Bagging Improves Reproducibility of Functional Parcellation

Abstract:
Developing methods that increase the reproducibility and reliability of neuroimaging measurement is an important challenge in clinical and cognitive neuroscience neuroimaging. One particular area of importance is estimation of functional areal organization, often studied through functional parcellation of the brain. Functional areal organization shows substantial variance across individuals, and creating more reproducible and reliable functional areal parcellations would allow for more generalizable estimates of brain organization. We apply bootstrap aggregation, or bagging, to the problem of improving reproducibility in functional parcellation. We use two test-retest datasets, one of 30 young adults scanned 10 times for ten minutes per scan, and another of 300 young adults scanned twice for six minutes per scan, to demonstrate that bagging provides functional parcellations with higher reproducibility and reliability compared to non-bagged functional parcellation. While increasing scan length and sample size have been regarded as the main methods of improving robustness of estimating functional organization, our results demonstrate that bagging can be used to boost the robustness of functional parcellation with as little as five minutes of scan time in as few as thirty subjects. These results imply bagging can be used to improve the robustness in acquisitions with a short scan time, which is commonplace in many established and ongoing studies and open source datasets. By testing an array of different reproducibility metrics, datasets, cluster levels, and acquisition lengths, we show where bagging can improve the reproducibility and reliability of functional parcellations. Overall, it seems that the use of this approach is beneficial in creating more reproducible clusters, and bagging should be applied when reproducibility of functional parcellations are under consideration.

Keywords: reliability, discriminability, accuracy, bagging, cluster ensembles, ensemble learning, striatum, thalamus, subcortical, basal ganglia, fMRI, bootstrap

1. Introduction

Reproducibility is an essential step towards building solid domain knowledge in the neurosciences, and recent efforts have focused on addressing problems that lead to lack of reproducibility. Given that functional parcellation remains a central goal for neuroscience, finding methods to enhance reproducibility of functional parcellation is important for improving robustness of findings. The past decade has yielded a high degree of convergence in the identification of large-scale functional networks that are highly reproducible across studies and laboratories, and can be measured at the individual level with moderate to high test-retest reliability. In contrast, efforts to delineate functionally-defined areal parcels have yielded more variable results across studies, with few focusing on the individual level (Glasser et al., 2016; Mejia et al., 2015). Among those that have attempted to quantify either the reproducibility or reliability of parcellation results, several have pointed to the need for optimization of methodologies, as the minimum data requirements needed to achieve desirable levels of reliability appear to supercede most data collections to date (Laumann et al., 2015; Xu et al., 2016). As such, there an increasing need for methods that can optimize the value of resting state acquisitions with less than 25 minutes of data that are either already established or currently underway.

Here, we draw upon advances in the machine learning literature, resampling methods and ensemble clustering in particular, to assess whether we can bring down the minimum data needed to achieve reproducible and reliable functional parcellations. The motivation behind bagging, a resampling technique that aggregates bootstrap samples (Breiman, 1996), is to reduce variability in the estimation process through averaging (Dudoit & Fridlyand, 2003). While originally designed for prediction, bagging has more recently become an important technique for ensemble clustering (Fischer & Buhmann, 2003; Hong, Kwong, Wang, & Ren, 2009; Jia, Xiao,

Liu, & Jiao, 2011; Li & Ding, 2008; Zhou, 2012). The essential notion of ensemble clustering is the added value of combining multiple cluster assignments into a single clustering (Strehl & Ghosh, 2002). In other words, the aggregation of cluster solutions themselves become the features for the consensus clustering. Numerous studies have found that the instability and sensitivity inherent in the cluster optimization procedure are significantly attenuated by cluster ensembles generated through bagging (Dudoit & Fridlyand, 2003; Hoyos-Idrobo, Schwartz, Varoquaux, & Thirion, 2015; Jia et al., 2011; Li & Ding, 2008; Strehl & Ghosh, 2002; Zhou, 2012), making this a promising area of research in the machine learning.

Here, we examine the impact of bagging on the reproducibility and reliability of functionally-defined cortical and subcortical parcellations generated using resting state fMRI data. We leverage and extend the Bootstrap Analysis of Stable Clusters (BASC) framework. BASC applies multi-level bagging to create group and individual-level functional parcellations. Though it was introduced nearly a decade ago (Bellec, Rosa-neto, Lyttelton, Benali, & Evans, 2010), the full merits of bagging are yet to be tested or appreciated in functional parcellation. Our strategy for establishing and assessing reproducibility is motivated by the COBIDAS report and broader outlines for the establishment of reproducibility in metrology and biomedical science (Jcgm & Others, 2008). These efforts have highlighted a range of challenges in reproducibility the field faces as we move towards more robust estimates of brain organization and biomarker discovery (Table 2; Zuo, Biswal, & Poldrack, 2019). In brief, we focus on creating parcels that improve 1-Between-Sample Reproducibility: similarity in group-level parcellations obtained between independent samples; 2-Between-Session Reproducibility: similarity of group level parcellations obtained in the same sample at different time points; 3-Reliability: consistency of differences among individuals on repeat assessment at distinct time points (See Methods 2.5-2.6 for more detailed definitions of these terms).

Keeping in line with extensive efforts at both cortical (Bellec et al., 2010; Craddock, James, Holtzheimer, Hu, & Mayberg, 2012; A. M. C. Kelly et al., 2009; C. Kelly et al., 2012; Margulies et al., 2007) and subcortical parcellation (Barnes et al., 2010; Choi, Yeo, & Buckner, 2012; Garcia-Garcia et al., 2017; Janssen, Jylänki, Kessels, & van Gerven, 2015), we perform functional parcellation at multiple resolutions in both cortex and subcortex to assess the impact of bagging on reproducibility and reliability of these parcellations. While several cortical parcellations are well established (Glasser et al., 2016; Yeo et al., 2011), subcortex is an area where parcellations are not only less well-studied (Choi et al., 2012; Janssen et al., 2015), but also are in need of methods to improve reproducibility and reliability given that they show notably lower reliabilities than the rest of the brain (Noble et al., 2017; O'Connor et al., 2017).

The present study leverages openly available datasets that are well suited for the establishment of minimum data requirements for reproducibility reliability (Table 1). We use the Hangzhou Normal University (HNU) test retest dataset (Zuo et al., 2014), which is comprised of 30 subjects each scanned 10 times for 10 minutes over the course of 30 days, and the Genomics Superstruct Project (GSP) dataset (Holmes et al., 2015), from which we select a subset of 300 participants with two 6 minute resting state scans, which we age and sex matched to the HNU dataset. Given the extensive computational burden of the current set of analysis, we run a full spectrum of parameters in the subcortical parcellation, and a complete, but reduced set on the cortical parcellation. **We demonstrate that compared to a standard clustering framework, bagging improves the reproducibility and reliability of both cortical and subcortical functional parcellations across a range of sites, scanners, individuals, scan lengths, samples, and clustering parameters. These results suggest that bagging may be an important component for achieving more robust functional parcellations in functional neuroimaging.**

2. Methods

2.1 Overview

We aimed to approach a full-scale assessment of the impact of bagging on between sample reproducibility, within-sample between-session reproducibility, and between-scan between-subject reliability. Towards this end, we apply PyBASC, a multi-level bagging approach for functional parcellation (See Methods Section 2.4 on PyBASC).

Data Source	Sample Size	Data Length	Age	Sex	Voxel Size	TR
HNU	30	10x 10-Minute	24.37, +/- 2.37	50/50 % F/M	3.4, 3.4, 3.4 mm	2000ms
GSP	300	2x 6-Minute	23.67, +/- 2.15	50/50 % F/M	3.0, 3.0, 3.0 mm	3000ms

Table 1: Overview of datasets used in the current study. HNU- Hanzhou Normal University. GSP- Genomics Superstruct Project.

2.2 HNU and GSP Data

We assessed the impact of bagging on test-retest reproducibility of functional parcellation using the HNU from CoRR (Zuo et al., 2014), and the Genomics Superstruct Project (GSP) (Holmes et al., 2015; Table 1). In the HNU dataset, 10 minute resting state scans were acquired in 30

young adults every three days for a month, for a total of 10 sessions per person. We also used 300 young adult participants age and gender matched to the HNU dataset from the Genomics Superstruct Project to assess reproducibility. From these 300 participants, we created 10 groups of 30 age and gender matched participants to assess within-study between-sample reproducibility. We also used all 300 participants to create a reference parcellation. We compared this reference to the parcellations of the HNU data to assess the impact of bagging on the between-sample between-site reproducibility of functional parcellation.

2.3 MRI Data

2.3.1 MRI Acquisition- *HNU. Anatomical*: 3D SPGR images were acquired on a GE 3T scanner with an 8 Channel head coil. Flip angle: 8 Degrees. TI: 450ms, TE: 60ms, TR: 8.06s, 180 slices, slice thickness 1mm. Acquisition time: 5:01. *Rest*: EPI images were acquired on a GE 3T scanner with an 8 Channel head coil. Flip angle 90 Degree. TE 30ms, TR 2000ms, 43 slices, slice thickness: 3.4 mm. Acquisition time: 10:00.

GSP Anatomical: T1 MEMPRAGE images were acquired on Siemens 3T Magnetom Tim Trio scanner with a 12 channel headcoil. Flip angle: 7 degrees. TI 1.1s, TE:1.5 / 3.4 / 5.2 / 7.0. TR 2.2s. 144 slices, slice thickness 1.2mm. Acquisition time: 2:12. *Rest*: T2*BOLD images were acquired on a Siemens 3T Magnetom Tim Trio scanner with a 12 channel headcoil. Flip angle 85 degrees. TE 30ms, TR 3.0s, 47 slices, 3mm slice thickness. Acquisition time: 6:12.

2.3.2 Preprocessing- *Anatomical:* 1) AFNI skull stripping. 2) MNI 152 2 mm template anatomical nonlinear registration with FNIRT. 3) Data transformed to 3mm. *Functional:* 1) Transformed to 3 mm. 2) Nuisance regression applied (white matter, CSF, global signal, motion, linear, and quadratic components of motion). 3) Friston 24-Parameter model used for volume realignment.

4) Nuisance band pass filtering between 0.01 and 0.1 Hz. 5) Spatial smoothing applied 6mm FWHM. Preprocessing in GSP and HNU structural and rest data were identical with the addition of de-spiking to the GSP resting state data. (HNU- CPAC 1.0; GSP- CPAC Version 1.2. See supplemental materials for the full CPAC preprocessing yaml file).

