**Test-retest reliability of a multivariate brain pattern in task-fMRI:** an analysis of the Neurological Pain Signature (NPS)

**Abstract** (count: 210 words)

Multivariate brain patterns developed by machine learning in fMRI studies outperform the traditional voxel-wise brain measures in predicting basic mental processes and clinical symptoms. As the multivariate brain patterns being more widely used as a biomarker to predict individual differences in mental processes, efforts to establish reliability and other measurement properties of candidate biomarkers will become more and more important. In this paper, we tested the reliability of a well-established multivariate brain pattern - the Neurologic Pain Signature (NPS) - that was trained to track pain induced by nociceptive input in three datasets with 11 studies. NPS showed good to excellent performance in the short-term (within one day) reliability (N = 321; ICC = 0.83 on average, ranging from 0.68 to 0.91). Reliability stayed at a comparable level giving the same number of trials with 1 week (N = 27) and 1 month (N = 120) interval between sessions. Excellent reliability was achieved with more than 60 trials per condition. Stimuli with larger effect size, such as higher temperature of heat stimuli, tend to produce measurement with higher reliability. Brain measures calculated in contrast with the baseline have higher reliability than in contrast with a control condition with low reliability itself. Our study demonstrated promising reliability performance of a multivariate brain pattern in the pain domain and provides a blueprint for studying reliability of other multivariate brain patterns in the future.

**Introduction** (count: 1092 words)

People differ in perceiving, thinking and acting in the environment. Understanding the individual differences in brain activities and its link with the behavior is a main focus of neuroimaging studies. With the development of fMRI techniques and data analysis methods, our knowledge regarding the brain-behavior association grows unprecedentedly. In classic studies, functions and processes are thought to be encoded in isolated brain regions of interest. Traditional brain mapping treats each local region or ‘voxel’ separately (Scoville & Milner,1957; Fodor, 1983) and conducts tests in a massive univariate way, also known as ‘voxelwise encoding model’ (Woo et al., 2017). In recent years, however, more and more evidence supports brain representations relevant for performance and clinical outcomes may be distributed across multiple regions and networks (). Design of traditional brain mapping only permits the inference that the presentation of a stimulus (or symptom) leads to an activation of a brain region, but does not allow the reverse inference, i.e., brain activities could predict the presentation of a stimulus (or symptom). As the machine learning being introduced to the neuroimaging studies (Haxby et al., 2002; ), there is a sea change to treat brain activities across multiple regions as biomarkers to quantitatively predict the basic mental processes (Wager et al., 2013; Woo et al., 2017; Yu et al., 2019; Chang et al, 2015; ) and clinical symptoms (for reviews, see Arbabshirani et al., 2017; Woo et al., 2017a). A brain biomarker is a spatial pattern of regression weights across brain regions, optimized by machine learning, which could be applied to brain activity maps obtained from new participants and make predictions with high sensitivity and specificity. This new way of thinking brings translational goals within reach finally (Woo et al., 2017).

As the biomarkers being more widely used to predict individual differences in mental processes, efforts to establish reliability and other measurement properties of candidate biomarkers will become more and more important. Reliability and validity are two of the most important measurement properties to evaluate the quality of a measurement (Streiner, 2003; Drost, 2011). Reliability is the ability of a measurement to give consistent results under similar conditions. Validity is to evaluate whether a measurement measures what it claims to measure. For example, IQ tests should measure intelligence instead of something else. A measurement score is composed of three parts: the true score of the construct of interest (T), systematic bias (B) and random error (E) associated with the measurement. Reliability equals the ratio of the variance of T + B divided by the variance of the total score (i.e., T + B + E), whereas validity equals the ratio of the variance of T divided by the total variance. Thus reliability places an upper limit on the validity of a measurement, though a reliable measure is not necessarily a valid measure due to the systematic bias (Streiner, 2003; Drost, 2011; Nobel et al., 2019).

Test-retest reliability is one type of reliability that assesses stability under repeated tests. There is a rich history of test-retest reliability studies of fMRI measurements, which have shown widely varied reliability performances in task-fMRI, such as averaged-voxel activation in brain regions of interest (ROI; Elliott, Knodt et al., 2020) and functional connectivity between ROIs (Letzen et al., 2016), and in resting state fMRI, such as individual edge-level connectivity (Noble et al., 2019), functional connectomics (Zuo and Xing, 2014), graph theory metrics (Braun et al., 2012; Cao et al., 2014; Termenon et al., 2016; Wang et al., 2011) and dynamic measures (Choe et al., 2017; Zuo et al., 2010a). Though most of the existing fMRI studies used univariate or mass univariate inferential strategies in their data analyses, some recent studies suggested that multivariate measures could improve the test-retest reliability substantially (Woo and Tor, 2016; Yoo et al., 2019). Several significant differences between univariate measures and multivariate brain patterns might influence their performance in the test-retest reliability. Firstly, univariate measures typically have not specified the location and topography with sufficient precision, thus limiting direct replications. For example, in meta-analyses, there is no consensus regarding how close findings should be considered as replications. By contrast, multivariate pattern signatures can specify a precise set of voxels and the topography of the relative expected activity levels across voxels, providing a basis for exact replication. Secondly, region of interest analyses, i.e., average of univariate responses within individual brain regions, are usually large, encompassing neurons with different functions, which dilutes signal and reduces their functional specificity. The multivariate pattern signatures, however, could capture fine-grained functional organization and can more accurately predict perceptions and behaviours ( ). Lastly, traditional univariate analyses mainly focus on which local regions highly related to stimuli or clinical outcomes. However, many features of neurologic and psychiatric disorders (e.g., pain, negative emotions, cognitive and social processes) are more likely encoded in distributed neural systems involving networks of many regions (Wager et al., 2013; Chang et al., 2015). In such situations, multivariate pattern signatures that could integrate contributions from different brain areas will likely be required for accurate prediction.

In the current study, we aim to systematically evaluate the test-retest reliability of a multivariate brain biomarker in the pain domain. Neurologic Pain Signature (NPS) is a well-validated multivariate brain biomarker, whose weights are optimized to be maximally predictive of pain based on fMRI signals (Wager et al., 2013). The NPS could accurately predict pain experience in response to noxious thermal (Wager et al., 2013; Brascher et al., 2016), mechanical (Krishnan et al., 2016), and electrical stimuli (Krishnan et al., 2016; Ma et al., 2016), but does not respond to non-noxious warm stimuli (Wager et al., 2013), threat cues (Wager et al., 2013; Krishnan et al., 2016; Ma et al., 2016), social rejection-related stimuli (Wager et al., 2013), observed pain (Krishnan et al., 2016), or aversive images (Chang et al., 2015). The signature does not measure a disorder, but rather a basic mental process regarding negative sensory and affective processes, which can serve as intermediate features that are altered in various disorders. For example, Lopez-Sola et al. (2016) found that enhanced NPS responses, combined with another brain signature related to non-painful sensory processing, discriminated fibromyalgia from pain-free controls with 93% accuracy. Before testing the reliability of NPS, we first validate the NPS response to pain stimuli by examining the effect size of NPS in response to noxious thermal stimuli in a large dataset with 9 published studies, more than 17,000 fMRI images related to single-stimulus epochs collected in over 320 participants. We then tested the short-term test-retest reliability of NPS in the same dataset, with the first and second half trials collected within one day. Lastly, we examined the long-term test-retest reliability of NPS in another two datasets with data collected in multiple sessions (N = 27, 1 week interval between 3 sessions; and N = 120, 1 month interval between 2 sessions).