Predictive Modeling Achieves High Test-Retest Reliability with Resting State Functional Connectivity

Chandra Sripada[1], Mike Angstadt[1], Saige Rutherford[1], Aman Taxali[1]
[1]*Department of Psychiatry, University of Michigan, Ann Arbor, MI*

**Correspondence: sripada@umich.edu*

Abstract

Test-retest reliability is critical for individual differences research. Thus, recent reports that found low test-retest reliability in fMRI have raised concern among researchers who aim to use brain imaging to predict psychologically- and clinically-important differences across people. These previous studies, however, have mostly focused on reliability of individual fMRI features (e.g., individual connections in resting state connectivity maps). Meanwhile researchers are increasingly employing multivariate predictive models that aggregate information across a large number of features to predict outcomes of interest, but the test-retest reliability of predicted outcomes of these models has not previously been systematically studied. Here we apply four kinds of predictive modeling methods to resting state connectivity maps from the Human Connectome Project dataset to predict 62 outcome variables. In contrast to reliability of individual resting state connections, we find reliability of the predicted outcomes of predictive models is much higher for two methods: Brain Basis Set (BBS) and Connectome Predictive Modeling (CPM). BBS had the overall highest reliability, with a mean reliability across predicted outcomes of 0.79 and a reliability of 0.75 or better (conventionally considered excellent) for 56 out of 62 outcome variables. We additionally identified three mechanisms that help to explain why predictive models have higher reliability than individual features. These results suggest at least one path forward for researchers aiming to utilize resting state connectivity for individual differences research: These researchers can potentially achieve higher test-retest reliability by making greater use of predictive models.

Introduction

Recent studies report troublingly low test-retest reliability of functional magnetic resonance imaging (fMRI) metrics including resting state functional connectivity[1] and task activation[2]. These studies have attracted substantial attention from clinical and translational neuroscientists because adequate test-retest reliability of fMRI is critical for its use in individual-differences research. If differences in functional imaging features (e.g., connectivity, activation, or other features) across individuals are not stable across scanning sessions, that is, if values of these imaging features lack consistency and/or agreement across sessions, then these features cannot serve as a basis for constructing predictively useful objective markers (i.e., “biomarkers”) of traits of interest.[3]

The current literature examining reliability in fMRI has been mostly focused on the reliability of *single* imaging features: individual connections in resting state connectivity maps and individual voxels or regions of interest in task activation maps. Neuroimaging researchers studying individual-differences are, however, increasingly moving away from univariate tests performed separately on each imaging feature and are instead utilizing multivariate predictive models[4–7] (hereafter "predictive models"). These methods aggregate information across thousands of distributed brain features, yielding a single overall "best guess" about the outcome of interest. Predictive models are now widespread in the field, and have been used to predict a range of psychologically- and clinically-relevant outcomes including cognitive skills[8–10], pain ratings[11–13], sustained attention[14,15], schizophrenia status[16], and depression subtype/treatment response[17], among many others. To date, however, the test-retest reliability of predicted outcomes derived from predictive models has not been evaluated.

This question is particularly interesting in light of well-known results from psychometrics that establish that aggregation of features, for example by taking sum scores or applying a weighting function, can yield a composite variable that is much more reliable than the individual items that make up the composite.[3,18] Standard predictive models widely used in the imaging field work in just this way: They aggregate features through application of a weighting function.[4] Thus, it is possible that like composite scores in psychology, predicted outcomes from such models will exhibit meaningfully higher reliability than individual features.

To test this hypothesis, we turned to the Human Connectome Project (HCP) dataset[19], which has two sessions of resting state fMRI data for a large sample of subjects. This dataset also has a large number of phenotypic outcome variables, allowing us to train predictive models across a number of psychological domains, including cognition, emotion, personality, and psychopathology. We examined four predictive modeling methods widely used in the neuroimaging field: lasso, elastic net, connectome predictive modeling (CPM)[20], and brain basis set (BBS)[21]. Results from our systematic comparison showed that two predictive modeling methods, BBS and CPM, yielded substantially higher test-retest reliability of predicted outcomes compared to individual connectivity features, with BBS showing the overall highest reliability.

2. Methods

2.1 Subjects and Data Acquisition

All subjects and data were from the HCP-1200 release[19,22]. Study procedures were approved by the Washington University institutional review board, and all subjects provided informed consent. Four resting state runs were performed (14.5 minutes each run) across two days, with two runs the first day and two runs the second day. Data was acquired on a modified Siemens Skyra 3T scanner using multiband gradient-echo EPI (TR=720ms, TE=33ms, flip angle = 52°, multiband acceleration factor = 8, 2mm isotropic voxels, FOV = 208x180mm, 72 slices, alternating RL/LR phase encode direction). T1 weighted scans were acquired with 3D MPRAGE sequence (TR=2400ms, TE=2.14ms, TI=1000ms, flip angle = 8, 0.7mm isotropic voxels, FOV=224mm, 256 sagittal slices). T2 weighted scans were acquired with a Siemens SPACE sequence (TR=3200ms, TE=565ms, 0.7mm isotropic voxels, FOV=224mm, 256 sagittal slices).