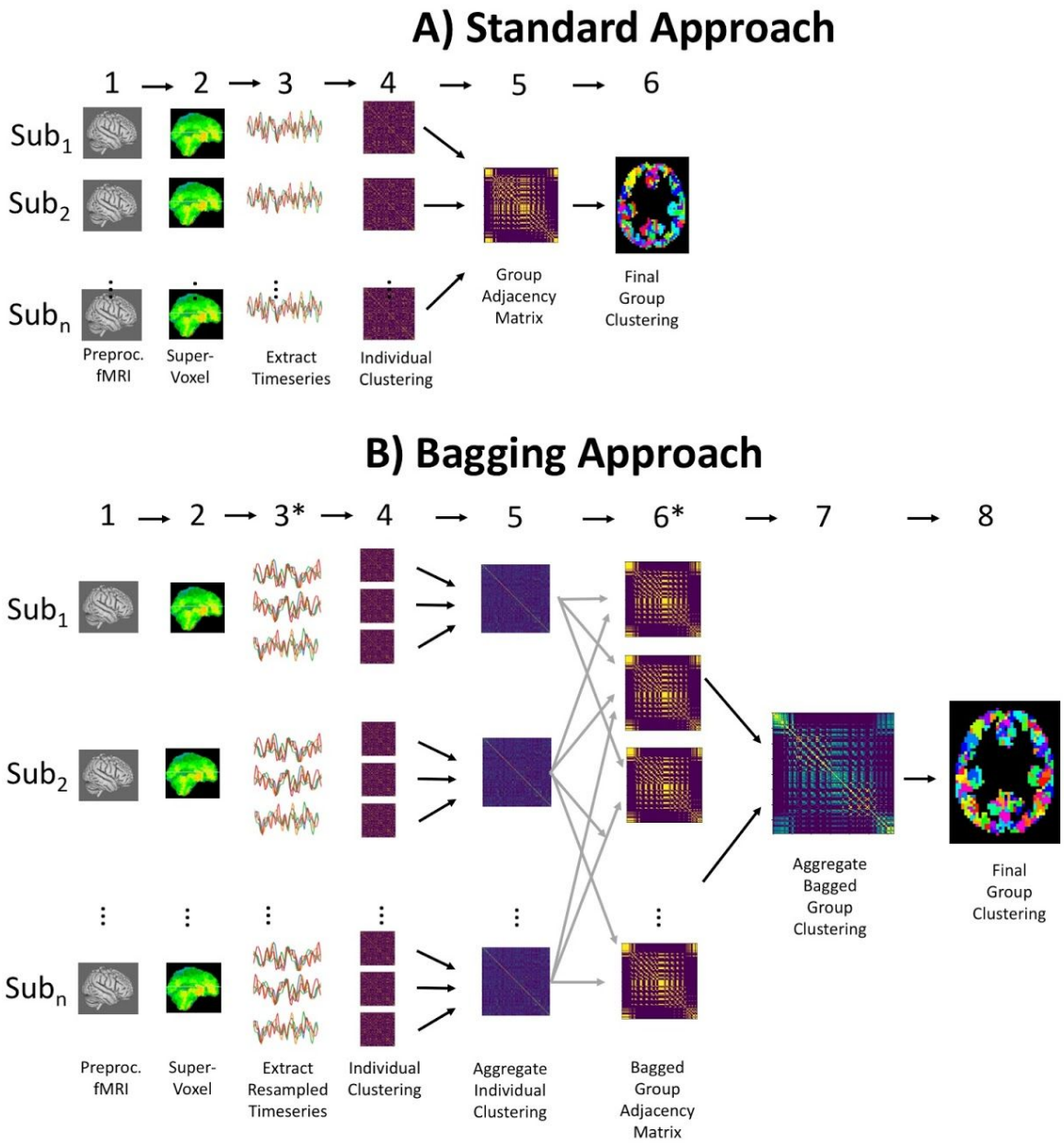

Figure 1- A) Outline of different steps in standard clustering approach. B) Creating cluster ensembles through bagging follows the same overall structure as A, with the addition of individual level time-series resampling at step 3, and bagging across individuals in step 6.

2.4 PyBASC Parcellation

When bagging is not applied, the structure of BASC follows commonly applied functional parcellation methods quite closely (Craddock et al., 2012). 1) We start with preprocessed functional MRI data; 2) this data is transformed from voxel representation to a supervoxel representation through data reduction 3) time series for each supervoxel is extracted, and a correlation matrix is calculated, 4) clustering is applied to extract a specific number of clusters (K) for each individual's correlation matrix and an adjacency matrix is created, and 5) adjacency matrices are averaged together and 6) clustered again to reveal the group level clustering.

In BASC, this approach is extended by bagging of the supervoxel time-series (Figure 1B; Step 3) to create multiple individual level adjacency matrices (4) across each bagged time-series. We then bagged across individuals to create resampled groups (6), which are then averaged to create the group mean adjacency matrix (7) that is clustered to create the final parcellation (8). This multilevel bagging approach has been previously applied in the original BASC implementation in Octave and has been explained in greater detail elsewhere (Bellec et al., 2010; Garcia-Garcia et al., 2017; See supplementary materials for more details about PyBASC). Here we have created a python based implementation of this framework, PyBASC, to conduct the following reproducibility and reliability assessments. Overall, PyBASC and BASC follow the same algorithmic structure, however there are a few differences worth noting: 1) Dimension reduction is conducted through a region growing algorithm in BASC, whereas in PyBASC Ward's criterion hierarchical clustering is applied. 2) Dimension reduction is applied at the group level in BASC and at the individual or group level in PyBASC.

In the current work, we applied parcellation to both cortical and subcortical regions across a range of number of clusters (Subcortical k = 10, 18; Cortical k = 7, 20, 197, 444). We chose k = 10, and 18 for the subcortical parcellation given our work and other literature has

repeatedly shown these parcellations fit best for the striatum and thalamus (Behrens et al., 2003; Choi et al., 2012; Garcia-Garcia et al., 2017; Johansen-Berg et al., 2005). We chose k = 7, 20, 197, and 444 following previous analysis with BASC showing these levels provided the most stable clusters (Bellec et al., 2010). Resampling involves a random element, and so to estimate the range of reproducibility produced by bagging we repeated the subcortical and cortical parcellations (20 and 10 times respectively), and compared them both to their respective reference parcellations.

Level of Generalization	**Participants**		**MRI**		**Data Used**	**Figures**
	Pop.	Sample	Scanner	Data		
Between Sample Reproducibility I	=	x	x	x	GSP 10 groups	3
Between Sample Reproducibility II	x	x	x	x	GSP 300 HNU Replication	2, 4
Between Session Reproducibility	=	=	=	x	HNU Reference & Replication GSP 10 groups TRT	5, 6, 7
Reliability	=	=	=	x	HNU Reference & Replication	8, 9, 10

Table 2- Analysis Descriptor. Level of Generalization depicts which type of generalizability we tested with each analysis, see Methods sections 2.5-2.6. Columns show which kind of parameters would be kept constant (=) or changed (x) across analyses in the current study.

2.5 Reproducibility

2.5.1 Concept- Reproducibility is the "closeness of agreement between results of measurements of the same measure and carried out under changed conditions of measurement" (Jcgm & Others, 2008).

2.5.2 Data Requirements- The changed conditions may include the method of measurement, observer, measuring instrument, reference standard, location, conditions of use, and time. Notably, statements of reproducibility must include description of the specific conditions changed. In other words, when discussing reproducibility, it's important to be specific about what is being changed and across which dimension(s) measurements are being reproduced. Reproducibility across samples requires different samples, but the same method and investigator can be used, which is called between-sample reproducibility. Reproducibility across sessions but not samples requires multiple assessments of the same sample, and is known as between-session reproducibility. Similarly, reproducibility across investigators requires different investigators but can use the same method and data, which is often referred to as computational reproducibility (Millman, Brett, Barnowski, & Poline, 2018).

To support our tests of between sample and between session reproducibility, we created several reference and replication datasets using the HNU and GSP test-retest datasets. In the HNU dataset, we created a test-retest dataset by combining the ten, 10-minute sessions into two five-session, 50-minute scans. Our reference and replication datasets were created through a concatenation of each participant's even and odd numbered sessions scans respectively. To assess the interaction of bagging and length of acquired data on reproducibility, we performed parcellation across the first 3, 5, 10, 15, 20, and 25, and 50 minutes of the replication data. We split the 300 subjects of the GSP data into 10 groups of 30 subjects that were age and sex matched to the HNU dataset, and we used these 10 groups to assess both between-sample

and between-session reproducibility. We also concatenated both 6 minute sessions and used all 300 subjects of the GSP dataset to create a large reference parcellation, and we compared this reference to the replication sets of the HNU data.

2.5.3 Testing Methods- To measure between-session reproducibility, we assessed the similarity between the reference and each of the replication datasets using spatial correlation of these group-level mean adjacency matrices, and the adjusted Rand Index of the group-level cluster labels. Given that bagged clustering is non-deterministic, we assessed the variance in our parcellations by repeating our parcellations 20 times for each replication dataset across each amount of scan time and bootstrap aggregation for the subcortical parcellation. For the cortical parcellation we repeated the parcellations 10 times. Since non-bagged clustering is deterministic, we needed a method to assess the variance in between-session reproducibility of the parcellations. To overcome this issue, we created 20 new datasets for each amount of the scan lengths by resampling the 50 minute RP dataset. See supplementary methods for details on this resampling method. We applied PyBASC without bagging across each of these 20 datasets to estimate the variance between-session reproducibility for the non-bootstrapped condition.

We assessed the effect of bagging on between-sample within-study reproducibility of functional parcellation by comparing the parcellations between 10 age and sex-matched groups of 30 subjects from the GSP dataset. Since we had 10 groups, we did not need to resample and repeat our bagged or non-bagged parcellations as above. We used each of the two 6 minute scans for each group, as well as a concatenated 12 minute scan. We compared the similarity of group-level parcellations between the 10 groups for each scan length. Second, we assessed the effect of bagging on between-sample between-study reproducibility. We used the same 12 minute scans from all 300 GSP subjects together to create a reference parcellation. For

assessing cortical parcellation reproducibility, we compared the GSP reference to the 10 minute HNU data. For assessing subcortical parcellation reproducibility we assessed the GSP parcellation against a range of HNU scan lengths.

