMRI is attractive as an alternative technique to assess brain function. Furthermore, studies have shown that rs-fMRI may require less time compared to task-fMRI to obtain comparable data.

As one example, rs-fMRI derived motor networks have been shown to be comparable to motor regions activated during task-fMRI, as well as motor regions identified during direct cortical stimulation (DCS) (Qiu TM et al Acta Neurochir 2014 Dec;156(12):2295-302). The concordance between rs-fMRI derived language networks and task-fMRI activated language regions is more variable; earlier reports suggested moderate concordance (Tie et al), however more recent reports demonstrate high subject-level variability (Sair, Duffau). While motor networks are relatively easily identified from rs-fMRI due to the relative invariance of the anatomical-functional relationship of the motor system across individuals, identification of the language network may be more challenging due to relatively high variance in localization of language areas across individuals (Sanai et al) and the similarity of elements of the language network with the spatial distribution of other networks such as the ventral attention network (Corbetta M, Patel G, Shulman GL Neuron 2008 Mayb 8;58(3):306-24).

If rs-fMRI is to be used as an *alternative* to task-fMRI, the accuracy of identification of the language network solely from rs-fMRI must be assessed when presented with multiple network correlation maps, which is the output generated from commonly used methods of rs-fMRI analysis, such as Independent Component Analysis (ICA). While automated methods of network identification are currently being developed, we were interested in the ability of humans to correctly identify the language network. We examined the human accuracy of identifying subject-level rs-fMRI language networks using their task-fMRI language activation maps as a gold standard. We hypothesize that there is low accuracy in human identification of rs-fMRI language networks when language rrelated task-fMRI activation maps are unavailable.

**MATERIALS AND METHODS**

**Participants**

The radiology information system (RIS) was interrogated for any patient who underwent fMRI for presurgical brain mapping between 1/1/2009 and 7/1/2015. 79 patients with intracranial neoplasms were identified for whom both language task-fMRI and rs-fMRI were available for the same imaging session. 21 patients had a prior history of an invasive intracranial procedure (surgery and/or biopsy) and were excluded to minimize confounding the effect of susceptibility. Because individual language paradigms commonly only activate subsets of the global language system, we only included patients who had completed three different language tasks (Silent Word Generation [SWG], Sentence Completion [SC], and Rhyming) during the same session. Patients who had suboptimal activation on any one of the three language task-fMRIs, assessed subjectively as is routine in clinical fMRI cases, were excluded (15 patients). After exclusion, data from 43 subjects were available for analysis.

**Handedness**

The Edinburgh Handedness Inventory was used to determine patient handedness (Oldfield RC neuropsychologia 1971).

**Imaging**

Images were acquired on a 3.0 Tesla Siemens Trio Tim MRI (Siemens Medical Solutions, Erlangen, Germany) using a 12-channel head matrix coil. For both task-fMRI and rs-fMRI, T2\* weighted BOLD images were acquired using 2D gradient echo echo-planar imaging: TR=2000ms, TE=30ms, flip angle = 90 degrees, FOV= 24 cm, acquisition matrix = 64 x 64 x 33, slice thickness = 4mm, slice gap = 1 mm, interleaved acquisition. Instructions for rs-fMRI were: keep your eyes closed, don’t move, and don’t think of anything in particular. 180 volumes were acquired for rs-fMRI (6 minutes). 3 dimensional T1 weighted structural images were also obtained: TR=2300ms, TI=900ms, TE=3.5ms, flip angle 9 degrees, FOV=24cm, acquisition matrix = 256x256x176, slice thickness 1mm.

**Task-fMRI Paradigms**

As is routine at our institution, we instructed the patients and performed practice sessions outside the scanner prior to fMRI to ensure understanding of the tasks. Real-time fMRI maps were monitored by the neuroradiologist administering the study to assess for global data quality. Any task with suboptimal activation assessed subjectively were repeated per our protocol; for final analysis the single best run of each task was chosen. The Prism software suite was used for stimulus presentation (Prism Clinical Imaging, Elm Grove, Wisconsin). A block design of either 30s (Rhyming) or 20s (SWG, SC) alternative task and control block was used for imaging time of 3 (Rhym) or 4 minutes (SWG, SC).