2.2 Data Preprocessing

Processed volumetric data from the HCP minimal preprocessing pipeline that included ICA-FIX denoising were used. Full details of these steps can be found in Glasser[23] and Salimi-Korshidi[24]. Briefly, T1w and T2w data were corrected for gradient-nonlinearity and readout distortions, inhomogeneity corrected, and registered linearly and non-linearly to MNI space using FSL's FLIRT and FNIRT. BOLD fMRI data were also gradient-nonlinearity distortion corrected, rigidly realigned to adjust for motion, fieldmap corrected, aligned to the structural images, and then registered to MNI space with the nonlinear warping calculated from the structural images. Then FIX was applied on the data to identify and remove motion and other artifacts in the timeseries. These files were used as a baseline for further processing and analysis (e.g. MNINonLinear/Results/rfMRI_REST1_RL/rfMRI_REST1_RL_hp2000_clean.nii.gz from released HCP data). Images were smoothed with a 6mm FWHM Gaussian kernel, and then resampled to 3mm isotropic resolution.

The smoothed images then went through a number of resting state processing steps, including a motion artifact removal steps comparable to the type B (i.e., recommended) stream of Siegel et al.[25]. These steps include linear detrending, CompCor[26] to extract and regress out the top 5 principal components of white matter and CSF, bandpass filtering from 0.1-0.01Hz, and motion scrubbing of frames that exceed a framewise displacement of 0.5mm. Subjects with more than 10% of frames censored were excluded from further analysis, leaving 966 subjects.

2.3 Connectome Generation

We calculated spatially-averaged time series for each of 264 4.24mm radius ROIs from the parcellation of Power et al.[27]. We then calculated Pearson's correlation coefficients between each ROI. These were then transformed using Fisher's r to z-transformation.

2.4 Inclusion/Exclusion Criteria

Subjects were eligible to be included if they had: 1) structural T1 data and had 4 complete resting state fMRI runs (14m 24s each); 2) full necessary behavioral data; 3) no more than 10% of frames censored. To avoid confounding due to intra-familial similarity, we additionally randomly selected one individual from each sibship. This left 389 unrelated individuals to enter our main analysis.

2.5. ICC for Individuals Connections

Test-retest reliability for each connection of the connectome was assessed with intra-class correlation (ICC) statistic, specifically type (2,1) according to the scheme of Shrout and Fleiss[28].

2.6. Training Predictive Models

Lasso

Lasso is a form of regularized linear regression using an L^1-norm penalty that tends to shrink coefficients to 0. In this way it can act as both a regularization and feature selection method. The amount of regularization is controlled via the lambda term in the objective function:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Lambda is typically chosen via cross validation to select a value that minimizes the cross validated error. On each fold of our 10-fold cross validation (see below), we used 5-fold cross validation within the training data to select the best lambda value for lasso regression.

Elastic Net

Elastic net is a mix between an L^1 penalized regression (i.e. lasso above) and an L^2 penalized regression (i.e. ridge regression). It attempts to balance the sometimes overly aggressive feature selection of lasso by mixing it with ridge regression. There is an additional hyperparameter, alpha, that controls the balance of the lasso and ridge penalties. The objective function for Elastic Net is:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \right\}$$

As before, within our overall 10-fold cross-validation (see below), we performed 5-fold cross-validation in the training data to select the best values for alpha and lambda for elastic net.

Both elastic net and lasso were performed using the MATLAB lasso function with 40 values of lambda automatically generated from a geometric sequence by MATLAB. The values of alpha tested were [0.01, 0.1, 0.325, 0.55, 0.775, 1].

Connectome Predictive Modeling (CPM)

Connectome predictive modeling (CPM)[20] is a predictive modeling method that has been used widely in fMRI with a variety of outcome variables[29,14,30–32]. In brief, CPM is first trained with every edge of the connectome to identify edges that are predictive of the phenotype of interest above some prespecified level (e.g., Pearson's correlation with significance of $p < 0.01$). The sum of connectivity values for these specified edges is then calculated for each test subject, and these sums serve as predicted outcome scores that are correlated with the actual outcome scores. CPM typically treats positively and negatively predictive edges differently, fitting separate models for each, and so we follow this practice as well.

Brain Basis Set (BBS)

Brain Basis Set (BBS) is a predictive modeling approach developed and validated in our previous studies[21,15,8,9] (see also studies[11–13] by Wager and colleagues for a broadly similar approach). BBS is similar to principal component regression[33,34], with an added predictive element. In a training partition, PCA is performed on a subjects *x* connections matrix using the pca function in MATLAB, yielding components ordered by descending eigenvalues. Expression scores are then calculated for each of *k* components for each subject by projecting each subject's connectivity matrix onto each component. A linear regression model is then fit with these expression scores as predictors and the phenotype of interest as the outcome, saving **B**, the *k x 1* vector of fitted coefficients, for later use. In a test partition, the expression scores for each of the *k* components for each subject are again calculated. The predicted phenotype for each test subject is the dot product of **B** learned from the training partition with the vector of component expression scores for that subject. The parameter k was set at 75 because prior studies[21] showed that larger values tend to result in overfitting and worse performance.

2.7. HCP Outcome Variables

We used a total of 62 outcome variables from the HCP dataset (choice of these variables was guided by [35], and a list of these variables is available in the Supplement). Two outcome variables were derived from factor analysis of HCP variables, and they are discussed in detail in our previous report[21]. In brief, a general executive factor was created based on overall accuracy for three tasks: *n*-back working memory fMRI task, relational processing fMRI task, and Penn Progressive Matrices task. A speed of processing factor was created based on three NIH toolbox tasks: processing speed, flanker task, and card sort task (all age-adjusted performance), similar to a previous report[36].