2.6 Reliability & Discriminability

2.6.1 Concept- Reliability is a metric of the intra-individual stability of a measurement across multiple occasions (Zuo & Xing, 2014), and can be measured with many different indices. Intraclass correlation (ICC) is a descriptive statistic that relates the within-subject to between subject variance to get an indication of univariate reliability (Shrout & Fleiss, 1979). ICC is most commonly considered as the intra-individual variability relative to the inter-individual variability. Discriminability is multivariate metric of reliability that takes the full set of features across all observations into account to compute a multivariate index of reliability.

2.6.2 Data Requirements- Calculating test retest reliability requires at least two observations of the same individuals in the same group. In the current study, we compute ICC using two measurements from each individual in both the HNU and GSP samples.

2.6.3 Testing Methods- We assessed whether bagging would improve both the univariate and multivariate reliability of our functional parcellations, and how length of scan acquisition would have an impact on these results. Using the same set of functional parcellations across multiple datasets calculated in Section 2.5 Reproducibility, namely the reference and multi-length replication datasets, we calculated the ICC for each cell in the individual level mean adjacency matrix data, and discriminability for the individual mean adjacency matrix as a whole.

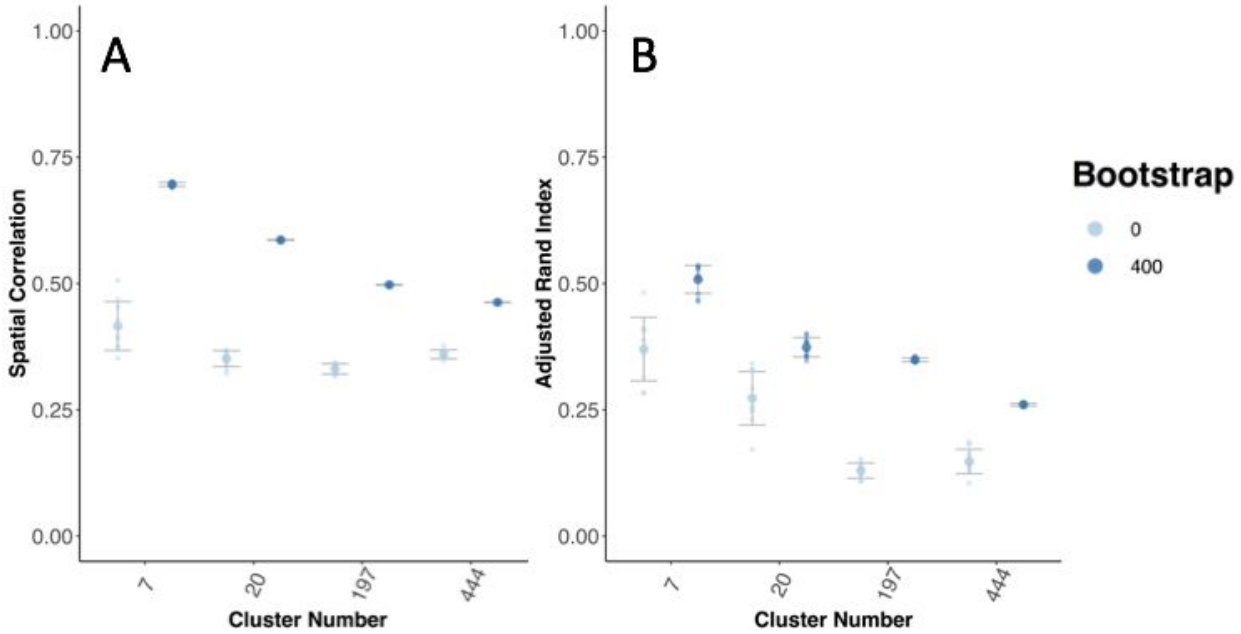

Figure 2- Bagging improves the reproducibility of cortical parcellation between the GSP and HNU datasets. We compared a reference parcellation created from the 12 minute combined scan from 300 individuals from the GSP to 10 repeated parcellations of 10 minutes of HNU data. A) Spatial correlation between the group mean adjacency matrix for the GSP and HNU parcellations. B) Adjusted Rand Index between the cluster labels for the GSP and HNU parcellations.

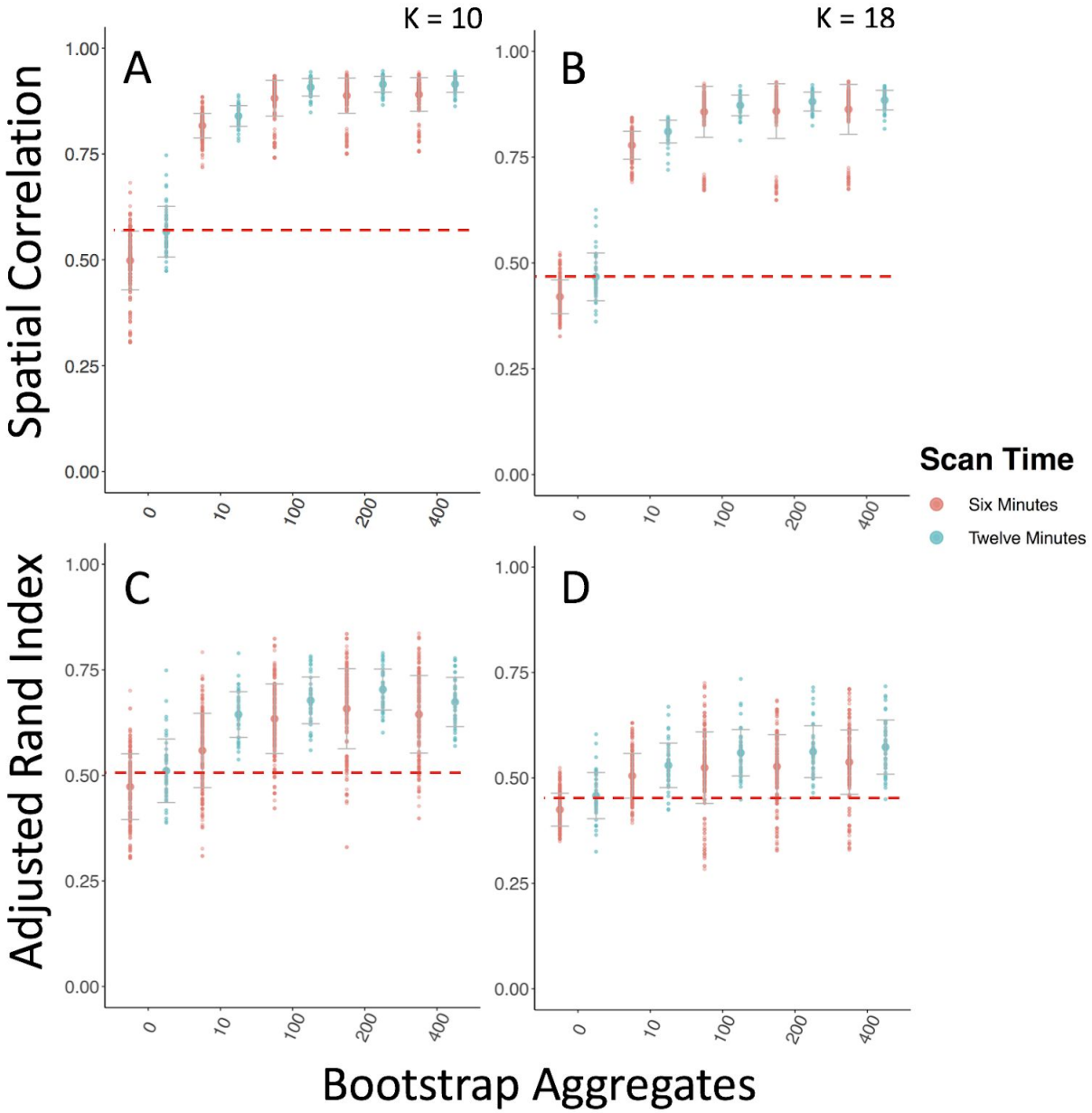

Figure 3- Bagging improves between sample within study reproducibility of subcortical parcellation. Comparing the parcellations from 10 groups of 30 subjects from the GSP sample to one another. Dotted red line shows the highest mean reproducibility of the standard parcellation approach. A-B Spatial correlation between mean group adjacency matrix of 10 GSP groups with 10 and 18 clusters. C-D Adjusted Rand Index between cluster labels of 10 GSP groups with 10 and 18 clusters.