**Image Processing and Analysis**

Statistical Parametric Mapping (SPM) version 8 (Wellcome Department of Imaging Neuroscience, University College London, UK) and custom MATLAB (Mathworks, Natick, MA) scripts were used to process the fMRI.

**Task-fMRI**

Task-fMRI underwent slice timing correction (STC) followed by motion correction (MC). Images were normalized to a Montreal Neurological Institute (MNI-152) template and spatially smoothed using a 6mm full width at half maximum (FWHM) Gaussian kernel. A general linear model analysis was performed using a canonical hemodynamic response function (HRF) convolved with the boxcar function for each task, without utilizing model derivatives or global intensity normalization. A 128 second high pass filter was used. An autoregressive model (AR1) was used to account for temporal auto-correlations. No confound matrix was used. A contrast design matrix set to detect activation across all three tasks compared to rest was used for each subject. SPM T-contrast maps were generated without clustering or multiple comparison correction as is typical for our clinical cases.

**Rs-fMRI**

Rs-fMRI underwent STC followed by MC. The ArtRepair toolbox [Mazaika et al., 2009] was then used to detect volumes with large shifts in global average signal intensity related to scan-to-scan motion; both the outlier volumes and additional volumes recommended for de-weighting were tagged for subsequent removal from analysis (i.e. for “scrubbing”). Rs-fMRI was linearly detrended, and following co-registration of rs-fMRI and T1 weighted images, physiological nuisance regression of rs-fMRI was performed utilizing the CompCor method [Behzadi et al., 2007] using signal extracted from eroded white matter and cerebrospinal fluid masks. After bandpass filtering from 0.01 to 0.1 Hz, smoothing was performed with a 6mm FWHM Gaussian kernel. Finally, images tagged by ArtRepair were removed (“scrubbed”) from the rs-fMRI volumes.

The Group Independent Component Analysis of fMRI Toolbox (GIFT, Medical Image Analysis Lab, <http://mialab.mrn.org/software/gift>) was used to generate ICA maps for each subject using 20 (ICA20) and 50 (ICA50) target components, utilizing the InfoMax algorithm with ICASSO set at 5 repeats. The language network was identified by first sorting the ICA components using multiple regression in GIFT with the task-fMRI SPM T-maps as the reference template. Then, the component that demonstrated the highest spatial overlap with the task-fMRI maps localized to Broca and Wernicke activation was selected as the rs-fMRI language network. Of note, in all cases there was one ICA component that best represented the primary language network for both ICA orders. This ICA component was labeled as the rs-fMRI language network map for each subject.

**Rs-fMRI language network identification**

Three participants independently reviewed the rs-fMRI ICA maps to identify the potential language network. All were blinded to the task-fMRI activation maps. The reviewers ranged in fMRI experience: 17 years, 2 years, and 6 years. Images were presented to each reviewer using the orthogonal viewer in FSLView (v 3.1, FMRIB, Oxford, UK) with each subject’s rs-fMRI maps overlayed onto their T1 weighted images. Review was performed independently for ICA20 and ICA50. The reviewers were allowed to modify image contrast and thresholds. Reviewers ranked up to three top choices for the candidate rs-fMRI language map, as well as their confidence in their assessment ranging from 1 (highly confident) to 5 (not confident).

**Statistical tests**

For descriptive statistics, percent correct answers were calculated for each rs-fMRI choice (first, second, third) for each ICA order (ICA20 and ICA50) from the proportion of correct answers to the total number of 43 correct choices. Corresponding 95% confidence intervals for each proportion were calculated based on a binomial distribution.

To determine whether higher number of choices improves accuracy and account for the dependency of inter-rater reading of identical images, a one-sided sign test was used as a non-inferiority test between the three different selection scenarios (top choice, top two choice, and top three choices). (Reference).

To compare the accuracy between ICA20 and ICA50, a two-sided sign test for each selection scenario and reviewers was used. To measure the concordance of rs-fMRI language network map identification between and within the three reviewers, intra-rater and inter-rater Kappa statistics were calculated.

For inferential statistics, univariate and multivariate logistic regression with robust variance was utilized to assess the association between accuracy (percent correct) and factors of ICA order, reviewers, and number of choices. Univariate and multivariate odds ratios were calculated and reported with their 95% confidence interval from robust variances.