2.8 Train/Test Split and Calculation of ICC for Predictive Models

All predictive models were trained and tested in a 10-fold cross validation scheme. We calculated ICCs for predicted outcomes for these predictive models as follows: On each fold, we trained a predictive model on the train partition for session 1 data. We then used this trained model to generate predicted outcomes for the test partition in both session 1 and session 2, and we calculated the ICC of this pair of predicted outcomes. We next did this same procedure in the other direction: We trained the predictive model on the train partition for session 2 data, generated predicted outcomes for the test partition in both session 2 and session 1, and calculated their ICC. We repeated this process on all 10 folds and averaged over all 20 ICCs (2 for each fold). We used ICC type (2,1) according to the scheme of Shrout and Fleiss[28].

For elastic net and lasso, which have an additional tunable parameter, we tuned this parameter in an embedded cross-validation procedure within the train partition. But otherwise, we followed the procedure described above.

3. Results

3.1 For Two Predictive Modeling Methods, BBS and CPM, Mean Reliability of Predicted Outcomes Across 62 Phenotypes Was Substantially Higher Than Mean Reliability for Individual Connections

The left side of Figure 1 shows test-retest reliabilities for individual connections of the connectome (in red). Mean reliability was 0.44, but the spread was remarkably wide (standard deviation 0.19). The next five plots (in blue) show test-retest reliabilities for the predicted outcomes of predictive models trained on 62 HCP outcome variables. BBS and CPM (both positive and negative) have notably higher mean reliabilities than elastic net and lasso (see also Table 1).

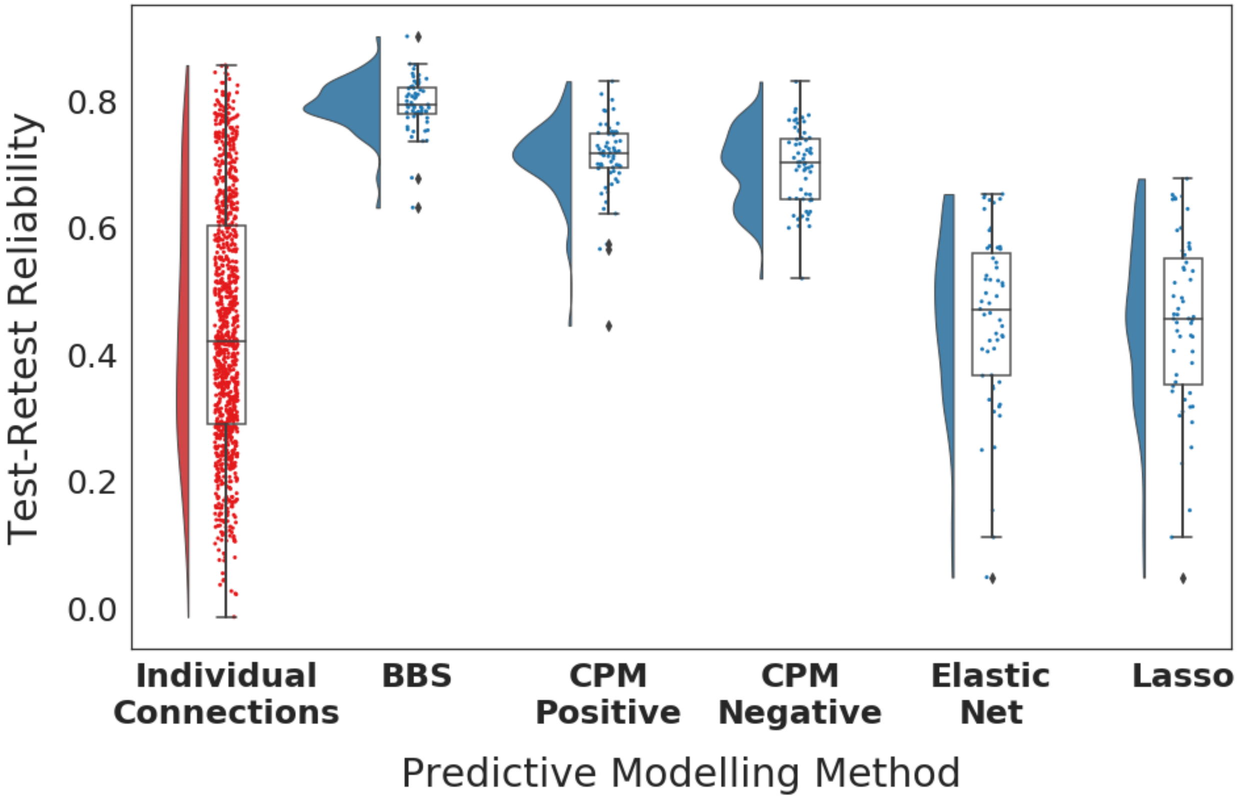

Figure 1: Distribution of Test-Retest Reliabilities for Individual Connections and for Predicted Outcomes of Predictive Models*. Reliabilities for individual connections were calculated over all connections in the resting state connectome. Mean reliability for individual connections was relatively low, but the range was wide. Reliabilities for predictive models were calculated over 62 different outcome variables available in the HCP dataset. Mean reliabilities for BBS and CPM (both positive and negative) were notably higher than for elastic net and lasso.*

Table 1 shows summary statistics for test-retest reliability as well as predictive accuracy for the predictive models. Two points are noteworthy. First, BBS and CPM had higher predictive accuracy than Elastic Net and Lasso. Second, we divided the 62 cognitive tasks into two categories: 1) Cognitive tasks, which were tasks from NIH Toolbox and Penn Neurocognitive Battery, as well as fMRI tasks of working memory (N-back), abstract reasoning (Relational Task), and math calculation (math condition of the Language Task); and 2) All the other tasks. For all four predictive modeling methods, predictive accuracy was notably higher for cognitive tasks.