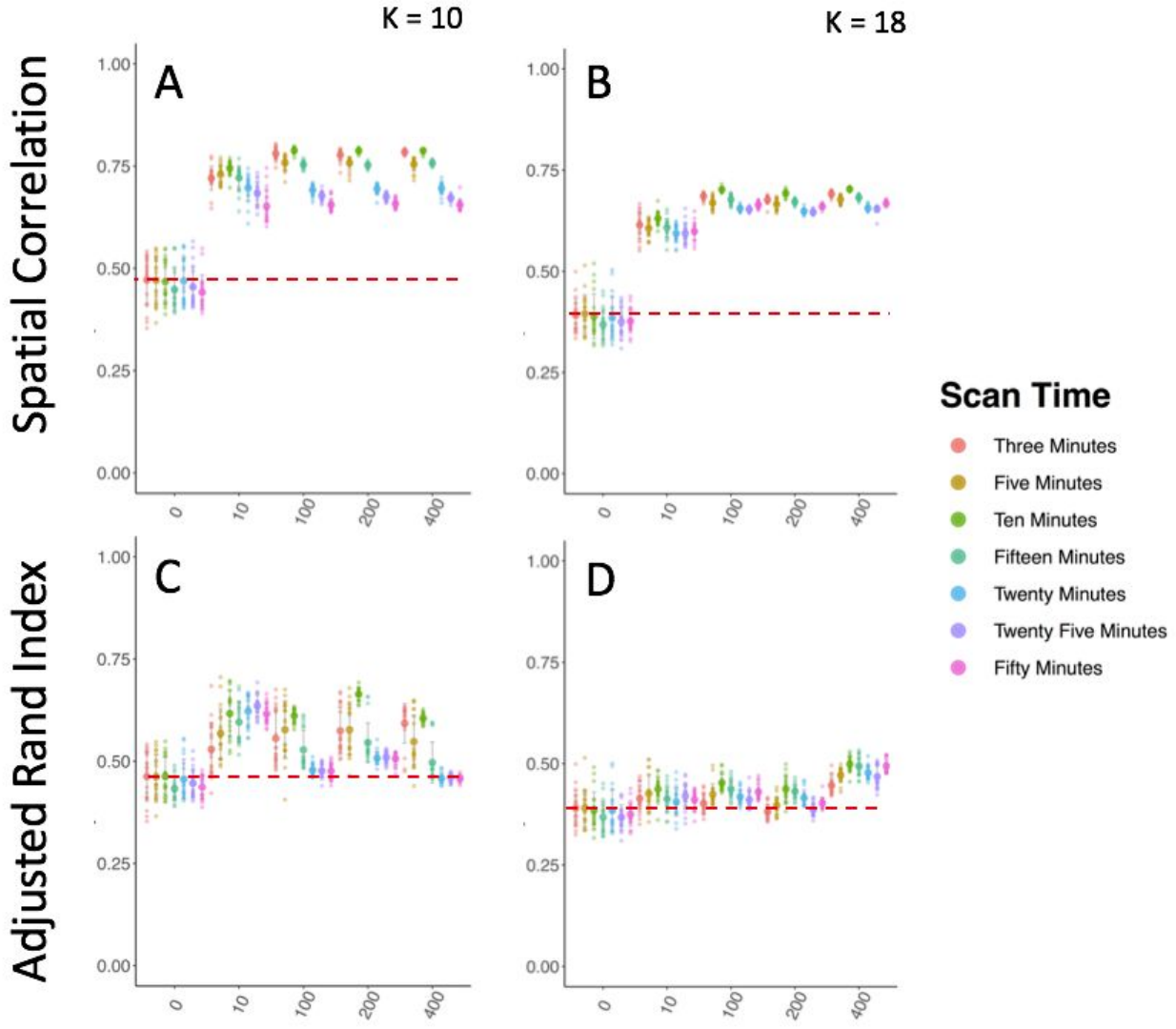

Figure 4: Bagging improves the reproducibility of subcortical brain parcellation between the GSP and HNU datasets, though results more mixed than the cortical results. We compared a reference parcellation created from the 12 minute combined scan from 300 individuals from the GSP to 10 repeated parcellations of 10 minutes of HNU data. A) Spatial correlation between the group mean adjacency matrix for the GSP and HNU parcellations. B) Adjusted Rand Index between the cluster labels for the GSP and HNU parcellations.

3 Results

3.1 Between-Sample Reproducibility

We first wanted to assess the impact of bagging on between-sample reproducibility, and we tested this impact using both subcortical and cortical parcellations on both HNU and GSP datasets. We hypothesized that bagging should improve the reproducibility of the parcellations, and that it may also have an impact on the consistency of the replication parcellations as well. We conducted a range of tests to assess these concepts. First, using the 300 individuals from the GSP study to create a reference cortical parcellation, we assessed the correlation of the group mean adjacency matrices (Figure 2 A), and the adjusted rand index (ARI) of the group cluster labels (Figure 2 B). We demonstrated that cortical parcellations were more reproducible between studies using bagging; for example, for K = 444, bagging improves ARI (Kruskal-Wallis test comparing 400 vs 0 bootstraps: chi-squared = 14.30; $p < 0.0005$). We also found that bagging improved the consistency of reproducibility, meaning that across runs bagging produced more cluster solutions whose varied less between one another (Test of unequal variance of ARI 400 vs 0 bootstraps: $F = 94.8$, $p < 0.0001$).

Second, we investigated the effect of bagging on the between-sample, within-study reproducibility of our subcortical parcellations (Figure 3). We split the 300 GSP subjects into 10 groups of 30, and computed the correlation of the group mean adjacency matrices (Figure 3 A, 3 B), and the ARI of their group cluster labels between groups. We found that while increasing the time of the scan from 6 to 12 minutes significantly improved reproducibility when measured by both correlation and ARI (K = 10; Correlation: Kruskal-Wallis chi-squared = 58.517, df = 1, p-value < 0.0001; ARI: Kruskal-Wallis chi-squared = 10.788, df = 1, p-value < 0.005), we found that the impact of bagging improved reproducibility more than increasing the scan length (K =

10; Correlation: Kruskal-Wallis chi-squared = 401.26, df = 1, p-value < 0.0001; ARI: Kruskal-Wallis chi-squared = 278.9, df = 1, p-value < 0.0001). We also found that bagging significantly decreased the variance in the reproducibility estimates for both 6 minute and 12 minute scans, but had a greater impact on the 12 minute scans (6 min ARI variance 0 vs 400 bootstraps: F = 0.716, num df = 267, denom df = 267, p-value < 0.01; 12 min ARI variance 400 vs 0 bootstraps F = 1.6712, num df = 89, denom df = 89, p-value < 0.05).

Third, we found that bagging had a significant impact on the between-sample-between study reproducibility of subcortical parcellation, comparing the GSP reference to the HNU dataset. Both correlation and ARI of the parcellations improves with scan time from 10 to 50 minutes as expected. However, we expected the parcellation labels to improve equally for both the 10 and 50 minute replication samples. What we found instead, is that while both improved significantly from 0 to 400 bootstraps, the 10 minute HNU dataset improved much more than the 50 minute (K = 10, 10 minute: t = -13.087, df = 21.389, p-value < 0.0001; 50 minute: t = -3.2407, df = 21.75, p-value < 0.005). When visualized (see supplemental materials), we see this difference is largely due to the fact that both the GSP reference and the 10 minute HNU scans split the bilateral caudate into two clusters, whereas the 50 minute HNU parcellation did not. In fact, to 50 minute HNU data yielded a a seemingly better quality parcellation overall, with less variance on the edges between parcels, and more homotopic similarity, which can be expected given homotopic connectivity is known to be especially high. Of note here, is that the clusters created have minimal spatial constraints applied to them, implying that the improvement in reproducibility does not come from the clusters being forced into the similar spatially constrained configuration. In fact we see many clusters in the HNU 50 minute Reference and Replication

solution that are anatomically distinct but functionally united bilateral homologues (ie the left and right putamen cluster together, the left and right caudate cluster together, etc).

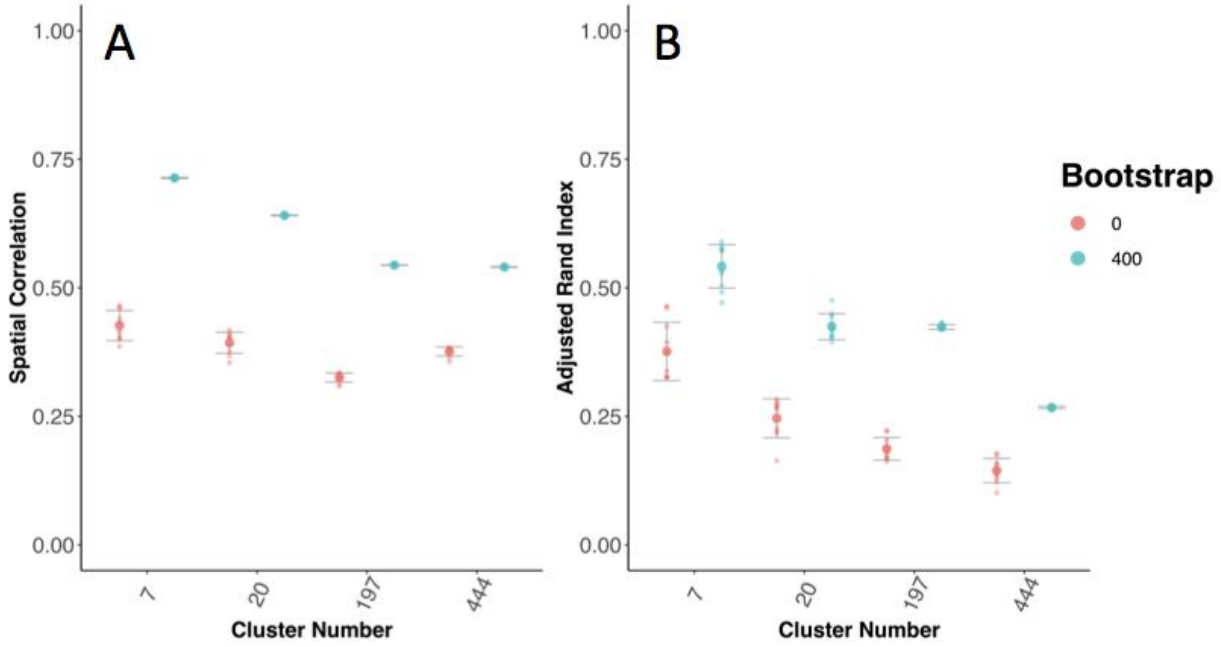

Figure 5- Bagging improves the between-session reproducibility of cortical parcellation in the HNU dataset. We compared a reference (Ref) parcellation created from a 50 minutes of scan time from the even numbered sessions to the HNU replication (Rep) dataset, created from 50 minutes of scan time from the odd numbered sessions. A) Spatial correlation between the group mean adjacency matrix for the Ref and Rep HNU parcellations. B) Adjusted Rand Index between the cluster labels for the the Ref and Rep HNU parcellations.