To determine whether increasing one choice from top 2 to top 3 choices would help improve accuracy or not, logistic regression using top two choices instead of top choice was used (Supplementary table 4). We further explore a probable reason of the inter-rater differences by substituting reviewer variable with years of fMRI experience (Supplementary table 5).

**A Fisher’s exact test was performed for each reviewer and ICA order to determine the association between handedness and percent correct, utilizing the top choice, top two choices, and top three choices**.

**To determine whether there was an association between the confidence rating and the correct choice,** Mantel-Haentzel Chi-square test for trend and Fisher’s exact test **was performed.**

All statistical analyses were performed in R version 3.2.3. The statistically significant level were determined at p-value below 0.05.

We aimed to discover the association between the probability of correct identification of the language network and some potential factors including the reviewers, ICA type and each selection scenario.

This study can be seen as the cross-sectional study that each subject received multiple reviews in different scenarios. Let denote the correctness of identification for subject by reviewer using ICA type and top choices . takes values 1 or 0, indicating whether the identification is correct or not. The cross-sectional structure motivates the generalized mixed effect model as follows,

where follows the normal distribution independently, denoting the random intercept. The covariates represent the fixed effect of reviewer . The effect of ICA20 and ICA50 is denoted by ,respectively. The covariates denote the effect of three different selection scenarios (top one, two and three). The term denotes the handedness for subject . To ensure the identifiability of the model, we introduce constraints .

We made the statement by the likelihood ratio test that there were no significant interaction terms between each pairs of the fixed effect. We also found that the effect of ICA type and handedness does not help to improve the model fitting significantly, so we excluded it from the model. In other words, there was no significant effect for the ICA type and handedness given other variable. We also performed the test for necessity of random intercept effect ( versus ) by parametric bootstrapping. We found the significant improvement (p-value = 0) by fitting the model accounting for the subject-specific random effect.

The final model we fit is the following,

where follows the normal distribution independently.

The conditional logistic regression model given subject-specific effect was constructed to explore the association between the probability of correct identification and the confidence rating adjusted for the effect of each reviewers and ICA types. It turned out there is no significant association. To measure the concordance of rs-fMRI language network map identification between and within the three reviewers, intra-rater and inter-rater Kappa statistics were calculated.

All statistical analyses were performed in R version 3.6.2.

**Results**

Patients comprised of 29 males and 14 females with mean age of 41 years (minimum 18 and maximum 69). 33 patients are right-handed, 8 left handed, and 2 ambidextrous.

Tables 1 and 2 illustrate the marginal association between correctness percentage and the reviewers, ICA type, selection scenarios. Table 1 demonstrates the percent correct for each reviewer for ICA20 and ICA50, assessing their top choice (1), the top two choices (1+2), or the top three choices (1+2+3). Reviewer 1, with 17 years of fMRI experience, demonstrated the highest overall accuracy with 72% [95% confidence interval (CI): 66.1-77.4%] correct responses across all conditions. Reviewers 2 and 3, with 2 and 6 years of experience respectively, had overall similar percent correct responses (50% [95% CI: 44.1-56.6%] and 55% [95% CI: 48.7-61.2%]). For each reviewer, highest accuracy was obtained using ICA50 and top three choices (81%, 65% and 60% for reviewer 1, 2, and 3 respectively). Conversely, the lowest accuracy was obtained also using ICA50 however limiting each reviewer to the top choice (58%, 35%, and 42%).

Significant differences in accuracy were seen between selecting only the top choice vs selecting top three choices (Table 2), with the exception of ICA20 for reviewer 3. In both ICA conditions and across all raters, no significant difference in accuracy was seen between selection of top two choices vs. top three choices. Comparing single top choice to top two choices, mixed findings were noted between ICA orders and raters. Across the raters, inclusion of three choices compared to one choice improved accuracy by 13% for ICA 20 and 24% for ICA50. When examining the converse problem – whether ICA order increases accuracy – no significant difference was found for any of the reviewers in any of the three choice conditions (Supplementary Table 2).