	BBS	CPM Positive	CPM Negative	Elastic Net	Lasso
Mean Reliability	0.79	0.71	0.69	0.45	0.45
SD	0.04	0.06	0.06	0.14	0.14
Min	0.63	0.44	0.52	0.05	0.05
Max	0.90	0.83	0.83	0.65	0.68
Mean Accuracy (*r*)	0.10	0.08	0.07	0.03	0.02
cognitive tasks	0.17	0.14	0.10	0.12	0.10
all other tasks	0.08	0.06	0.06	0.00	-0.02

Table 1: Summary Statistics for Test-Retest Reliability and Accuracy of Predicted Outcomes of Predictive Models. Reliability is measured with the intraclass correlation (ICC) statistic. Reliability was higher for BBS and CPM compared to elastic net and lasso. Accuracy is measured with Pearson's correlation between predicted and actual outcome variable. Interestingly cognitive tasks, mostly from the NIH Toolbox and Penn Neurocognitive Battery, yielded better accuracy than all other (non-cognitive) tasks.

3.2 BBS Had the Highest and Most Consistent Test-Retest Reliabilities Across Outcome Variables

Figure 2 shows test-retest reliabilities across the 62 outcome variables for BBS, CPM-positive, and Elastic Net. For BBS, performance was consistently high, and the method had higher ICCs than CPM-positive, the next best performer, for 61 out of 62 phenotypes. Reliability is conventionally classified as excellent when it exceeds 0.75[37], and BBS exceeded this value for 56 out of 62 outcome variables.

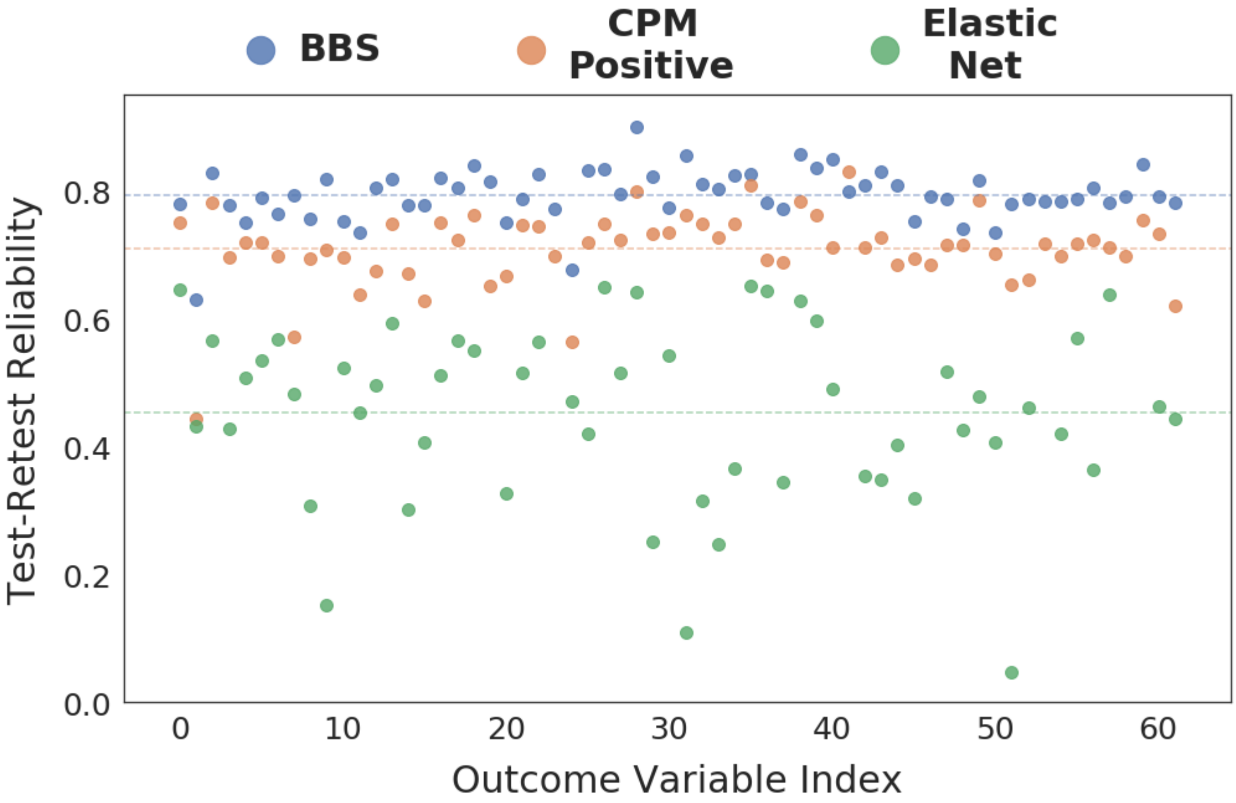

Figure 2: Performance of BBS, CPM, and Elastic Net Across the 62 Outcome Variables. *Reliability is conventionally classified as excellent when it exceeds 0.75, and BBS exceeded this value for 56 out of 62 outcome variables. Dashed lines show the mean reliability for the associated predictive modeling method.*

3.3. Multiple Mechanisms Likely Explain Why BBS and CPM Have a Large Boost in Reliability

We examined several mechanisms that could help to explain why predicted outcomes of predictive models are more reliable than individual features, focusing specifically on BBS and CPM.