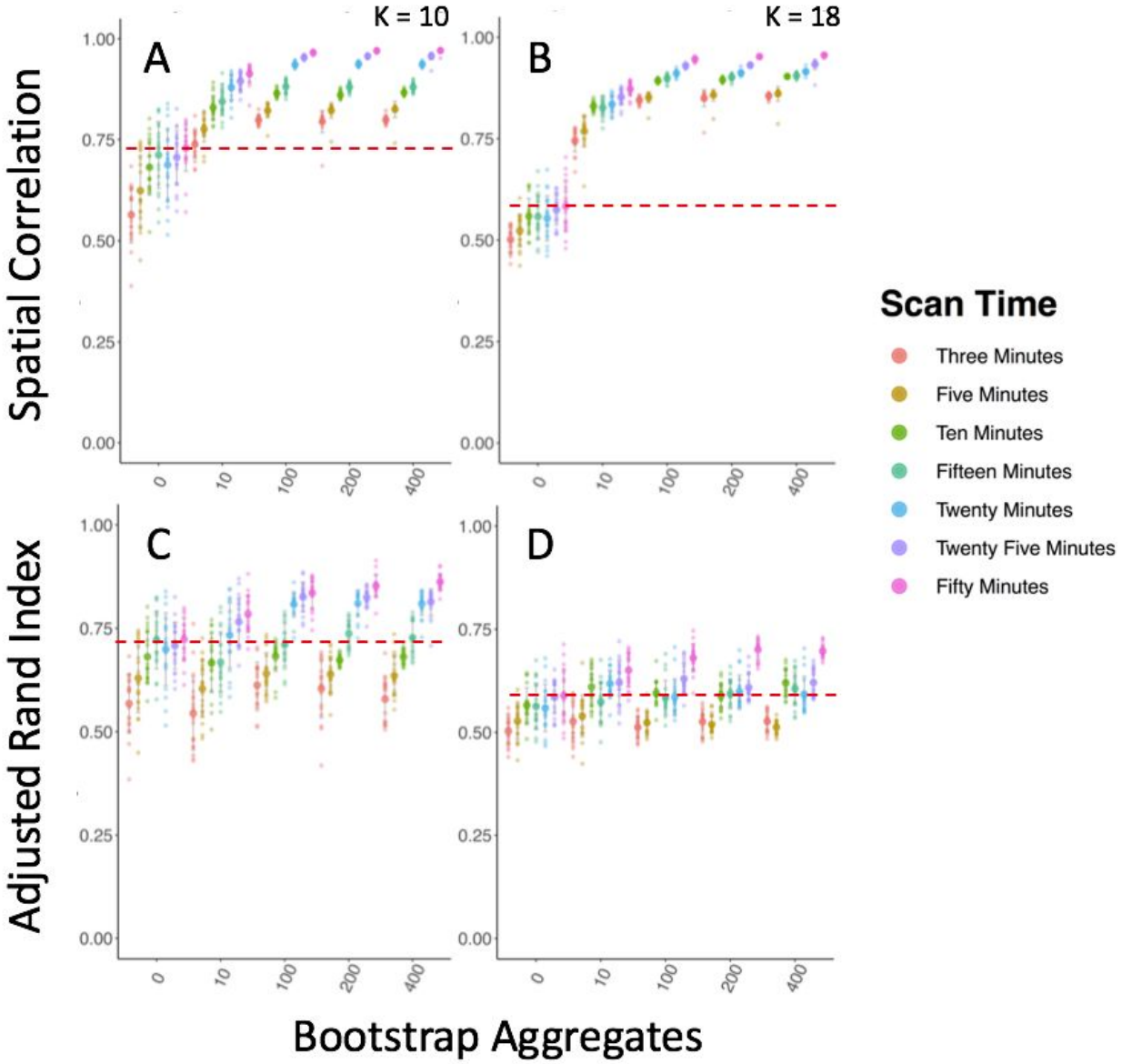

Figure 6 Bagging improves the between-session reproducibility of subcortical parcellation in the HNU dataset. We compared a reference (Ref) parcellation created from a 50 minutes of scan time from the even numbered sessions to the HNU replication (Rep) dataset, created from 50 minutes of scan time from the odd numbered sessions. A) Spatial correlation between the group mean adjacency matrix for the Ref and Rep HNU parcellations. B) Adjusted Rand Index between the cluster labels for the the Ref and Rep HNU parcellations.

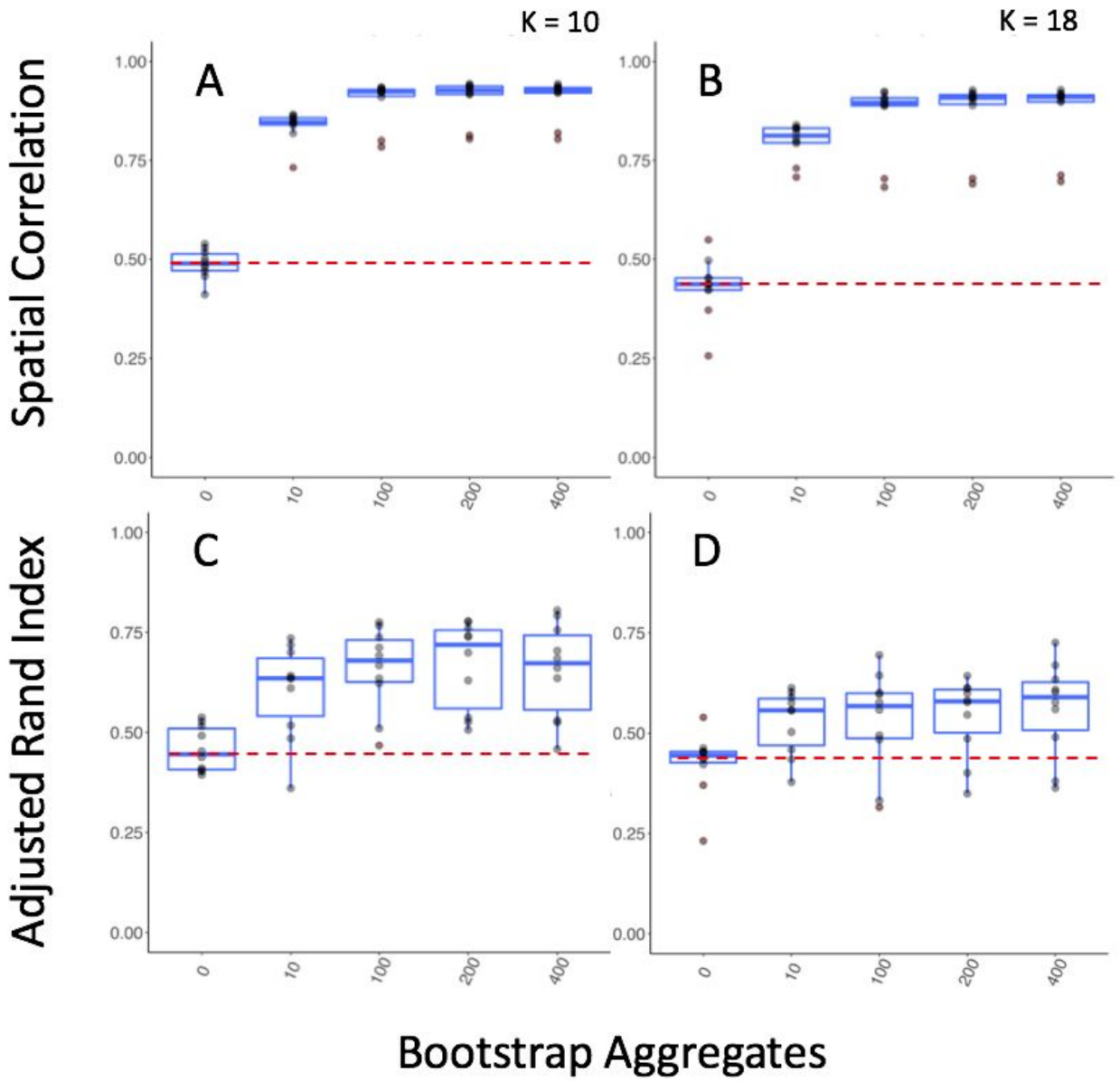

Figure 7 Bagging improves between-session reproducibility of subcortical parcellation in the GSP dataset. We split the 300 subjects of the GSP dataset into 10 groups of 30, and compare the parcellations of their first and second functional scan to one another. A-B show the improvement in bagged parcellations measured by spatial correlation of the group mean adjacency matrix for 10 and 18 clusters respectively. C-D show the improvement in bagged parcellations measured by the adjusted Rand index of the cluster labels for 10 and 18 clusters respectively.

3.2 Between-Session Reproducibility

Whereas creating parcellations with high between-sample reproducibility is key for generalizing scientific discovery, we wanted to assess whether bagging could help create parcellations with higher between-session reproducibility, which is key for decreasing uncertainty in measuring indications of change over time in a sample of interest. We tested between-session reproducibility using both the subcortical and cortical parcellations of the HNU and GSP datasets. First, we found that our cortical parcellations had higher between-session reproducibility with bagging than without in the HNU dataset (Figure 5). Comparing the K=444 parcellation, we see significant improvements in both between-session reproducibility of group mean adjacency matrices ($t = -57.86$, $df = 9.2157$, p-value = < 0.0001), and group cluster labels ($t = -16.353$, $df = 9.1687$, p-value < 0.0001). If parcellations have higher between-session reproducibility on average, but have high variance across repeated parcellations we might not have a good grasp on the extent to which a particular parcellation will match across time. Therefore, it's also important for the variance in the between-session reproducibility of our parcellations to decrease. We found that bagging also significantly decreases the variance in between-session reproducibility of the group mean adjacency matrix ($F = 83.456$, num df = 9, denom df = 9, p-value < 0.0001).

Second, we also tested the effect of bagging on the between-session reproducibility of subcortical parcellations in the HNU dataset (Figure 6). We found that both bagging and scantime length have a highly significant impact on the between-session reproducibility of parcellations. For instance, comparing 5 and 50 minute group mean adjacency matrix correlation and cluster label ARI with 0 bootstraps (K=10; Correlation: Kruskal-Wallis chi-squared = 14.137, df = 1, p-value < 0.0005; ARI: Kruskal-Wallis chi-squared = 12.176, df =

1, p-value < 0.0005), and with 400 bootstraps (K=10; Correlation: Kruskal-Wallis chi-squared = 29.268, df = 1, p-value <0.0001; ARI: Kruskal-Wallis chi-squared = 29.271, df = 1, p-value <0.0001). Notably, the impact of time was enhanced in the bagging condition, suggesting bagging and scan time can work synergistically to provide parcellations with high between-session reproducibility. We also see that the improvement in between-session reproducibility from bagging is significantly greater than increasing scan length. For example, for the 0 bootstrap condition, increasing the scan time from 20 minutes to 50 minutes does not yield a significant improvement in between-session reproducibility of group mean adjacency matrix correlation or ARI (K=10; Correlation: Kruskal-Wallis chi-squared = 2.9041, df = 1, p-value >0.05; ARI: Kruskal-Wallis chi-squared = 0.79683, df = 1, p-value > 0.05), but increasing the bagging from 0 to 400 has a significant effect on the between-session reproducibility of the 20 minute data (K=10; Correlation: Kruskal-Wallis chi-squared = 29.268, df = 1, p-value < 0.0001; ARI: Kruskal-Wallis chi-squared = 17.579, df = 1, p-value < 0.0001). We also found that bagging had a significant impact on reducing the variance in between-session reproducibility across multiple parcellations, replicating our results with the cortical parcellation (20 Minutes; K = 10; Correlation: F = 311.14, num df = 19, denom df = 19, p-value < 0.0001; ARI: F = 9.7003, num df = 19, denom df = 19, p-value < 0.0001).

