Alternative metrics (I2C2 and Kernel) were not used because they fail to offer significant advantage over the selected measures.

\Introduction

One of the foremost goals of neuroimaging work is to identify biomarkers of psychiatric illnesses – signatures or patterns in data that can help predict symptoms onset or suggest possible mechanisms. Resting state functional connectivity is a key part of this search, as many have suggested that it may contain useful information on internal neural function. However, clinically useful neuroimaging remains out of reach. There is growing concern that the reliability of neuroimaging data is critical limitation of its utility. The “crisis of reproducibility” has been especially incisive in this field due to results that indicate consistently poor reliability in these data.

Test-retest reliability is a term within generalizability theory that describes the stability of a value over repeated tests. Low test-retest reliability suggests that an observed value may be far from its “ground truth” value due to variance captured within the test. As such, reliability places a limit on the validity of an observed association, because the variance in the observed data is unrelated to the true values.

The measurement of test-retest reliability can be done through different metrics, each of which reflect unique forms of reliability subject to interpretation. These metrics can be univariate – looking at the reliability of each test item or measurement individually, or multivariate – looking at the stability of the overall test or scan.

We aim to determine the test-retest reliability of functional connectivity in a cohort of adolescents with and without depression. By characterizing the stability of their connectomes over a one-year period, we can begin to understand how age and psychiatric illness might affect reliability. In addition, we apply a range of analyses to explore how different metric may reflect different facets of reliability and offer unique perspectives on the data.

\subsection\*{Correlations with clinical measures and motion}

\noindent We also compared these continuous measures to the clinical questionnaires administered at baseline and after one year. These included the Mood and Feelings Questionnaire (MFQ), SHAPS, ARI – 1 week, and SCARED. We compared the continuous reliability measures derived from each metric to both the mean value between the two visits and the change in value across visits. Similarly, medication status was categorized as a binary value (taking/not taking) for psychiatric medications and nonpsychiatric medications. To examine the effect of changing medications on reliability, we compared the change in medication status to the continuous measures. As head motion has a significant impact on reliability, we correlated the measures with average motion and average max framewise displacement. We also explored possible associations between participant age and different reliability measures.

**Results**

\subsection\*{Univariate Reliability}

\noindent Edgewise reliability of functional connectivity was poor for both depressed and healthy volunteers. However, the depressed participants had higher mean ICC, indicating improved reliability. Regions of the brain with more and less reliable edges can be seen in Fig. X. The largest groupwise differences occurred in these connections XXX.

\subsection\*{Multivariate Reliability}

\noindent Multivariate metrics reflected improved reliability compared to univariate. Fingerprinting values were moderate (XXX), but significantly greater than chance (Poisson (1) test, see XXX). Fingerprinting accuracy did not differ between groups. Both groups were highly discriminable (better than chance XXX) and did not differ.

\subsection\*{Effect size analysis}

Edge-level effect sizes between depressed and healthy participants were not correlated with any edge-level reliability measures. This suggests that the reliability of an edge is not related to its significance in groupwise differences.

\subsection\*{Correlations with clinical measures and confounds}

Reliability measures were not correlated with average symptoms or change in symptoms after correcting for multiple comparisons. There were also no correlations with motion measures or age.