Selection of High Variance Features

One mechanism that could improve reliability specifically in BBS is that it leverages features with high inter-subject variance, and test-retest reliability is positively related to inter-subject variance[3,38]. BBS uses dimensionality reduction with PCA, and the PCA algorithm finds components in descending order of variance explained, with the first component explaining the most variance in the data.[39] Consistent with this idea, we found the mean variance of the individual connections in the connectome was 0.04. But mean variance of expression scores of the first 75 PCA components is 4.4, more than 100 times larger.

Selection of Correlated Features

Another mechanism that could improve reliability in both BBS and CPM is selecting correlated features and aggregating over them. It is well known from classical test theory that the sum of a set of positively correlated features will have higher reliability than the features themselves[3] (see the Supplement for a general equation linking reliability of a weighted sum to the statistical properties of the individual features). BBS selects correlated features through the use of PCA: Each component consists of a weighted set of features that are jointly co-expressed across subjects (in proportion to the loadings), and thus these features are correlated across subjects.[39] Supporting this idea, we found that the mean test-retest reliability of the first 75 PCA components is 0.66, much higher than the mean reliability of individual features (which is 0.44). Of note, this boost in reliability for components likely reflects both the operation of this second mechanism (selecting correlated features) as well as the first (selecting high variance features).

CPM also selects correlated features, but in a different way. CPM performs a search for features that are correlated with the outcome variable up to a desired level of statistical significance (e.g., $p< 0.01$). Since these features are all correlated with the behavioral variable of interest, they will also tend to be correlated with each other as well. Consistent with this idea, we found that mean pairwise intercorrelation of all features across the connectome is 0.005, while the mean pairwise intercorrelation of CPM-selected feature set is 0.05 for CPM positive and 0.05 for CPM negative.

Selection of Valid Features

Assume that there is true variance in the connectome that relates to an outcome variable of interest (that is, there are stable, non-noise connectomic differences that correlate with an outcome variable). Both BBS and CPM select features that are correlated with the specified outcome variable. Given our assumption, then, features selected by BBS and CPM will correspondingly be enriched with respect to these valid, stable connectomic differences. This enrichment with respect to true variance will boost test-retest reliability.

To demonstrate the role of this third mechanism for boosting reliability, we permuted subject labels for the 62 outcome variables 100 times, in effect creating random outcome variables for each subject. We then computed test-retest reliability for BBS and CPM trained on these randomized outcome variables. We found mean reliability for BBS and CPM in predicting the randomized outcome variables was 0.65 and 0.54, respectively. This is notably lower than their respective mean reliabilities in predicting real outcome variables (0.79 for BBS and 0.72 for CPM). This result suggests selection of valid features does in fact play a role in boosting reliabilities for both BBS and CPM.

4. Discussion

This is the first study to systematically investigate test-retest reliability of multivariate predictive models applied to resting state connectivity maps. We found that in contrast to reliability of individual resting state connections, reliability of the predicted outcomes of predictive models is much higher for two modeling methods, BBS and CPM. We also found that BBS was the overall best performer: For 56 out of 62 outcome variables, reliability of BBS predicted outcomes was better than 0.75, conventionally considered excellent. Test-retest reliability is critical for the use of fMRI in individual differences research. Our results suggest more widespread use of predictive models can help address questions about reliability that have been raised in recent reports and that remain a serious concern for the neuroimaging field.

Test-Retest Reliability and Units of Analysis

Previous studies of reliability in resting state fMRI have mostly examined individual connections.[40–42] While results have varied, a recent meta-analysis[1] found reliability was typically relatively low at 0.29. Broadly consistent with this result, we found mean reliability of individual connections in the HCP dataset was 0.44. Several studies examined larger, more complex units of analysis and found higher levels of reliability. For example, Noble and colleagues[40] examined mean connectivity within intrinsic connectivity networks such as default mode network and fronto-parietal network. They found reliabilities were modestly higher for networks than for individual connections (range 0.35 to 0.60 for networks). Similarly, a modest boost in reliability appears to be observed with higher-order metrics such as graph theory metrics[43,44]. Predictive models, which aggregate across a still wider range of features using trained feature weights arguably represent a still higher, more complex unit of analysis. In the present study, we found clear evidence of substantially higher test-retest reliability for predicted outcomes of predictive models (for BBS and CPM in particular). Overall, these results suggest that test-retest reliability differs substantially across units of analysis, as well as the types of aggregation methods that were utilized to generate the higher-level units.

Should We Be Pessimistic or Optimistic About Using Resting State Connectivity for Individual Differences Research?

Given high mean reliability for predicted outcomes of predictive models and much lower mean reliability for individual connections, should we be an optimistic or pessimistic about reliability of resting state connectivity? While we acknowledge both perspectives capture part of the overall picture, we briefly suggest there is more reason for optimism.

Test-retest reliability is most critical for research that seeks to use imaging features to predict individual differences, for example, translational neuroimaging research that aims to construct brain-based biomarkers. This is because reliability is mathematically related to predictive validity, the ability of measures to predict an outcome. According to an important formula in classical test theory, reliability sets a ceiling on predictive validity, and as reliability falls, the maximum achievable predictive validity drops with it.[3]

But, critically, if one's goal is in fact prediction of individual differences of some outcome variable of interest (e.g., behaviors or symptoms), focusing on individual connections of the connectome is unlikely to be a fruitful approach. This is because for most constructs of interest—general cognitive ability, neuroticism, pain, autism, it is unlikely that any single connection contains much discriminative information about the construct. Rather, it is likely that this discriminative information resides in distributed changes across widespread network connections. To capture this diffuse, distributed information, univariate tests are less effective than multivariate methods such as predictive models[4,45]. Consistent with this idea, Wager, Woo, Chang and their colleagues have shown in a series of studies with task-based fMRI that effect sizes are generally substantially larger with predictive models than with univariate statistical tests applied to individual imaging features[12,13,38,45].