While the cortical parcellation demonstrated that only 10 minutes of data was needed for bagging to improve between-session reproducibility in the HNU data, with the subcortical parcellation we saw that bagging did not improve the reproducibility for less than 20 minutes of data. This may be due to random elements of sampling those 3, 5, 10, or 15 minute portions of the HNU dataset, or differences in the parcellation favored by lower amounts of data compared to higher amounts, as we saw in comparing the HNU 50 minute parcellation to the GSP reference parcellation in section 3.1. However, we found that with the GSP dataset, where we

compare the effect of bagging on between-session reproducibility of 10 separate groups of 30 subjects, we see that bagging does provide a significant impact in between-session reproducibility with as little as 5 minutes of data (Figure 7), suggesting that long scans are not required for bagging to have a significantly beneficial impact on parcellation between-session reproducibility of the group mean adjacency matrix and cluster labels (K = 10; Correlation: Kruskal-Wallis chi-squared = 14.286, df = 1, p-value < 0.0005; ARI: Kruskal-Wallis chi-squared = 10.566, df = 1, p-value < 0.005). We found that while between-session reproducibility was improved with bagging, these reproducibility estimates were not improved as was the case with longer datasets (K = 10; Correlation: F = 0.5374, num df = 9, denom df = 9, p-value > 0.05; ARI: F = 0.22301, num df = 9, denom df = 9, p-value < 0.05).

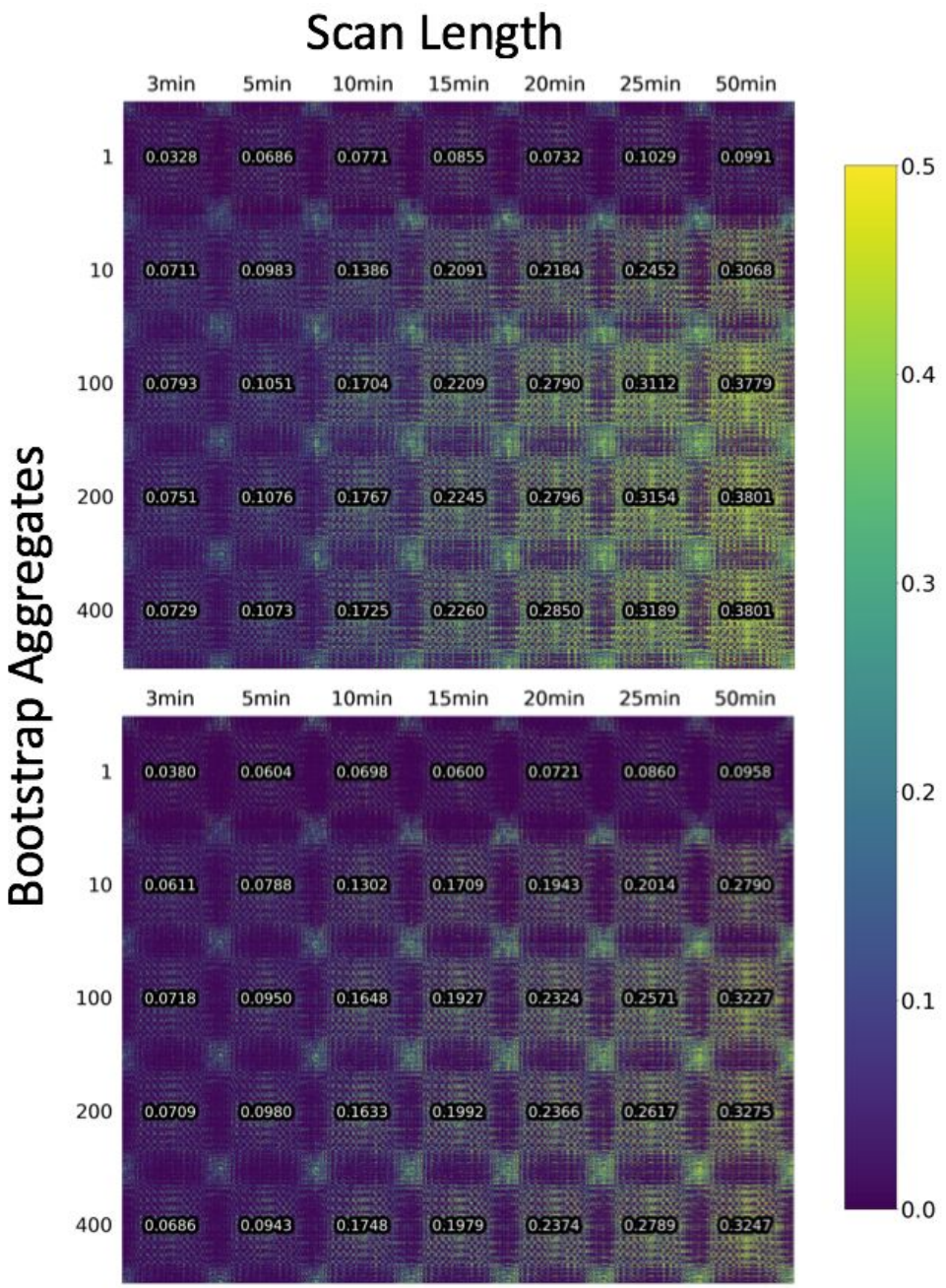

Figure 8 Univariate Reliability. Bagging improves the ICC of parcellations. The top and bottom correspond to the K = 10 and K = 18 results respectively. The rows show the different numbers of bootstrap aggregates, and the columns show the length of time used. Each cell shows the average ICC for all voxel-voxel pairs in the clustering.

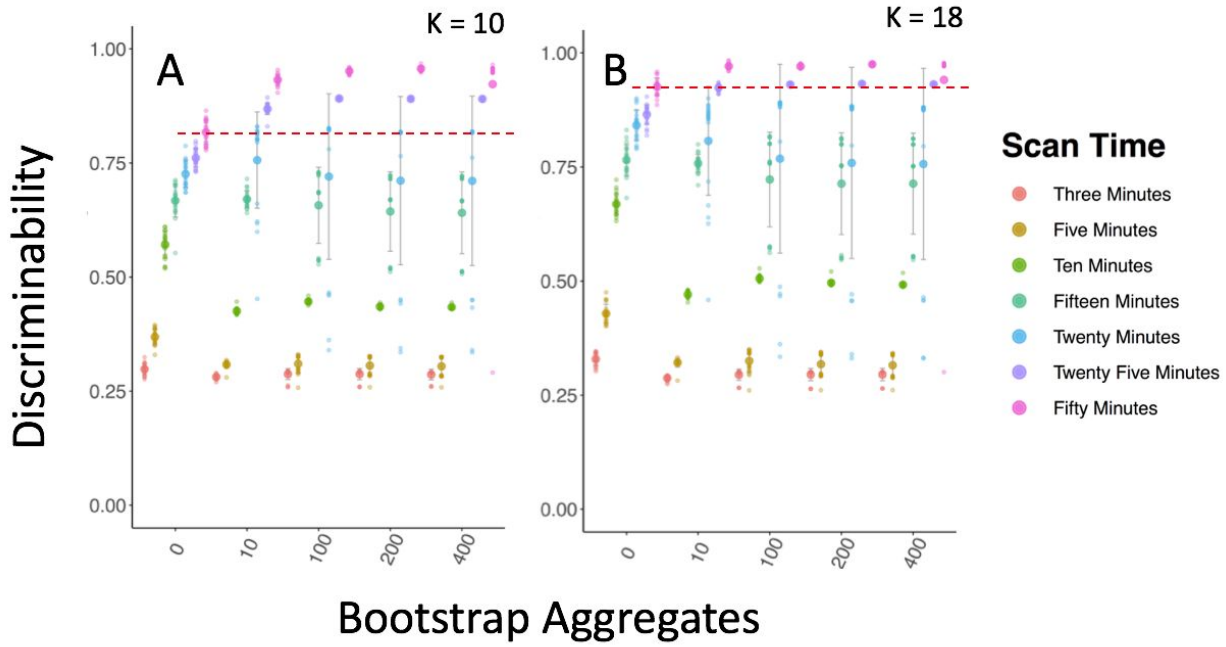

Figure 9 Multivariate Reliability. Figure A and B show the discriminability results for the subcortical parcellation for K = 10 and K = 18 respectively. The dashed red line shows the top mean performance for the non-bagged parcellations. Bagging leads to improved discriminability for longer scan times.