Kappa statistics demonstrated overall fair concordance between reviewer 1 vs. reviewers 2 and 3 (Kappa=0.35-0.40). Poor concordance was observed between reviewers 2 and 3 (Kappa =0.16). When limiting analysis to the top choice, again reviewer 1 demonstrated fair concordance with reviewers 2 and 3 (Kappa=0.21-0.47) and there was poor concordance between reviewers 2 and 3 (Kappa =0.028). Similar findings are seen when analyzing concordance of any of the top three choices (reviewer 1 vs 2 and 3 Kappa = 0.31-0.40, and reviewer 2 vs 3 Kappa = 0.12).

Based on the model (2), conditional on the subject, the odds of correct identification for reviewer 2 is 0.26 [95% CI: 0.17-0.41] times the odds of reviewer 1 given the same selection scenario. Similarly, the odds ratio of reviewer 3 against reviewer 1 is 0.34 [95% CI: 0.22-0.53] and the odds ratio of odds ratio of reviewer 3 against reviewer 2 is 1.31 [95% CI: 0.86-1.99]. Note that there is no significant difference between odds of reviewer 2 and reviewer 3 given the same selection scenario conditional on the subject.

From the same model, the odds of correct identification for selecting the top 2 choices is 2.39 [95% CI: 1.56-3.69] times the odds of top 1 choice given the same reviewer and conditional the subject. The odds ratio of selecting top 3 choices against top 1 choice is 3.05 [95% CI: 1.98-4.76], and the odds ratio of including top 3 choices against top 2 is 1.28 [95% CI: 0.83-1.98]. The difference between odds of selecting top 3 and top 2 choice is not statistically significant.

To assess the performance of the model, we make the plot of the receiver operating characteristic (ROC) curve (Figure 2) and compute the area under the curve (AUC). The ROC curve was constructed by the leave-one-out procedure that we sequentially treated one subject as the test set and leave the data of the rest subject as the train set for the model fitting. The AUC = 0.849 [95% CI: 0.823 – 0.875] indicates an 84.9% chance that the model will correctly identify the language network based on the rs-fMRI.

Compared to selecting one top choice, the inclusion of the top 2 and top 3 choices improved the odds of getting the correct response by 1.87 [95% CI: 1.31-2.68] and 2.23 [95% CI: 1.55-3.21] times respectively (p<0.001, Table 4). Significant differences in accuracy were also found between reviewer 1 vs reviewer 2 and 3 with odds ratio of 0.39 [95% CI 0.27-0.57] and 0.47 [95% CI: 0.33-0.68], respectively (p<0.0001, Table 4).

Additional logistic model with years of fMRI experience demonstrated odds of getting the correct choice to be 1.06 for each 1 year increase in experience (i.e., a 1 year increase in fMRI improved accuracy by 6%). However, due to the small sample size the implication of this analysis is limited.

**No significant association was seen between handedness and percent correct in any setting (Supplementary Tables 3-5).**

**Mean confidence rating ranged from 1.6 (Reviewer 3, ICA50) to 2.9 (Reviewer 2, ICA 50). In general, there was no association between the confidence rating of each reviewer and whether the answer was correct, when correcting for multiple comparisons (Supplementary Tables 6-9).**

**Discussion**

While a large number of studies illustrate the potential promise of utilizing rs-fMRI as a clinical tool in various scenarios such as disease identification and stratification (e.g. neurodegenerative disorders Seeley Neurodjd), prediction of disease progression (Petrella DMN 2011), treatment response (Avissar M et al Brain Stimul 2017 Jul 13), or prognosis (Sair Radiology in press), few studies have successfully demonstrated the feasibility of using rs-fMRI at the single subject level. The primary area where this promise holds is in presurgical brain mapping, where early reports demonstrated a level of concordance between rs-fMRI and task-fMRI and DCS that was encouraging (prior references). While more recent studies question the validity of this claim when examining larger number of subjects (Sair, Duffau’s paper) due to high variability of accuracy across subjects, nevertheless in select cases, rs-fMRI may be considered a viable option for presurgical brain mapping; indeed, at least one institution has included rs-fMRI in their presurgical brain mapping paradigm without obtaining task-fMRI (Lee MH et al Topics in Magnetic Resonance Imaging 2016 Feb 25(1) 11-8). For this purpose, obtaining highly accurate intrinsic brain network data is paramount to avoid adverse outcomes following surgery.