In short, then, predictive models are arguably a more important tool for individual differences research in fMRI than univariate tests applied to individual imaging features. If this is correct, then poor reliability of individual imaging features may not be a major concern. Rather, a more optimistic interpretation is available: Predictive models, which are the critically important tools we need for individual differences research in neuroimaging, *do* appear to have adequate levels of test-retest reliability (at least with certain methods such as BBS and CPM).

Why Predictive Models Have Better Test-Retest Reliability

We examined several factors that can explain why predictive models such as BBS and CPM have better reliability compared to individual connections. One mechanism is that these predictive models select valid features, i.e., features that vary due to stable underlying differences across individuals. Moreover, we showed reliability of predictive models suffers when outcome variables are randomized (and thus this mechanism is blocked). A second mechanism is that BBS and CPM both select and then aggregate correlated features, but in different ways: BBS selects correlated features directly as part of its PCA procedure, while CPM selects them indirectly through the fact that the features it selects are all correlated with the behavioral outcome of interest. A third mechanism that boosts reliability that is more unique to BBS is selection of high variance features. These three mechanisms interact in complex ways. Moreover, the relative roles of these three mechanisms in boosting reliability for any particular dataset and outcome variable appear to be hard to specify. The size of the boost depends in complex ways on the variance/covariance structure of the imaging features and the precise patterns with which these features correlate with the outcome variable.

Implications for Test-Retest Reliability of Task fMRI

Recent reports also find poor reliability in task fMRI[2]. This result may be seen as particularly discouraging because many researchers have thought that tasks, because they involve carefully controlled manipulations of psychological constructs, might be an especially effective way of detecting differences in these constructs across individuals.[46] Could predictive models play a similar role in boosting reliability with task-based fMRI? We believe the answer is likely to be yes. The three mechanisms we identified for why predictive models boost reliability are quite general and reflect basic statistical properties of these models. There is no obvious reason to expect that they will operate only with resting state connectivity maps and not with task activation maps. Moreover, one study found initial evidence that predictive models do in fact produce a boost in reliability in the task setting: Woo and Wager[38] report that model-based predictions of pain ratings during a nociceptive stimulation task had higher reliability (ICC=0.72) than three regions of interest known to be associated with pain (ICCs: 0.54 to 0.59). To further investigate this issue, in a companion report, we perform a systematic comparison with task activation maps of voxel- and region of interest-level reliability versus reliability of predicted outcomes of predictive models.

Limitations

This study has several limitations. First, we assessed four popular predictive modeling methods, and we found sizable differences in their test-retest reliabilities for predicted outcomes (with the starkest differences between BBS and CPM on the one hand and elastic net and lasso on the other). There are a large number of other predictive modeling methods that we did not study and future work can systematically compare them. Second, while we examined a large number of outcome variables (62 in total), there are of course a vast number of outcome variables that we could not test with the HCP dataset (e.g., pain ratings, schizophrenia status, depression-treatment response, etc.). As more comprehensive data sets become available, it would be useful to extend these results to a still broader range of outcome variables. Third, it bears emphasis that test-retest reliability is a statistic that is specific to a given population. Most relevant for the present purposes, it is highly sensitive to the inter-individual variance in imaging features.[3] The HCP dataset consists of a fairly homogenous sample of psychologically healthy young adults. It is possible that reliability will be higher in fMRI, at both the individual feature-level as well as the predictive model-level, if more heterogenous samples are considered, as this could potentially boost inter-individual variance in imaging features.[38]

Conclusion

In sum, this study is the first to systematically assess the test-retest reliability of predicted outcomes of predictive models applied to resting state connectivity maps. In contrast to the somewhat bleak conclusions of recent studies about reliability of individual imaging features, we found that least some predictive modeling methods, specifically BBS and CPM, demonstrate consistently high test-retest reliability.

Competing Interests

The authors declare no conflicts of interest.

Funding

CS was supported by R01MH107741, U01DA041106, and a grant from the Dana Foundation David Mahoney Neuroimaging Program.

References

1. Noble S, Scheinost D, Constable RT. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *NeuroImage*. September 2019:116157. doi:10.1016/j.neuroimage.2019.116157

2. Elliott ML, Knodt AR, Ireland D, et al. Poor test-retest reliability of task-fMRI: New empirical evidence and a meta-analysis. *bioRxiv*. 2019:681700.

3. Nunnally Jr JC. Introduction to psychological measurement. 1970.

4. Woo C-W, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci*. 2017;20(3):365-377. doi:10.1038/nn.4478

5. Scheinost D, Noble S, Horien C, et al. Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage*. 2019.

6. Orru G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci Biobehav Rev*. 2012;36(4):1140-1152.

7. Klöppel S, Abdulkadir A, Jack Jr CR, Koutsouleris N, Mourão-Miranda J, Vemuri P. Diagnostic neuroimaging across diseases. *Neuroimage*. 2012;61(2):457-463.