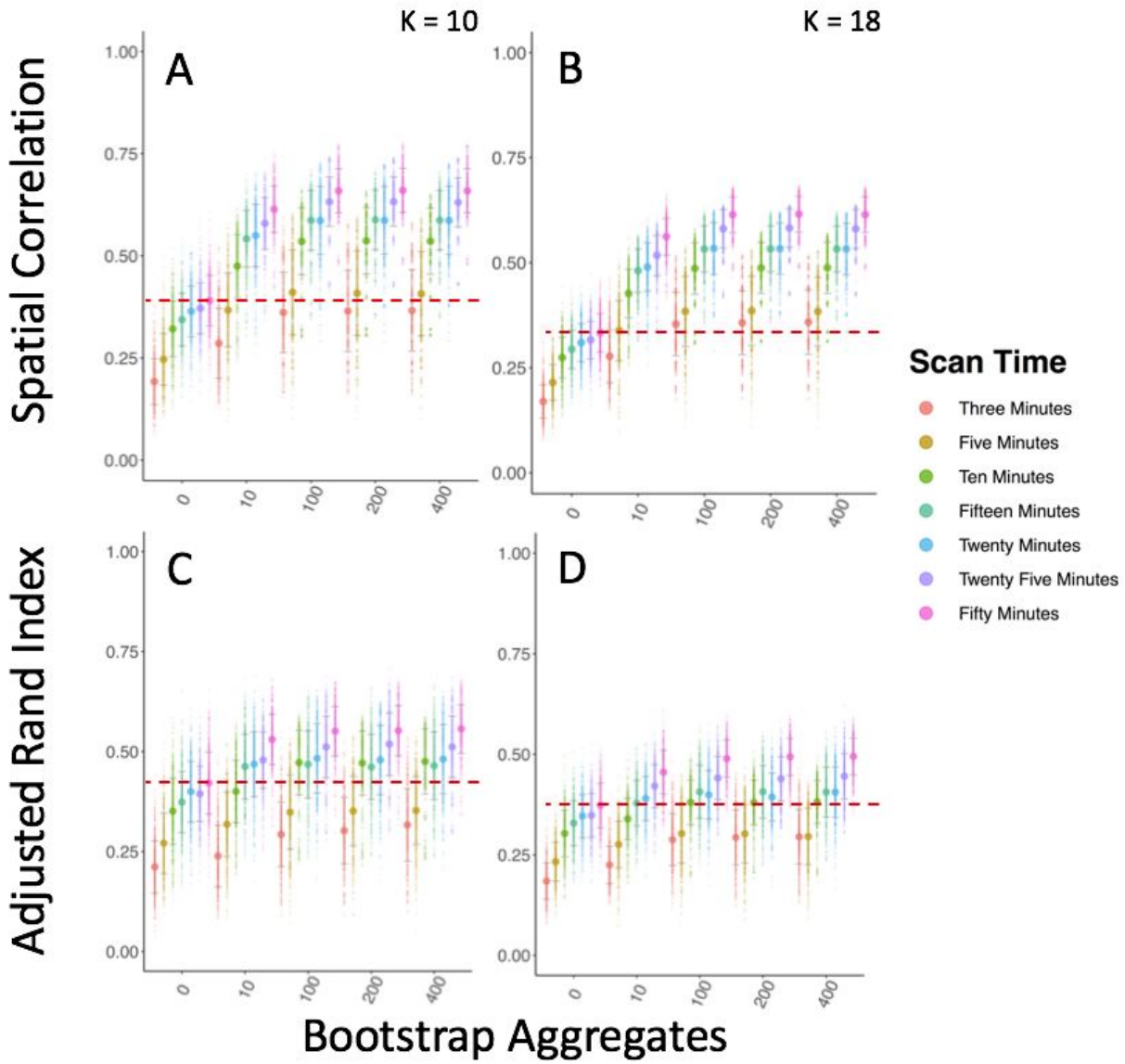

Figure 10 - Overall this figure shows how bagging improves coherence between individual and group parcellations. A and B show the improvement in correlation between the group-mean adjacency matrix and each individual's mean adjacency matrix for the K=10 and K=18 subcortical parcellation respectively. C & D show the improvement of the adjusted Rand Index between the clusters labels of the group and individual level parcellations.

3.3 Reliability

We found that both scan length and bagging led to improvements in the univariate estimates of ICC of the voxel-wise adjacency matrices (Figure 8). Each of the entries in this matrices is the number of times across individual-level bootstraps that two voxels are put into the same cluster. As such, the individual mean adjacency matrix is a representation of the average clustering solution across bootstrap aggregations for a given individual. This is a noisy measure that is improved substantially by bagging across participants to create group-level parcellations. However, we see that bagging improves the ICC of the mean adjacency matrix more than increasing scan time. For example, for K = 10, average ICC improves from 0.07 to 0.09 when increasing scantime from 20 to 50 minutes with 0 bootstraps, whereas 20 minute ICC improves from 0.07 to 0.285 from 0 to 400 bootstraps. The individual voxel-voxel ICCs were low-to-moderate, reflecting that each of the voxel-voxel relationships that make up the individual parcellations may not be reliable; however, it is still possible that when taken as a whole, the parcellation patterns in these adjacency matrices would be reliably distinct between individuals. We tested this using the multivariate metric of reliability called discriminability. This method allows us to consider the extent to which individual parcellation patterns are unique to that individual compared to the rest of the group. We found that discriminability of the parcellations for 0 bootstraps was heavily impacted by the amount of data acquired, and was quite high for 25 and 50 minutes of data (25 minute mean = 0.761; 50 minute mean = 0.817). We found that discriminability was significantly increased through bagging but not for all lengths of data. It seems that individual-level signal was not strong enough in 3-15 minutes of data to improve the discriminability of parcellation through bagging indata (all $p > 0.05$); however we did see significant improvements from 0 to 400 bootstraps for 20 minutes (Kruskal-Wallis chi-squared = 7.1762, df = 1, p-value < 0.01), 25 minutes (Kruskal-Wallis chi-squared = 29.44, df = 1, p-value

< 0.0001); and 50 minutes (Kruskal-Wallis chi-squared = 24.584, df = 1, p-value < 0.0001), with average discriminability reaching even higher for 20 (mean = 0.71), 25 (mean = 0.89), and 50 (mean = 0.92) minutes.

Finally, when making generalizations from group-level information to the individual, it's critically important that the group-level information is representative of the individual, a concept known as ergodicity (Adolf & Fried, 2019; Fisher, Medaglia, & Jeronimus, 2018). Such group-to-individual generalizability is critical for deploying any group-defined models on individual level data, and as such is a central for any kind of biomarker discovery process. Given that recent work has demonstrated that individual-level brain networks can differ substantially from group level estimates (Gordon, Laumann, Adeyemo, & Petersen, 2015; Laumann et al., 2015) we wanted to test whether bagging could improve the group-to-individual generalizability of parcellation, improving their potential for scientific and clinical applications. We found that bagging led to significantly more generalizable parcellations (Twenty minute parcellation; 0 vs 400 bootstraps; K = 10; Correlation: Kruskal-Wallis chi-squared = 800.81, df = 1, p-value < 0.0001; ARI: Kruskal-Wallis chi-squared = 248.43, df = 1, p-value < 0.0001), and increasing scan time did as well (K=10; Correlation: Kruskal-Wallis chi-squared = 50.985, df = 1, p-value <0.0001; ARI: Kruskal-Wallis chi-squared = 16.593, df = 1, p-value <0.0001). We also found that bagging led to an overall decrease in the variance of the group-individual similarity (K = 10; Correlation:F = 0.56163, num df = 599, denom df = 599, p-value = 2.404e-12; ARI: F = 0.85283, num df = 1199, denom df = 1199, p-value = 0.005884), but this decreased variance was less pronounced here than in other analyses.

4. DISCUSSION

4.1 Overview- We find that bagging improves the reproducibility of group-level functional parcellations with as little as five minutes of data, providing a potentially valuable means of bringing down data requirements for reproducible parcellation. Bagging enhanced parcellations are also more reliable on the individual level, and with long enough scans we see increases in the discrimination of individual-level parcellations, meaning that the more detailed, individual-specific attributes of parcellations can be enhanced through bagging. We find that bagging enhanced parcellations also have higher between-session reproducibility, meaning that they are better suited for within-sample repeated measures that are commonly used in intervention or other longitudinal studies. Finally, we found that bagging improved group-to-individual generalization of parcellations, which is key for using any common parcellation across participants. Overall, this indicates that bagging enhanced parcellations outperform the standard approach on a wide variety of reproducibility, and reliability indices, and should be considered for further implementation in cutting-edge parcellation approaches for potential improvements in making more robust measurements of the connectome (Glasser et al., 2016; Laumann et al., 2015; Xu et al., 2016).

4.2 Implications Given that bagging can have positive impacts on multiple measures of reproducibility and reliability, this may have important implications for analyses using clustering in MRI data. The cluster ensemble literature has shown that bagging and other methods may improve clustering. Other approaches include combining cluster solutions across multiple numbers of clusters (Kuncheva & Hadjitodorov, 2004), varying random parameters of the clustering algorithm, or combining multiple clustering techniques (Hu & Yoo, 2004; Lancichinetti & Fortunato, 2012). This implies that there may be other avenues to pursue ways in which both ensemble methods or bagging can be used to improve the generalizability of functional

parcellations and other MRI approaches that use clustering. For example, prior work has demonstrated the utility of using clustering for tissue class segmentation in lesion detection (Bosc et al., 2003; Juang & Wu, 2010), and these efforts may be furthered by the use of cluster ensemble methods for detecting and evaluating differences lesions more reproducibly across samples and timepoints.

The impact of bagging on reproducibility observed between completely independent samples is particularly promising, as it suggests a means of improving the generalizability of parcellation atlases across sites, scanners, and populations. The choice of a parcellation is an important step in the creation of a connectome, and it has a significant impact on the sensitivity of the resulting connectome to phenotypic differences (Abraham et al., 2017). Recent work has demonstrated that up to 60% of the variance in the edges of a connectome derive from the variance in the match of the reference parcellation to the data in question (Bijsterbosch, Beckmann, Woolrich, Smith, & Harrison, 2019); this is problematic because when parcellations do not generalize well between samples, the resulting connectome edges will conflate differences between groups network structure with differences in underlying parcellation fit (Zuo & Xing, 2014). In this way, enhancing the reproducibility of the parcellation fit between samples may also increase sensitivity of network edges to phenotypic differences between groups and individuals. This is an important step for creating the robust brain-behavior associations required for using brain networks as biomarkers.

Moreover, demonstrations of the current work that bagging improves the reliability and reproducibility of functional parcellations suggests that bagging may be an important method to explore further for biomarker development. With improved reliability of parcellations we may be able to also elucidate brain-behavior relationships with more robustness than previously possible. The strength of the association between two variables is limited to the reliability of the

variables in question (Zuo and Xing 2014). With improvements in between-session reproducibility of measurement we improve the ability to clinically assess changes over time that may occur in a clinical trial or behavioral intervention, or in the case of a longitudinal development study, such as the ongoing ABCD study (Volkow et al., 2018). As measurement error decreases over time, it becomes more feasible for estimating the change over time that occurs as the result of a process of interest rather than uncontrolled variance.

An important recurrent finding was that bagging decreased the variance in the reproducibility estimates. While the importance of improving reproducibility of parcellations are key for improving the fit of a parcellation across samples or over time, if the improvements are highly variable then it's unclear when a procedure such as bagging is helping or hurting for a given case. Seeing that bagging significantly reduced this variance is not only indicative that it is more likely to produce reproducible parcellations, but also that each time it is applied to a given dataset these parcellations are likely to be more in line with one another. We show that cluster ensembles have a better grasp on the central tendency of the data for each participant and the group as a whole, and by demonstrating that bagging improves the consistency of parcellations we show that bagging also helps prevent small perturbations in the data from having outsized effects on the parcellation results.