Two widely used methods of rs-fMRI analysis are associated with unique limitations. Seed based analysis (SBA) necessitates placing regions of interest (ROIs) in selected areas of the brain depending on the network to be defined. The advantage of this method is that anatomy can guide placement of ROIs to target specific networks if clear landmarks for ROI placement are available, such as for the motor network. However, in the setting of presurgical language mapping, several confounds appear. First, primary language areas are more widely distributed anatomically across subjects (Sanai et al), thus placement of an ROI in the inferior portion of the pars opercularis, for example, may work for one patient but not another. Second, often there are anatomical distortions in expected language regions in patients with large brain lesions, making it difficult to determine correct anatomic landmarks. Finally, damage to primary language regions due to tumor infiltration or prior surgery may also cause reorganization of language networks (Southwell DG et al J Neurosurg 2016 May 124(5):1460-9, Wang et al 2013), further limiting accurate ROI placement.

The alternative is to use a data-driven approach such as ICA. Here, rs-fMRI time series typically at the voxel level is separated in to maximally independent components. The resultant component maps each may represent one of the various intrinsic brain networks, or nuisance. To determine relevant maps, automated methods such as template-matching may be used, however as is the case with SBA in patients with large lesions, anatomical distortion or network reorganization may impede its accuracy. More commonly, relevant network maps are selected by visual inspection, taking into consideration potential changes in network topology.

Aside from (however related to) the problem of choosing the ideal number of target components in ICA (i.e. ICA order), for which there is no clear paradigm (Hui et al 2011), the issue of categorizing networks derived from ICA can be especially challenging, due to several factors. First, network maps do not necessarily break evenly across components, and using low ICA orders may cause merging of one or more networks (or even networks and noise), and using high ICA orders may result in fragmentation of networks into sub-networks. For our study, using task-fMRI maps as the target, we found that a single component for ICA20 and ICA50 best matched the target, with no fragmentation at these levels of ICA orders. More difficult to estimate is the possibility that at the lower ICA order of 20, non-language related brain regions may have merged into the language network component. While no difference was found overall between ICA20 and ICA50 accuracy across raters, in more detailed analysis, a significant difference was found between the two ICA orders when comparing accuracy of selecting one top choice vs selecting the three top choices, indicating that at higher ICA orders, increased specificity of the maps may result in a higher probability of incorrect selection. In practical terms, using ICA20 may result in correct identification of the language network more frequently when only selecting one component, however the specificity of the map itself may be lower than the correct ICA50 map, thus potentially introducing type I errors. Indeed, higher orders have shown to increase the correlation of language maps when directly comparing rs-fMRI maps to task-fMRI activation maps (Sair HBM).

The fact that despite a relatively long clinical and research fMRI career of 17 years, accurate identification of rs-fMRI language map peaked at only 81% across any condition for that reviewer, is worrisome for those that are employing this method for research or clinical use. Equally problematic is the fact that this accuracy was achieved when having three choices, and limiting the choice to the best guess decreased accuracy to less than 65%. The situation becomes more dire for the additional reviewers with fewer years of experience, overall not better than chance (50%). This significantly hampers the utility of rs-fMRI as an independent tool for presurgical brain mapping when a gold standard is not available for confirmation. An alternate use of rs-fMRI would be to supplement task-fMRI rather than replace it.

Automated methods of network identification may therefore be necessary when no gold standard is available. Using a multilayer perceptron (a neural network), Mitchell et al demonstrated that reliable intrinsic brain networks were able to be characterized in 13 patients with distorted brain anatomy with electrocortical stimulation as the gold standard (Mitchell TJ et al 2013). Further validation in larger sample sizes would be necessary to ensure that methods such as this have high reliability and reproducibility.

**Conclusion**

We demonstrate low accuracy of blinded identification of rs-fMRI language networks across reviewers with varied years of fMRI experience. This limits the utility of using rs-fMRI as an alternative to task-fMRI for language mapping, although complementary usage may be beneficial together with task-fMRI. For independent use of rs-fMRI language network identification in both research and clinical domains, alternate methods of network identification must be utilized.