8. Sripada C, Rutherford S, Angstadt M, et al. Prediction of Neurocognition in Youth From Resting State fMRI. *Mol Psychiatry*. accepted:495267. doi:10.1101/495267

9. Sripada C, Angstadt M, Rutherford S. Towards a “Treadmill Test” for Cognition: Reliable Prediction of Intelligence From Whole-Brain Task Activation Patterns. *bioRxiv*. January 2018. doi:10.1101/412056

10. Dubois J, Galdi P, Paul LK, Adolphs R. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *Philos Trans R Soc B Biol Sci*. 2018;373(1756). doi:10.1098/rstb.2017.0284

11. Wager TD, Atlas LY, Lindquist MA, Roy M, Woo C-W, Kross E. An fMRI-Based Neurologic Signature of Physical Pain. *N Engl J Med*. 2013;368(15):1388-1397. doi:10.1056/NEJMoa1204471

12. Chang LJ, Gianaros PJ, Manuck SB, Krishnan A, Wager TD. A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biol*. 2015;13(6):e1002180.

13. Woo C-W, Schmidt L, Krishnan A, et al. Quantifying cerebral contributions to pain beyond nociception. *Nat Commun*. 2017;8:14211.

14. Rosenberg MD, Finn ES, Scheinost D, et al. A neuromarker of sustained attention from whole-brain functional connectivity. *Nat Neurosci*. 2016;19(1):165-171. doi:10.1038/nn.4179

15. Kessler D, Angstadt M, Sripada C. Brain Network Growth Charting and the Identification of Attention Impairment in Youth. *JAMA Psychiatry*. 2016;73(5):481-489.

16. Watanabe T, Kessler D, Scott C, Angstadt M, Sripada C. Disease Prediction based on Functional Connectomes using a Scalable and Spatially--Informed Support Vector Machine. *NeuroImage*. 2014;96:183-202.

17. Drysdale AT, Grosenick L, Downar J, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med*. 2017;23(1):28.

18. He Q. Estimating the reliability of composite scores. In: Coventry: Ofqual; 2009.

19. Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. The WU-Minn Human Connectome Project: An overview. *NeuroImage*. 2013;80:62-79. doi:10.1016/j.neuroimage.2013.05.041

20. Shen X, Finn ES, Scheinost D, et al. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat Protoc*. 2017;12(3):506-518. doi:10.1038/nprot.2016.178

21. Sripada C, Angstadt M, Rutherford S, et al. Basic Units of Inter-Individual Variation in Resting State Connectomes. *Sci Rep*. 2019;9(1):1900. doi:10.1038/s41598-018-38406-5

22. WU-Minn HCP. 1200 Subjects Data Release Reference Manual. 2017.

23. Glasser MF, Sotiropoulos SN, Wilson JA, et al. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*. 2013;80:105-124. doi:10.1016/j.neuroimage.2013.04.127

24. Salimi-Khorshidi G, Douaud G, Beckmann CF, Glasser MF, Griffanti L, Smith SM. Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage*. 2014;90:449-468. doi:10.1016/j.neuroimage.2013.11.046

25. Siegel JS, Mitra A, Laumann TO, et al. Data Quality Influences Observed Links Between Functional Connectivity and Behavior. *Cereb Cortex N Y N 1991*. 2017;27(9):4492-4502. doi:10.1093/cercor/bhw253

26. Behzadi Y, Restom K, Liau J, Liu TT. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*. 2007;37(1):90-101. doi:10.1016/j.neuroimage.2007.04.042

27. Power JD, Cohen AL, Nelson SM, et al. Functional Network Organization of the Human Brain. *Neuron*. 2011;72(4):665-678. doi:10.1016/j.neuron.2011.09.006

28. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428.

29. Finn ES, Shen X, Scheinost D, et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat Neurosci*. 2015;18(11):1664-1671. doi:10.1038/nn.4135

30. Yoo K, Rosenberg MD, Hsu W-T, et al. Connectome-based predictive modeling of attention: Comparing different functional connectivity features and prediction methods across datasets. *NeuroImage*. 2018;167:11-22. doi:10.1016/j.neuroimage.2017.11.010

31. Beaty RE, Kenett YN, Christensen AP, et al. Robust prediction of individual creative ability from brain functional connectivity. *Proc Natl Acad Sci*. January 2018:201713532. doi:10.1073/pnas.1713532115

32. Lake EMR, Finn ES, Noble SM, et al. The functional brain organization of an individual predicts measures of social abilities in autism spectrum disorder. *bioRxiv*. March 2018:290320. doi:10.1101/290320

33. Park SH. Collinearity and optimal restrictions on regression parameters for estimating responses. *Technometrics*. 1981;23(3):289-295.

34. Jolliffe IT. A note on the use of principal components in regression. *Appl Stat*. 1982:300-303.

35. Kong R, Li J, Orban C, et al. Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. *Cereb Cortex*. 2018;29(6):2533-2551.

36. Carlozzi NE, Beaumont JL, Tulsky DS, Gershon RC. The NIH Toolbox Pattern Comparison Processing Speed Test: Normative Data. *Arch Clin Neuropsychol*. 2015;30(5):359-368. doi:10.1093/arclin/acv031

37. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6(4):284.

38. Woo C-W, Wager TD. What reliability can and cannot tell us about pain report and pain neuroimaging. *Pain*. 2016;157(3):511-513.

39. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*. 2010;2(4):433-459.