4.3 Limitations- The current manuscript is a work in progress, and there are several additional elements that deserve careful consideration. For example, we have not properly considered here the impact of motion on the reliability estimates, nor have we assessed the impact of individual level bagging versus group level bagging. It is also worth noting that the functional parcellations created in the current work are not intended to be new reference atlases that are more reproducible than others. Instead, we intend to demonstrate that bagging can improve the

fit and generalization of functional parcellations. Most likely, the application of such cluster ensemble methods on multi-modal structural and functional data would lead to further improvements on the test-retest reliability of the parcellation, such as those applied in recent advanced efforts (Glasser et al., 2016; Xu et al., 2016). Another limitation of the current study is that we did not exhaustively test the space of cluster methods, even though PyBASC inherently is capable of such extensions. Bagging and cluster ensembles are notable for being agnostic to clustering technique and have been successfully demonstrated in a wide range of clustering techniques however, and we feel this limitation is notable but not a significant one.

4.4 Future Directions- The current effort aimed at establishing the role of bagging in improving reproducibility and reliability of functional parcellations, but the number of tests were far from exhaustive. There are more ways that bagging and cluster ensembles could be leveraged to improve measurements of the brain. For instance, varying the time series window length, cluster technique, or resampling method may lead to better individual-level and group-level estimates. Some research has demonstrated that selecting diverse sets of cluster solutions for aggregation can outperform simply aggregating all cluster solutions. Such a diversity selection approach could also be employed in a similar context and may likely improve reproducibility and reliability of the cluster solutions as well.

References

Abraham, A., Milham, M. P., Di Martino, A., Cameron Craddock, R., Samaras, D., Thirion, B., & Varoquaux, G. (2017). Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage*, Vol. 147, pp. 736–745. https://doi.org/10.1016/j.neuroimage.2016.10.045

Adolf, J. K., & Fried, E. I. (2019). Ergodicity is sufficient but not necessary for group-to-individual generalizability. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(14), 6540–6541.

Barnes, K. A., Cohen, A. L., Power, J. D., Nelson, S. M., Dosenbach, Y. B. L., Miezin, F. M., … Schlaggar, B. L. (2010). Identifying Basal Ganglia divisions in individuals using resting-state functional connectivity MRI. *Frontiers in Systems Neuroscience*, *4*(June), 18.

Behrens, T. E. J., Johansen-Berg, H., Woolrich, M. W., Smith, S. M., Wheeler-Kingshott, C. A. M., Boulby, P. A., … Matthews, P. M. (2003). Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nature Neuroscience*, Vol. 6, pp. 750–757. https://doi.org/10.1038/nn1075

Bellec, P., Rosa-neto, P., Lyttelton, O. C., Benali, H., & Evans, A. C. (2010). NeuroImage Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *NeuroImage*, *51*(3), 1126–1139.

Bijsterbosch, J. D., Beckmann, C. F., Woolrich, M. W., Smith, S. M., & Harrison, S. J. (2019). The relationship between spatial configuration and functional connectivity of brain regions revisited. *eLife*, *8*. https://doi.org/10.7554/eLife.44890

Bosc, M., Heitz, F., Armspach, J. P., Namer, I., Gounot, D., & Rumbach, L. (2003). Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. *NeuroImage*, *20*(2), 643–656.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Choi, E. Y., Yeo, B. T. T., & Buckner, R. L. (2012). The organization of the human striatum estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *108*(8), 2242–2263.

Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., & Mayberg, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, *33*(8), 1914–1928.

Dudoit, S., & Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* , *19*(9), 1090–1099.

Fischer, B., & Buhmann, J. M. (2003). Bagging for path-based clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(11), 1411–1415.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(27), E6106–E6115.

Garcia-Garcia, M., Nikolaidis, A., Bellec, P., Craddock, R. C., Cheung, B., Castellanos, F. X., & Milham, M. P. (2017). Detecting stable individual differences in the functional organization of the human basal ganglia. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2017.07.029

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., … Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Gordon, E. M., Laumann, T. O., Adeyemo, B., & Petersen, S. E. (2015). Individual Variability of the System-Level Organization of the Human Brain. *Cerebral Cortex* , bhv239.

Holmes, A. J., Hollinshead, M. O., O'Keefe, T. M., Petrov, V. I., Fariello, G. R., Wald, L. L., … Buckner, R. L. (2015). Brain Genomics Superstruct Project initial data release with

structural, functional, and behavioral measures. *Scientific Data*, *2*, 150031.

Hong, Y., Kwong, S., Wang, H., & Ren, Q. (2009). Resampling-based selective clustering ensembles. *Pattern Recognition Letters*, *30*(3), 298–305.

Hoyos-Idrobo, A., Schwartz, Y., Varoquaux, G., & Thirion, B. (2015). Improving sparse recovery on structured images with bagged clustering. *Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop on*, 73–76. IEEE.

Hu, X., & Yoo, I. (2004). Cluster Ensemble and Its Applications in Gene Expression Analysis. *Proceedings of the Second Conference on Asia-Pacific Bioinformatics - Volume 29*, 297–302. Darlinghurst, Australia, Australia: Australian Computer Society, Inc.

Janssen, R. J., Jylänki, P., Kessels, R. P. C., & van Gerven, M. A. J. (2015). Probabilistic model-based functional parcellation reveals a robust, fine-grained subdivision of the striatum. *NeuroImage*, *119*, 398–405.

Jcgm, J., & Others. (2008). Evaluation of measurement data—Guide to the expression of uncertainty in measurement. *Int. Organ. Stand. Geneva ISBN*, *50*, 134.

Jia, J., Xiao, X., Liu, B., & Jiao, L. (2011). Bagging-based spectral clustering ensemble selection. *Pattern Recognition Letters*, *32*(10), 1456–1467.

Johansen-Berg, H., Behrens, T. E. J., Sillery, E., Ciccarelli, O., Thompson, A. J., Smith, S. M., & Matthews, P. M. (2005). Functional–Anatomical Validation and Individual Variation of Diffusion Tractography-based Segmentation of the Human Thalamus. *Cerebral Cortex* , *15*(1), 31–39.

Juang, L.-H., & Wu, M.-N. (2010). MRI brain lesion image detection based on color-converted K-means clustering segmentation. *Measurement*, *43*(7), 941–949.

Kelly, A. M. C., Di Martino, A., Uddin, L. Q., Shehzad, Z., Gee, D. G., Reiss, P. T., … Milham, M. P. (2009). Development of Anterior Cingulate Functional Connectivity from Late Childhood

to Early Adulthood. *Cerebral Cortex* , *19*(3), 640–657.

Kelly, C., Toro, R., Di Martino, A., Cox, C. L., Bellec, P., Castellanos, F. X., & Milham, M. P. (2012). A convergent functional architecture of the insula emerges across imaging modalities. *NeuroImage*, *61*(4), 1129–1142.

Kuncheva, L. I., & Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, *2*, 1214–1219 vol.2. ieeexplore.ieee.org.

Lancichinetti, A., & Fortunato, S. (2012). Consensus clustering in complex networks. *Scientific Reports*, *2*, 336.

Laumann, T. O., Gordon, E. M., Adeyemo, B., Snyder, A. Z., Joo, S. J., Chen, M.-Y., … Petersen, S. E. (2015). Functional System and Areal Organization of a Highly Sampled Individual Human Brain. *Neuron*, *87*(3), 657–670.

Li, T., & Ding, C. (2008). Weighted Consensus Clustering. In *Proceedings*. *Proceedings of the 2008 SIAM International Conference on Data Mining* (pp. 798–809). Society for Industrial and Applied Mathematics.

Margulies, D. S., Kelly, A. M. C., Uddin, L. Q., Biswal, B. B., Castellanos, F. X., & Milham, M. P. (2007). Mapping the functional connectivity of anterior cingulate cortex. *NeuroImage*, *37*(2), 579–588.

Mejia, A. F., Nebel, M. B., Shou, H., Crainiceanu, C. M., Pekar, J. J., Mostofsky, S., … Lindquist, M. A. (2015). Improving reliability of subject-level resting-state fMRI parcellation with shrinkage estimators. *NeuroImage*, *112*, 14–29.

Millman, K. J., Brett, M., Barnowski, R., & Poline, J.-B. (2018). Teaching Computational Reproducibility for Neuroimaging. *Frontiers in Neuroscience*, *12*, 727.

Noble, S., Spann, M. N., Tokoglu, F., Shen, X., Constable, R. T., & Scheinost, D. (2017).

Influences on the Test–Retest Reliability of Functional Connectivity MRI and its Relationship with Behavioral Utility. *Cerebral Cortex* , *27*(11), 5415–5429.

O'Connor, D., Potler, N. V., Kovacs, M., Xu, T., Ai, L., Pellman, J., … Milham, M. P. (2017). The Healthy Brain Network Serial Scanning Initiative: a resource for evaluating inter-individual differences and their reliabilities across scan conditions and sessions. *GigaScience*, *6*(2), 1–14.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations-uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Strehl, A., & Ghosh, J. (2002). Cluster Ensembles --- A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research: JMLR*, *3*(Dec), 583–617.

Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J., … Weiss, S. R. B. (2018). The conception of the ABCD study: From substance use to a broad NIH collaboration. *Developmental Cognitive Neuroscience*, *32*, 4–7.

Xu, T., Opitz, A., Craddock, R. C., Wright, M. J., Zuo, X.-N., & Milham, M. P. (2016). Assessing Variations in Areal Organization for the Intrinsic Brain: From Fingerprints to Reliability. *Cerebral Cortex* , *26*(11), 4192–4211.

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., … Buckner, R. L. (2011). *The organization of the human cerebral cortex estimated by intrinsic functional connectivity*. 1125–1165.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press.

Zuo, X.-N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., … Milham, M. P. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific Data*, *1*, 140049.

Zuo, X.-N., Biswal, B. B., & Poldrack, R. A. (2019). *Reliability and Reproducibility in Functional Connectomics*. Frontiers Media SA.

Zuo, X.-N., & Xing, X.-X. (2014). Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: a systems neuroscience perspective. *Neuroscience and Biobehavioral Reviews*, *45*, 100–118.

Zuo, X.-N., & Xing, X.-X. (2014). Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: a systems neuroscience perspective. *Neuroscience and Biobehavioral Reviews*, *45*, 100–118.