40. Noble S, Spann MN, Tokoglu F, Shen X, Constable RT, Scheinost D. Influences on the Test–Retest Reliability of Functional Connectivity MRI and its Relationship with Behavioral Utility. *Cereb Cortex*. 2017;27(11):5415-5429. doi:10.1093/cercor/bhx230

41. Birn RM, Molloy EK, Patriat R, et al. The effect of scan length on the reliability of resting-state fMRI connectivity estimates. *NeuroImage*. 2013;83:550-558. doi:10.1016/j.neuroimage.2013.05.099

42. Shehzad Z, Kelly AM, Reiss PT, et al. The Resting Brain: Unconstrained yet Reliable. *Cereb Cortex*. 2009. doi:bhn256 [pii] 10.1093/cercor/bhn256

43. Braun U, Plichta MM, Esslinger C, et al. Test–retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. *Neuroimage*. 2012;59(2):1404-1412.

44. Termenon M, Jaillard A, Delon-Martin C, Achard S. Reliability of graph analysis of resting state fMRI using test-retest dataset from the Human Connectome Project. *Neuroimage*. 2016;142:172-187.

45. Reddan MC, Lindquist MA, Wager TD. Effect size estimation in neuroimaging. *JAMA Psychiatry*. 2017;74(3):207-208.

46. Matthews PM, Honey GD, Bullmore ET. Neuroimaging: Applications of fMRI in translational medicine and clinical practice. *Nat Rev Neurosci*. 2006;7(9):732.

Supplement

Composite Reliability Formula (relates test-retest reliability of a composite variable to statistical properties of the variables that are summed; based on He and colleagues[18]).

$$r = 1 - \frac{\sum_{i=1}^{n} w_i^2 \sigma_{e,X_i}^2}{\sum_{i=1}^{n} w_i^2 \sigma_{X_i}^2 + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} w_i w_j \sigma_{X_i,X_j}}$$

r = reliability of the composite variable
w_i = assigned weight to feature i
σ_{e,X_i}^2 = error variance of feature i
$\sigma_{X_i}^2$ = variance of feature i
σ_{X_i,X_j} = covariance between feature i and feature j

Outcome Variables from HCP Dataset

Number	Variable Name	Description
1	GenExec	General Executive Factor
2	ProcSpeed	Processing Speed Factor
3	PMAT24_A_CR	Fluid Intelligence (PMAT)
4	ASR_Extn_T	Adult Self Report - Externalizing
5	ASR_Intn_T	Adult Self Report - Internalizing
6	ASR_Attn_T	Adult Self Report - Attention
7	NEOFAC_O	Openness (NEO)
8	NEOFAC_C	Conscientiousness (NEO)
9	NEOFAC_E	Extraversion (NEO)
10	NEOFAC_A	Agreeableness (NEO)
11	NEOFAC_N	Neuroticism (NEO)
12	DDisc_AUC_40K	Delay Discounting
13	ProcSpeed_AgeAdj	Processing Speed
14	PicSeq_AgeAdj	Visual Episodic Memory
15	CardSort_AgeAdj	Cognitive flexibility (DCCS)
16	Flanker_AgeAdj	Inhibition (Flanker task)
17	ListSort_AgeAdj	Working Memory (list sorting)
18	ReadEng_AgeAdj	Reading (pronounciation)
19	PicVocab_AgeAdj	Vocabulary (picture matching)

20	SCPT_SEN	Sustained Attention - Sens.
21	SCPT_SPEC	Sustained Attention - Spec.
22	IWRD_TOT	Verbal Episodic Memory
23	VSPLOT_TC	Spatial orientation
24	MMSE_Score	Cognitive status (MMSE)
25	PSQI_Score	Sleep quality (PSQI)
26	Endurance_Unadj	Walking endurance
27	GaitSpeed_Comp	Walking Speed
28	Dexterity_Unadj	Manual dexterity
29	Strength_Unadj	Grip strength
30	Odor_Unadj	Odor identificaiton
31	PainInterf_Tscore	Pain Interference Survey
32	Taste_Unadj	Taste intensity
33	Mars_Final	Contrast Sensitivity
34	Emotion_Task_Face_Acc	Emotional Face Matching
35	Language_Task_Math_Avg_Difficulty_Level	Arithmetic
36	Language_Task_Story_Avg_Difficulty_Level	Story comprehension
37	Social_Task_Perc_Random	Social Cognition - random
38	Social_Task_Perc_TOM	Social Cognition - interaction
39	WM_Task_Acc	Working Memory (n-back)
40	ER40_CR	Emot. Recog. - Total
41	ER40ANG	Emot. Recog. - Angry
42	ER40FEAR	Emot. Recog. - Fear
43	ER40HAP	Emot. Recog. - Happy
44	ER40NOE	Emot. Recog. - Neutral
45	ER40SAD	Emot. Recog. - Sad
46	AngAffect_Unadj	Anger - Affect
47	AngHostil_Unadj	Anger - Hostility
48	AngAggr_Unadj	Anger - Aggression
49	FearAffect_Unadj	Fear - Affect
50	FearSomat_Unadj	Fear - Somatic Arousal
51	Sadness_Unadj	Sadness
52	LifeSatisf_Unadj	Life Satisfication
53	MeanPurp_Unadj	Meaning & Purpose
54	PosAffect_Unadj	Positive Affect
55	Friendship_Unadj	Friendship
56	Loneliness_Unadj	Loneliness
57	PercHostil_Unadj	Perceived Hostility
58	PercReject_Unadj	Perceived Rejection
59	EmotSupp_Unadj	Emotional Support
60	InstruSupp_Unadj	Instrument Support
61	PercStress_Unadj	Perceived Stress
62	SelfEff_Unadj	Self-Efficacy